Traffic Risk Mining From Heterogeneous Road Statistics

Koichi Moriya, Shin Matsushima, and Kenji Yamanishi

***Abstract*— At present, a large amount of traffic-related data is obtained manually and through sensors and social media, e.g., traffic statistics, accident statistics, road information, and users' comments. In this paper, we propose a novel framework for mining traffic risk from such heterogeneous data. Traffic risk refers to the possibility of occurrence of traffic accidents. Specifically, we focus on two issues: 1) predicting the number of accidents on any road or at intersection and 2) clustering roads to identify risk factors for risky road clusters. We present a unified approach for addressing these issues by means of feature-based non-negative matrix factorization (FNMF). In particular, we develop a new multiplicative update algorithm for the FNMF to handle big traffic data. Using real-traffic data in Tokyo, we demonstrate that the proposed algorithm can be used to predict traffic risk at any location more accurately and efficiently than existing methods, and that a number of clusters of risky roads can be identified and characterized by two risk factors. In summary, our work can be regarded as the first step to a new research area of traffic risk mining.**

***Index Terms*— Learning systems—unsupervised learning, machine intelligence—pattern analysis.**

I. Introduction

A. Background of Traffic Risk Mining

THE fact that published traffic data are becoming increasingly varied and heterogeneous is noteworthy. Indeed, the data may include not only traffic statistics but also information collected through a variety of sensors and social networks. Recent years have witnessed efforts to use such traffic data for a wider range of purposes, such as safety management, driver support, traffic infrastructure design, and disaster prevention. For example, a private service called SAFETY MAP is available on the internet [1] (see Figure 1). It shows various traffic statistics, such as the frequencies of accidents and braking at different locations on the road map. It also shows comments posted online by drivers and pedestrians for these locations. Such information has been used by Japanese local governments to identify high-risk locations and adopt safety measures accordingly (see [2]). However, such services present a number of problems.

1) Necessity of Completing Risk Information: The first problem is the unavailability of data for all of the locations on the map. Hence, there are some locations for which it is not possible to predict the traffic accident risk regardless of their high potential risk. This problem has prompted us to consider developing a methodology for completing and predicting the risk information at any location by using data from other locations.

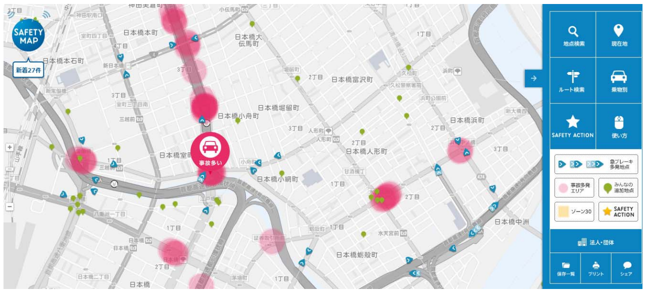


Fig. 1. Screen shot of SAFETY MAP [1]. The density of each red circle indicates how many traffic accidents occurred at that location previously. Users have the option of posting their opinions online, and this facility has been used for actual decision making by Japanese local governments.

2) Necessity of Discovering Global Knowledge of Traffic Risk: The second problem is the isolated nature of the traffic risk information associated with each location, which complicates the development of a comprehensive view of traffic risk. If the traffic risk information could be shared among locations with similar road conditions, it would be possible to extract their common traffic risk patterns. Thus, it would be possible to understand the degree of risk involved by referring to the risk patterns. Therefore, it is necessary to combine the information from all the locations in order to obtain a global perspective that would be useful for ranking the risk associated with different locations.

This paper describes our efforts toward solving the two above-mentioned problems for real traffic data. Our contributions are twofold. First, we propose a novel framework for traffic risk mining that involves the combined use of data from heterogeneous sources, such as traffic statistics, sensor data, and social networking data. Thus, we provide a unified methodology for solving the two above-mentioned problems. Second, we report the traffic mining results obtained for real datasets, including (i) the results for predicting the number of accidents on roads and at intersections and (ii) the results related to knowledge discovery concerning traffic risk factors. Finally, we discuss the usefulness of these results for road design and safety management. In summary, we establish a new research area of traffic risk mining.

B. Novelty and Significance of This Work

The novelty and significance of this paper can be summarized as follows.

Manuscript received November 28, 2017; revised May 5, 2018; accepted June 16, 2018. Date of publication September 13, 2018; date of current version November 9, 2018. This work was supported by JST CREST under Grant JPMJCR1304. The Associate Editor for this paper was R. Trasarti. *(Corresponding author: Shin Matsushima.)*

The authors are with the Department of Information Science and Technology, The University of Tokyo, Tokyo 133-8656, Japan (e-mail: koichi_moriya@mist.i.u-tokyo.ac.jp; shin_matsushima@mist.i.u-tokyo.ac.jp; yamanishi@mist.i.u-tokyo.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2018.2856533

1) A Novel Framework for Traffic Risk Mining Using Feature-Based Non-Negative Matrix Factorization: We propose a framework for mining traffic risk information from heterogeneous datasets consisting of traffic statistics (number of accidents, traffic volume, roadway information, brake data, and social opinion.) The framework has two main functions: (A) predicting traffic risk (i.e., the number of traffic accidents) at any location, and (B) knowledge discovery of traffic risk patterns by clustering roadways. We introduce a novel methodology for achieving both (A) and (B) simultaneously by means of *feature-based non-negative matrix factorization* (FNMF).

2) A Multiplicative Update Formula for FNMF: We develop a new multiplicative update algorithm for FNMF, which facilitates efficient and accurate processing of large amounts of traffic data. Moreover, the proposed method and update algorithm can be used to process other big data. This algorithm enables us to predict the number of accidents more accurately and efficiently than existing methods. Furthermore, we can cluster risky locations with respect to several risk factors.

3) Empirical Demonstration of Our Methodology Using Real Traffic Datasets: We demonstrate the effectiveness of our methodology using real traffic data in Tokyo, Japan. We show that our methodology can predict the number of accidents with a mean absolute error of 0.50. We also show that roads and intersections in Tokyo with high traffic risk can be characterized in terms of two risk factors. In summary, our work can be regarded as the first step toward a new research area of traffic risk mining.

This paper extends the original work presented in [17] in the following aspects:

- Thorough description of the multiplicative updates and derivation of the multiplicative update formula for generalized Kullback–Leibler (KL) divergence minimization.
- Verification of traffic risk mining by comparison with linear regression model.
- Development of methods for evaluating and characterizing features of extracted clusters using stochastic decision olist.
- Detailed cluster analysis using the proposed method and characteristics of actual images of locations.

C. Previous Work

Several studies related to traffic data mining have been reported. Most of these studies have focused on flow prediction and path exploration (see e.g., [9], [16]). Bashah and Hill [4], whose work involved data mining of traffic accidents, analyzed the causes of accidents using prediction methods. Krishnaveni and Hemalantha [13] focused on prediction and characterization of the severity of injury resulting from traffic accidents. Chang and Chen [5] analyzed the factors underlying frequent accidents. Bayam *et al.* [3] analyzed the relationship between drivers' age and accidents. Chong *et al.* [7] adopted machine learning methods to model the severity of injury due to traffic accidents.

Matrix factorization has been applied to prediction problems in which the prediction of unknown data is considered as the completion of missing data (see [11]); its applications include recommendation [15], image recognition [21], and power spectrum identification for voice data [18]. In addition, it has been used for inference from traffic data [19].

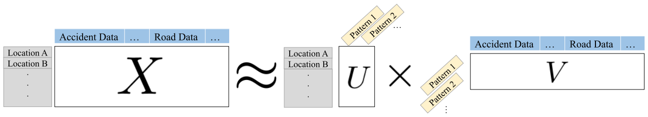


Fig. 2. Concept of matrix factorization-based framework: we aim to represent i-th row of the data matrix, $X_{i:}$, by $\sum_{k=1}^{K} u_{ik} V_{k:}$, a linear combniantion of rows of V denoted by $V_{k:}$. By doing so, we can interpret V as a set of row vectors that represents typical patterns for $X_{i:}$, and u_{ik} as the strength of the k-th pattern contained in $X_{i:}$. In our work, $X_{i:}$ and $V_{k:}$ represents each location and its typical pattern, respectively.

The remainder of this paper is organized as follows. In Section 2, we review existing work based on the use of non-negative factorization and its application to prediction and clustering. In Section 3, we introduce our methodology for analyzing integrated traffic data and derive a new multiplicative update formula for a variant of feature-based matrix factorization. In Section 4, we empirically demonstrate the effectiveness of our methodology using real traffic data. In Section 5, we develop a method for evaluating and characterizing the obtained clusters and perform cluster analysis on the obtained results. Finally, in Section 6, we conclude the paper.

II. Matrix Factorization for Prediction and Clustering

A. Matrix Factorization-Based Framework

The matrix factorization method is a well-established technique for extracting latent information from data. The family of matrix factorization methods includes principal component analysis (PCA), canonical correlation analysis (CCA), and non-negative matrix factorization (NMF). In this paper, we use a matrix factorization-based method to extract risk factors from data originating from high-risk locations for predicting the number of accidents at unseen locations and categorizing high-risk locations into clusters (see Figure 2). More specifically, we factorize a matrix in which each row corresponds to a high-risk location and each column corresponds to its attributes. We aim to realize a factorization that represents each high-risk location by a combination of representative patterns of high-risk locations, with corresponding coefficients for each pattern.

Matrix factorization is known as a useful methodology for extracting components capable of explaining the underlying structure of a data matrix of interest. Two methods based on this methodology are non-negative matrix factorization (NMF) and feature-based matrix factorization (FMF). Our model, which is introduced in Section 3, is based on these two methods. Next, we review these two methods and describe how they can be applied to prediction and clustering problems.

B. Non-Negative Matrix Factorization

In particular, NMF focuses on the analysis of non-negative matrices. Imposing non-negative constraints on the model parameters allows for enhanced interpretability of the extracted components and clustering using the estimated parameters.

Given a matrix $X \in \mathbb{R}_+^{N\times M}$, we aim to find U, V such that $X \approx UV$. In typical cases where $\mathrm{rank}(X) \ll N, M$ holds, we expect that X can be approximated by $U \in \mathbb{R}_+^{N\times K}$, $V \in \mathbb{R}_+^{K\times M}$ with a small rank, $K \ll N, M$. The parameter estimation problem can be formulated as the following minimization problem using an approximation measure $E(x_{ij}, \hat{x}_{ij})$:

$$\text{minimize} \sum_{i,j} E(x_{ij}, \hat{x}_{ij}), \tag{1}$$

$$\text{subject to } \hat{x}_{ij} = \sum_{k=1}^{K} u_{ik} v_{kj}. \tag{2}$$

We denote the (i, j)-th element of X by x_{ij} and so forth for U and V. For the approximation measure E, the mean squared error and generalized KL divergence are well known [8]:

$$\mathrm{MSE}(x_{ij}, \hat{x}_{ij}) = \left(x_{ij} - \hat{x}_{ij}\right)^2,$$
$$\mathrm{KL}'(x_{ij}, \hat{x}_{ij}) = \left(x_{ij} \log \frac{x_{ij}}{\hat{x}_{ij}} - x_{ij} + \hat{x}_{ij}\right).$$

Furthermore, an iterative method with a multiplicative update formula is well known [14], [25] and widely used. In the case of $\mathrm{MSE}(x_{ij}, \hat{x}_{ij})$, we perform the following update until the values of the parameters converge:

$$u_{ik} \leftarrow u_{ik} \frac{\sum_j x_{ij} v_{kj}}{\sum_{j,l} u_{il} v_{kj} v_{lj}}, \quad v_{kj} \leftarrow v_{kj} \frac{\sum_i x_{ij} u_{ik}}{\sum_{i,l} v_{lj} u_{ik} u_{il}}.$$

This update formula can be derived using the expectation maximization (EM) algorithm, which is not necessarily derived from a statistical problem. The key idea for understanding the EM algorithm in this way is that the E-step provides an upper bound for the target objective function, which is derived from Jensen's inequality, whereas the M-step minimizes the derived upper bound.

The update formula in matrix form can be validated from the perspective of the EM algorithm, i.e., minimization with respect to u_{ik} in the M-step is independent of the other elements of U given a fixed V and vice versa. Thus, we can compute $U^{(t+1)}$ using $\left(U^{(t)}, V^{(t)}\right)$ and $V^{(t+1)}$ using $\left(U^{(t+1)}, V^{(t)}\right)$, which results in the following update formula:

$$U^{(t+1)} = U^{(t)} \odot \frac{X(V^{(t)})^{\mathrm{T}}}{U^{(t)}V^{(t)}(V^{(t)})^{\mathrm{T}}},$$
$$V^{(t+1)} = V^{(t)} \odot \frac{(U^{(t+1)})^{\mathrm{T}}X}{(U^{(t+1)})^{\mathrm{T}}U^{(t+1)}V^{(t)}},$$

where $\odot$ denotes element-wise multiplication and fractional expressions of matrices represent element-wise division.

C. Feature-Based Matrix Factorization

When additional information is available about a particular element, row, or column of X, it is desirable to incorporate such information to explain the elements of X, in addition to using the term $\sum_j u_{ik} v_{kj}$. Chen *et al.* extended the general matrix factorization model to the following form, which is referred to as feature-based matrix factorization (FMF) [6]:

$$\begin{aligned}
&\text{minimize} \sum_{i,j} (x_{ij} - \hat{x}_{ij})^2 + r,\\
&\text{subject to } \hat{x}_{ij} = \mu + \sum_{i'=1}^{N} a_{i'} \alpha_{i'}(i, j)\\
&\qquad + \sum_{j'=1}^{M} b_{j'} \beta_{j'}(i, j) + \sum_{l=1}^{L} c_l \gamma_l(i, j)\\
&\qquad + \sum_{k=1}^{K} \left(\sum_{i'=1}^{N} u_{i'k} \alpha_{i'}(i, j) \right) \left(\sum_{j'=1}^{M} v_{kj'} \beta_{j'}(i, j) \right),\\
&\qquad r = \lambda_1 \sum_{i,k} u_{ik}^2\\
&\qquad + \lambda_2 \sum_{k,j} v_{kj}^2 + \lambda_3 \sum_l c_l^2 + \lambda_4 \sum_i a_i^2 + \lambda_5 \sum_j b_j^2.
\end{aligned}$$

α, β, and γ constitute auxiliary information that is used to estimate the elements of X, and a, b, and c are the corresponding parameters to be estimated. μ is the mean value of x_{ij}. a, b are dependent only on i, j, respectively, whereas c is independent of i, j. The last term is an extended form of matrix factorization (2). Each λ is a hyperparameter dependent on the problem. When we set $\mu = 0, c \equiv \emptyset, a \equiv \emptyset, b \equiv \emptyset, \alpha_{i'}(i, j) = \begin{cases} 0 & (i' \neq i) \\ 1 & (i' = i) \end{cases}, \beta_{j'}(i, j) = \begin{cases} 0 & (j' \neq j) \\ 1 & (j' = j) \end{cases}, \lambda = 0$, the model coincides with standard matrix factorization. Thus, using this formulation, we can learn more complex models compared to those using matrix factorization. However, the new form of the model is no longer able to derive multiplicative updates. Therefore, gradient descent was used for parameter estimation in [6], which is expected to be much slower compared to multiplicative updates.

D. Prediction by Completion

Here, it is assumed that all the elements in X are occupied by observed values. Prediction can be performed by incorporating missing values of the elements of matrix X and setting target entries to be predicted for the missing values. By completing the missing entries of X using an estimated parameter, we can obtain predictions for the respective missing entries.

Zhang *et al.* [24] proposed weighted non-negative matrix factorization (WNMF) to process a data matrix with missing values by rewriting the objective function of matrix factorization (1) as $\sum_{x_{ij} \in X^o} E\left(x_{ij}, \hat{x}_{ij}\right)$, where X^o is the set of observed values in X. We define the weight matrix W with the same size as X as an indicator to distinguish between observed or missing values:

$$W = (w_{ij}), w_{ij} = \begin{cases} 1 & x_{ij} \in X^o \\ 0 & x_{ij} \notin X^o. \end{cases}$$

Then, the update formula can be represented as

$$U^{(t+1)} = U^{(t)} \odot \frac{(W \odot X)\,(V^{(t)})^{\mathrm{T}}}{\left(W \odot (U^{(t)}V^{(t)})\right)(V^{(t)})^{\mathrm{T}}},$$
$$V^{(t+1)} = V^{(t)} \odot \frac{(U^{(t+1)})^{\mathrm{T}}\,(W \odot X)}{(U^{(t+1)})^{\mathrm{T}}(W \odot U^{(t+1)}V^{(t)})}.$$

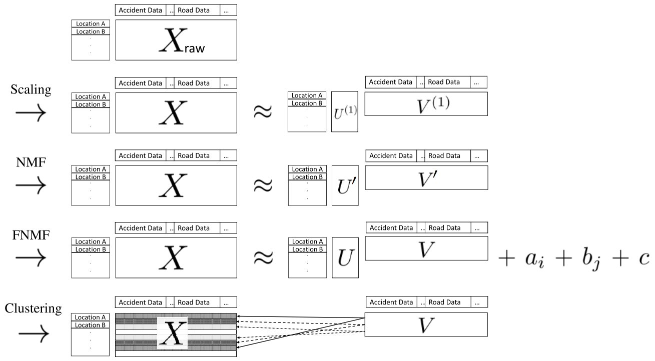


Fig. 3. Proposed framework for traffic risk mining.

This formulation can also be applied to other models, such as FMF.

E. Clustering by Factorization

We are motivated to extract patterns and clusters of traffic locations that share the same risk patterns. Clustering using NMF has been studied especially for document clustering tasks [12], [20], [22]. As NMF approximates i-th row of the data matrix $X_{i:}$ by a linear combniantion of rows of V, we can interpret V as a set of row vectors that represents typical patterns for $X_{i:}$, and u_{ik} as the strength of the k-th pattern contained in $X_{i:}$. When the i-th row of the data matrix represents the number of occurrences of each type of accident at a high-risk location i, we can interpret $V_{k:}$ as a typical pattern of high-risk locations, and u_{ik} as the degree to which the i-th location belongs to the k-th pattern. Clustering of the rows can be performed according to this interpretation. Xu *et al.* [22] proposed a clustering method using the result of NMF for labeling each row i by the label l_i on the basis of the following rule:

$$l_i = \operatorname*{argmax}_{k=1,\dots,K} \; u_{ik}.$$

Note that this can be done only in the case of NMF. Without the non-negativity, each parity can be chosen arbitrarily.

III. Proposed Method

In this section, we introduce our method for traffic risk mining, which consists of the following three processes:

- Form a matrix by unifying all the datasets and scaling them appropriately, and find the initial value for U, V by using the k-means method only with respect to traffic accident data.
- After several updates of non-negative matrix factorization, learn a variant of feature-based matrix factorization by using a new multiplicative update formula.
- Perform clustering by using information of the estimated parameters and determine the number of clusters on the basis of the data related to accidents in each cluster.

The entire flow of the framework is shown in Figure 3. In the following sections, we discuss the details with respect to each of these three processes. For details of the data, see Section IV-A.

A. Scaling and Initialization

In most cases, data that are suitable for NMF processing, such as review data, voice data, and image data, have a matrix with the same order of values for each row and column. By contrast, for data matrices with values whose order varies from column to column, the minimization procedure

$$\text{minimize} \sum_{i,j} w_{ij}(x_{ij} - \hat{x}_{ij})^2$$

would have a major influence on summands with larger values of x_{ij}, which is likely to result in summands with relatively small values of x_{ij} being ignored. Therefore, if NMF is applied naively to such a data matrix, we obtain parameters that result in small errors for columns of larger orders, while generating comparatively large errors between x_{ij} and its prediction, where the order of x_{ij} is relatively small. Therefore, shrinking the order of the values of such columns with a large number of values would facilitate improved performance in terms of prediction of the accident statistics by suppressing the influence of columns for which we are not interested to make predictions. However, shrinking these columns excessively could cause them to be ignored, which would ultimately have a negative impact on the prediction accuracy.

Our aim is to classify accident patterns and predict missing entries using heterogeneous roadway information. We use datasets consisting of traffic flow, brake rates, and other roadway statistics, such as intersection density, velocity limits, and the number of lanes. Thus, scaling appears to be an issue when we perform matrix factorization with respect to a matrix integrating such attributes. We adopt a scaling rule,

$$x^s_{ij} = \frac{\alpha x_{ij}}{\max_i x_{ij}} \quad (\forall j \in S, i = 1, \dots, n),$$

where x^s_{ij} is the value after scaling, $S \subset \{1, \dots, M\}$ is the set of columns that has to be scaled, and α is a small positive number ($1 \le \alpha \le 3$). Note that the order of the scaled values, $0 \le x^s_{ij} \le \alpha$, is always less than the order of the number of traffic accidents, $0 \le x_{ij} \le 10$. The issues here are to determine (1) which columns should be in S, and (2) what value must be set for α. For (1), we adopt the following scaling rule:

- Values provided as percentages are divided by 100 unless the maximum observed value is less than 10.0.
- Traffic flows per hour are divided by the maximum values in the day/night time for each category.
- Values representing the length of streets, total flows, and velocity limits are divided by the maximum of the respective rows
- Values representing the number of lanes, density of intersections, and other values that are always smaller than 1 are not scaled.

For (2), we determined the value of α by performing comparative experiments, the details of which appear in Section IV-B.

In general, matrix factorization significantly depends on the choice of the initial parameters. Therefore, we specified a value for the initial parameter with the aim of controlling the quality of the results. We applied the k-means method to the part of the matrix that contains traffic accident data, such that the factorization reflected the different types of risks. Given the mean vectors $c_k \in \mathbb{R}^{m_a \times 1}$ for each cluster k and the labels for each row $l_i = \operatorname{argmin}_k \left\| x_{i,1:m_a} - c_k \right\|$ as a result of the k-means method, we define $\hat{C} = \left(\hat{c}_1^{\mathrm{T}}, \hat{c}_2^{\mathrm{T}}, \ldots, \hat{c}_K^{\mathrm{T}}\right)^{\mathrm{T}}$ by $\hat{c}_k = \sum_{i:l_i=k} x_{i,m_a+1:M}$, and we set $V^{(1)} = [C, \hat{C}]$ as the initial value for V. Here, $x_{i,1:m_a}$ is the part corresponding to traffic accident statistics, i.e., $m_a < M$ is the number of attributes pertaining to traffic accidents, and $C = \left(c_1^{\mathrm{T}}, c_2^{\mathrm{T}}, \ldots, c_K^{\mathrm{T}}\right)^{\mathrm{T}}$. For U, we define

$$U^{(1)} = (u_{ik}), \quad u_{ik} = \frac{\|x_{i,1:m_a} - c_k\|_2^{-1}}{\sum_l \|x_{i,1:m_a} - c_l\|_2^{-1}}.$$

This expression reflects the similarity between each traffic accident data value and the mean vector of each cluster.

B. Feature-Based Non-Negative Matrix Factorization and Its Multiplicative Update

In Section II-B, we have seen that NMF can be used for both prediction and clustering owing to the non-negativity of the parameters and that it allows for efficient multiplicative updates. On the other hand, in Section II-C, we have seen that FMF can represent a much more complex model; however, a less efficient method of gradient descent has been used. We consider the combination of both models to establish a more complex model than NMF to allow clustering and multiplicative updates. We consider the following model, which we refer to as feature-based non-negative matrix factorization (FNMF).

$$\text{minimize} \sum_{i,j} w_{ij} E(x_{ij}, \hat{x}_{ij}), \tag{3}$$

$$\text{subject to } \hat{x}_{ij} = c + a_i + b_j + \sum_{k=1}^{K} u_{ik} v_{kj}, \tag{4}$$

$$u_{ik} \geq 0, \quad v_{kj} \geq 0, \quad \hat{x}_{ij} \geq 0. \tag{5}$$

Note that the non-negativity constraints are only applied to $U, V, \hat{X}$, and not to a, b, c. This model is strictly broader than the NMF model when K is set to the same value. In the sense of the argument presented in Section II-B, the EM algorithm is unable to process negative values of parameters directly; therefore, we propose an analytical form for updating a, b, and c as well as U and V.

First, we consider the mean squared error for the approximation measure E. Therefore, we minimize

$$\sum_{i,j} w_{ij} (x_{ij} - \hat{x}_{ij})^2.$$

The minimization problem (3) is equivalent to the maximization problem of the function $J(U, V, a, b, c)$ defined by

$$\sum_{i,j} \Bigg(2(x_{ij} - a_i - b_j - c) \sum_k u_{ik} v_{kj} - \left(\sum_k u_{ik} v_{kj} \right)^2 - a_i^2 - b_j^2 - c^2 - 2a_i b_j - 2a_i c - 2b_j c + 2(a_i + b_j + c) x_{ij} \Bigg).$$

Setting the parameter before and after an update to (u_{ik}^0, v_{kj}^0) and (u_{ik}, v_{kj}), respectively, we define $h_{ijk}^0 = \frac{u_{ik}^0 v_{kj}^0}{\sum_{l=1}^K u_{il}^0 \; v_{lj}^0}$. Then, by Jensen's inequality,

$$\begin{aligned} -\left(\sum_k u_{ik} v_{kj} \right)^2 &= -\left(\sum_k h_{ijk}^0 \frac{u_{ik} v_{kj}}{h_{ijk}^0} \right)^2 \\ &\geq -\sum_k h_{ijk}^0 \left(\frac{u_{ik} v_{kj}}{h_{ijk}^0} \right)^2 \\ &= -\sum_k \frac{1}{h_{ijk}^0} \left(u_{ik} v_{kj} \right)^2. \end{aligned}$$

Therefore, we have $J(U, V, a, b, c) \geq J'(U, V, a, b, c)$, where $J'(U, V, a, b, c)$ is equal to

$$\sum_{i,j} \Bigg(\sum_k \left(2(x_{ij} - a_i - b_j - c) u_{ik} v_{kj} - \frac{1}{h_{ijk}^0} \left(u_{ik} v_{kj} \right)^2 \right) - a_i^2 - b_j^2 - c^2 - 2a_i b_j - 2a_i c - 2b_j c + 2(a_i + b_j + c) x_{ij} \Bigg).$$

We aim to monotonically increase the objective function by maximizing $J'(U, V, a, b, c)$. By incorporating the constraints (5), we can write the Lagrangian $L(U, V, a, b, c, \alpha, \beta, \gamma)$ as

$$J' - \sum_{i,k} \alpha_{ik} u_{ik} - \sum_{k,j} \beta_{kj} v_{kj} - \sum_{i,j} \gamma_{ij} \left(\sum_k u_{ik} v_{kj} + a_i + b_j + c \right).$$

By differentiating J' with respect to u_{ik}, we obtain

$$u_{ik}^{(t+1)} = u_{ik}^{(t)} \frac{\sum_j (x_{ij} - a_i - b_j - c) v_{kj}^{(t)}}{\sum_{j,l} u_{il}^{(t)} v_{kj}^{(t)} v_{lj}^{(t)}}.$$

Note that the update formula of v_{kj} uses $u_{ik}^{(t+1)}$ instead of $u_{ik}^{(t)}$ when v_{kj} is updated after u_{ik}. Then, we obtain the update of v_{kj}:

$$v_{kj}^{(t+1)} = v_{kj}^{(t)} \frac{\sum_i (x_{ij} - a_i - b_j - c) u_{ik}^{(t+1)}}{\sum_{i,l} (u_{ik}^{(t+1)})^2 / u_{ik}^{(t)} v_{lj}^{(t)} u_{il}^{(t)}}.$$

By the Karush–Kuhn–Tucker (KKT) conditions, we can see that $\alpha_{ik} u_{ik} = 0$. If $u_{ik} > 0$, then $\alpha_{ik} = 0$, and we get $\frac{\partial}{\partial u_{ik}} L(U, V, a, b, c, \alpha, \beta, \gamma) = \frac{\partial}{\partial u_{ik}} J'(U, V, a, b, c) + \alpha_{ik} = 0$. Otherwise, we can get this equation satisfied by setting $u_{ik} = 0$ and α_{ik} to an appropriate positive number. Thus, we obtain the update for u_{ik}; similarly, we get the updates for v_{kj}. Next, by differentiating J' with respect to a_i, we get

$$a_i \leftarrow \sum\nolimits_j \left(x_{ij} - b_j - c - \sum\nolimits_k u_{ik} v_{kj} \right) / M.$$

If there exist i and j such that $\sum_{k=1}^K u_{ik} v_{kj} + a_i + b_j + c < 0$ holds, we can satisfy the KKT conditions by setting $\sum_{k=1}^K u_{ik} v_{kj} + a_i + b_j + c = 0$. Thus, we can set a_i to $\max_j \sum_{k=1}^K u_{ik} v_{kj} + a_i + b_j + c$. We can derive the update for

b_i and c similarly. When we incorporate the missing values, the above-mentioned argument holds similarly.

Finally, we introduce the weight matrix for prediction, and the update formula becomes

$$\begin{aligned}
\bar{Y} &= (\bar{y}_{ij}), \quad \bar{y}_{ij} = w_{ij}(x_{ij} - a_i - b_j - c),\\
Y^{(t)} &= U^{(t)}V^{(t)},\\
P &= \frac{\bar{Y}(V^{(t)})^{\mathrm{T}}}{(W \odot Y^{(t)})(V^{(t)})^{\mathrm{T}}},\\
Q &= \frac{(U^{(t+1)})^{\mathrm{T}}\bar{Y}}{(U^{(t+1)} \odot P)^{\mathrm{T}}(W \odot Y^{(t)})},\\
U^{(t+1)} &= U^{(t)} \odot P,\\
V^{(t+1)} &= V^{(t)} \odot Q,\\
u_{ij}^{(t+1)} &\leftarrow \max(0, u_{ij}^{(t+1)}), \quad v_{ij}^{(t+1)} \leftarrow \max(0, v_{ij}^{(t+1)}),\\
a_i^{(t+1)} &= \max\left(\frac{\sum_j w_{ij}\left(x_{ij} - y_{ij} - b_j^{(t)} - c^{(t)}\right)}{\sum_j w_{ij}},\right.\\
&\qquad \left.\times \max_j(-y_{ij} - b_j^{(t)} - c^{(t)})\right),\\
b_j^{(t+1)} &= \max\left(\frac{\sum_i w_{ij}\left(x_{ij} - y_{ij} - a_i^{(t+1)} - c^{(t)}\right)}{\sum_j w_{ij}},\right.\\
&\qquad \left.\times \max_i\left(-y_{ij} - a_i^{(t+1)} - c^{(t)}\right)\right),\\
c^{(t+1)} &= \max\left(\frac{\sum_{i,j} w_{ij}\left(x_{ij} - y_{ij} - a_i^{(t+1)} - b_j^{(t+1)}\right)}{\sum_{i,j} w_{ij}},\right.\\
&\qquad \left.\times \max_{i,j}(-y_{ij} - a_i^{(t+1)} - b_j^{(t+1)})\right).
\end{aligned}$$

Next, we consider the KL divergence for the approximation measure. We minimize

$$\sum_{i,j} w_{ij}\left(x_{ij}\log\frac{x_{ij}}{\hat{x}_{ij}} - x_{ij} + \hat{x}_{ij}\right).$$

The minimization problem is equivalent to the maximization problem of the function $J(U, V, a, b, c)$ defined by

$$\sum_{i,j}\left(x_{ij}\log\left(\sum_k u_{ik}v_{kj} + a_i + b_j + c\right) - \hat{x}_{ij}\right).$$

We set the parameter before and after an update to (u_{ik}^0, v_{ik}^0) and (u_{ik}, v_{kj}), respectively, and we set a_i, b_j, c likewise. We consider the sequence p_{ij} defined by

$$(p_{ij})_{k,k=1\ldots K+3} = \{u_{i1}v_{1j}, u_{i2}v_{2i}, \ldots, u_{iK}v_{Kj}, a_i, b_j, c\}.$$

If we write $(p_{ij})_k^0$, it denotes the parameter $(p_{ij})_k$ before an update. We define

$$h_{ijk}^0 = \frac{(p_{ij})_k^0}{\sum_{l=1}^{K+3}(p_{ij})_k^0}.$$

Then, by Jensen's inequality,

$$\begin{aligned}
\log\hat{x}_{ij} &= \log\sum_{k=1}^{K+3}(p_{ij})_k\\
&= \log\sum_{k=1}^{K+3}h_{ijk}^0\frac{(p_{ij})_k}{h_{ijk}^0}\\
&\geq \sum_{k=1}^{K+3}h_{ijk}^0\log\frac{(p_{ij})_k}{h_{ijk}^0}.
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
J(U, V, a, b, c) &\geq \sum_{ij}\left(x_{ij}\sum_k h_{ijk}^0\log\frac{(p_{ij})_k}{h_{ijk}^0} - \hat{x}_{ij}\right)\\
&= J'(U, V, a, b, c).
\end{aligned}$$

By differentiating J' with respect to u_{ik}, we get

$$u_{ik} = \frac{\sum_j\frac{x_{ij}u_{ik}^0\ v_{kj}^0}{x_{ij}^0}}{\sum_j v_{kj}}.$$

Here, $x_{ij}^0 = \sum_k u_{ik}^0\ v_{kj}^0 + a_i^0 + b_j^0 + c$. Likewise, we obtain the update formula of v_{kj}, a_i, b_j, c:

$$\begin{aligned}
v_{kj} &= \frac{\sum_i\frac{x_{ij}u_{ik}^0\ v_{kj}^0}{x_{ij}^0}}{\sum_i u_{ik}},\\
a_i &= \frac{a_i^0}{M}\sum_j\frac{x_{ij}}{x_{ij}^0},\\
b_j &= \frac{b_j^0}{N}\sum_i\frac{x_{ij}}{x_{ij^0}},\\
c &= \frac{c^0}{NM}\sum_{ij}\frac{x_{ij}}{x_{ij^0}}.
\end{aligned}$$

Finally, we introduce the non-negative subjections and the weight function. Note that the update formula stated above satisfies the non-negative subjections if we set non-negative initial values. The update formula becomes

$$\begin{aligned}
\hat{x}_{ij}^{(t)} &= \sum_k u_{ik}^{(t)}v_{kj}^{(t)} + a_i^{(t)} + b_j^{(t)} + c^{(t)},\\
y_{ij}^{(t)} &= \sum_k u_{ik}^{(t)}v_{kj}^{(t)},\\
u_{ik}^{(t+1)} &= \frac{\sum_j\frac{w_{ij}x_{ij}u_{ik}^{(t)}v_{kj}^{(t)}}{\hat{x}_{ij}^{(t)}}}{\sum_j w_{ij}v_{kj}^{(t)}},\\
v_{kj}^{(t+1)} &= \frac{\sum_j\frac{w_{ij}x_{ij}u_{ik}^{(t+1)}v_{kj}^{(t)}}{\hat{x}_{ij}^{(t)}}}{\sum_i w_{ij}u_{ik}^{(t+1)}},\\
a_i^{(t+1)} &= \frac{a_i^{(t)}}{\sum_j w_{ij}}\sum_j\frac{x_{ij}}{\hat{x}_{ij}^{(t)}},\\
b_j^{(t+1)} &= \frac{b_j^{(t)}}{\sum_i w_{ij}}\sum_i\frac{x_{ij}}{\hat{x}_{ij}^{(t)}},\\
c^{(t+1)} &= \frac{c^{(t)}}{\sum_{i,j} w_{ij}}\sum_{ij}\frac{x_{ij}}{\hat{x}_{ij}^{(t)}}.
\end{aligned}$$

In terms of computational time, our method requires more time for one iteration than NMF. For both of these methods, i.e., our update and the original form of NMF, the dominant computation in terms of order is the matrix multiplication, such as $Y = UV$ appearing in the update formula, which is practically $O(KNM)$. The other element-wise multiplication or comparison requires $O(NM)$ time to compute. The computation of the update of a, b, c requires us to pay the additional cost of $O(15NM)$ to perform the computation in NMF, which is comparable to $O(KNM)$ in our case for small values of K. In such cases, the computation time required for one iteration of our algorithm is longer than that required when the normal NMF method is used. Therefore, we took advantage of the NMF method by using its intermediate result as an initial value for our model. Its effect is examined in the experimental section.

C. Selection of the Number of Clusters

We aim to perform prediction and clustering using the FNMF model. Therefore, we must choose K such that both prediction and clustering are performed optimally. As K corresponds to the number of clusters, its role in terms of clustering is expected to be more important compared to prediction. Here, we specify K such that the obtained clusters appropriately represent the corresponding types of risk.

We choose K by applying certain information criteria to the clusters obtained using FNMF, related to only that part containing information about traffic accidents, as it reflects the pattern of risks in which we are interested. Given the final output of U by the proposed matrix factorization, we regard $l_i = \arg\max_k u_{ik}$ as the cluster of the corresponding row i. The set of rows for each cluster is defined as $C_k = \{i \mid l_i = k\}$. The mean $\hat{\mu}_k$ and covariance $\hat{\Sigma}_k$ related to the accident data in each cluster and the log-likelihood $p(X_{:,1:m_a}, l; K, \hat{\theta})$ can be estimated as follows:

$$\hat{\mu}_k = \frac{\sum_{i\in C_k} x_{i,1:m_a}}{|C_k|},$$

$$\hat{\Sigma}_k = \frac{\sum_{i\in C_k} (x_{i,1:m_a} - \mu_k)^{\mathrm{T}}(x_{i,1:m_a} - \mu_k)}{|C_k|},$$

$$p(X_{:,1:m_a}, l; K, \hat{\theta}) = \prod_i \Bigg(\frac{|C_{l_i}|}{N} \times \left((2\pi)^{m_a} |\hat{\Sigma}_{l_i}|\right)^{-\frac{1}{2}} \times \exp\left(-\frac{1}{2}(x_{i,1:m_a} - \hat{\mu}_{l_i})\hat{\Sigma}_{l_i}^{-1}(x_{i,1:m_a} - \hat{\mu}_{l_i})^{\mathrm{T}}\right)\Bigg).$$

Then, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be computed using the following equations [10]:

$$\mathrm{AIC}(K) = -2\log p(X_{:,1:m_a}, l; K, \hat{\theta}) + m_a(m_a+3)K + K, \tag{6}$$

$$\mathrm{BIC}(K) = -2\log p(X_{:,1:m_a}, l; K, \hat{\theta}) + \frac{m_a(m_a+3)K}{2}\sum_{k=1}^{K}\log|C_k| + K\log N. \tag{7}$$

The smaller the values of these criteria, the better is the clustering in terms of each criterion.

IV. Evaluation of the Proposed Method

In this section, we evaluate our method, which is based on the FNMF model. First, we describe the datasets that we used; then, we describe the approach that we used to integrate them. Second, we show that our scaling and initialization methods are effective in improving the prediction performance. Third, we show that the multiplicative update can find the parameters for our model in an efficient manner, thereby outperforming other methods in terms of prediction.

A. Integration of Datasets

We used the traffic accident dataset, which is collected and published by the Institute of Traffic Accident Research and Data Analysis (ITARDA). The data consist of statistics on traffic accidents at high-risk locations, where accidents occurred most frequently in 2012, and of individual records of accidents that occurred at these high-risk locations in Tokyo. It also provides the number of accidents for each category of accidents, i.e., day or night, number of injured persons, and additional information related to each location. Locations are categorized as either intersections or streets. Approximately 70% of 914 high-risk locations are at intersections. In total, 6201 individual records were collected for 914 high-risk locations.

The dataset containing the traffic flow data is collected and published by the Ministry of Land, Infrastructure, Transportation and Tourism. The statistics include the number and type of vehicles that pass through a particular intersection every hour daily. The roadway data statistics, including the number of lanes, width of lanes, and densities of intersections, are produced by ©Sumitomo Electric Industries, Ltd. These datasets are offered by ©Kokusai Kogyo Co., Ltd. The brake data represent the rate of brake application and are calculated with respect to those locations where sufficient flows are observed. Brake data are collected and offered by Honda Motor Company, Ltd.

All of these datasets were integrated to form one matrix. Each row of the matrix corresponds to a high-risk location listed in the accident dataset. For each high-risk location i, the part $x_{i,1:m_a}$ corresponds to an accident and the other statistics, such as traffic flow and brake rates, correspond to $x_{i,m_a+1:M}$. The values for which the corresponding statistics were missing were treated as missing values.

As statistics besides those related to accidents are always associated with roadways, they cannot be used directly for obtaining $x_{i,m_a+1:M}$ if the location i is an intersection. However, most of the accident records referring to accidents at intersections list the roadway on which the accident occurred. Using these records, we incorporated statistics for the rows corresponding to intersections. More specifically, given an intersection i and a set of accident records that contain roadway information A_i, we formed the statistics $x_{i,m_a+1:M}$ for intersections such that $x_{i,m_a+1:M} = \frac{\sum_{a\in A_i} x(a)}{|A_i|}$, where $x(a) \in \mathbb{R}^{1\times m-m_a}$ is the statistic corresponding to an

TABLE I
COMPARISON OF SCALING RULES

Method	MAE	MAE_most	MAE_pred	MAE_pred_most
No Scaling	10.1958	9.9705	9.9364	9.3541
$\alpha = 1$	**0.4608**	1.4518	0.5205	**1.4372**
$\alpha = 2$	0.4612	**1.4484**	0.5244	1.4820
$\alpha = 3$	0.4613	1.4523	**0.4996**	1.4575
Only accidents	0.4343	1.2595	-	-

accident a. When a statistical record for accidents does not exist in A_i, we treat the part $x_{i,m_a+1:M}$ as missing values. Thus, we obtained 41 categories of accidents ($m_a = 41$) and 291 attributes for the other roadway data from the 42-nd feature to the 332-nd feature, including flow statistics in the range from #89 to #331 and the brake rate as the 332-nd feature. Therefore, overall, a 914 × 332 matrix was formed.

B. Effect of Scaling and Initialization

We compared the prediction performance of our method using the data matrix described in the previous subsection by obtaining the results without scaling and with the scaling rules defined in Section III-A. The results are listed in Table I. Each factorization was performed 50 times and the absolute error in relation to the true value was measured in several ways. First, we calculated the absolute error among all of the accidents for each location and evaluated the average value for each location, denoted by MAE. Furthermore, as it is more important to correctly predict the accident category as being the category of accidents that occurred most frequently at each location, we evaluated the absolute error for the most frequent category at each location and measured the average among all the locations, which we denoted by MAE_most. In addition, we computed the prediction error in the following way: we selected one row and deleted all the information about traffic accidents from the chosen row, and the learned parameters were treated as missing values. We calculated the absolute error between the predicted value and the true value, and we took the average for 50 trials. This procedure was repeated for all the rows. We denoted the values for all the categories and the maximum as MAE_pred and MAE_pred_most, respectively.

We terminated the algorithm when the rate of decrease on the objective value was below 0.05%. We set $a_i = 0$, $b_j = 0, c = 0$ for the initial value and $K = 5$. A value of approximately 9.5 was obtained as the prediction error when no scaling was performed, which is of no realistic use. When the matrix factorization method is applied only to the accident data, the model parameter is estimated so that it models a particular part of the matrix. Compared to this situation, factorizing the entire X balances the error in the part of the accident data and the part of the other data. Thus, in principle, using the entire X is not considered useful for producing smaller error values with respect to the part of accident data compared to using only accident data, given the same parameter space. However, note that we cannot predict missing values by inferring from the knowledge of the other part, unless we incorporate the other data.

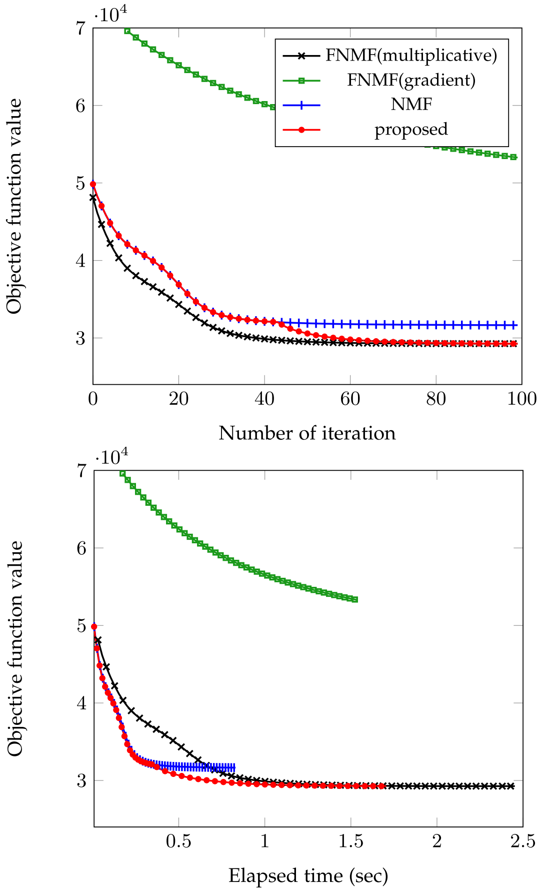


Fig. 4. (upper) Number of iterations vs. Accidents MSE (lower) Time vs. Accidents MSE.

When α is set to a moderate value, the accuracy is slightly worse than the value of "only accidents", but it is of the same order. Thus, we conclude that by introducing α, we can reduce the influence of the part other than the accident part to improve the performance in terms of accuracy. The value of α only affects the accuracy slightly. With respect to accident prediction, $\alpha = 3$ results in the best performance; therefore, we adopted this value for our calculations.

C. Comparison of Prediction Performance

The aim of this section is to demonstrate the efficiency of our update formula. This was done by comparing the proposed algorithm with the existing gradient descent for FMF, the multiplicative update for NMF, and linear regression. In addition, we compared our update without starting from the NMF parameters. The objective function, MAE with respect to accidents, and the prediction error are plotted in Figure 4, 5, and 6,

TABLE II

OBTAINED VALUES OF PERFORMANCE MEASURES FOR RESPECTIVE METHODS

Method	MSE	Accident MAE	Missing Accident MAE
Proposed	2.9417e+004	**0.6665**	**0.4861**
FNMF(Multiplicative)	**2.9404e+004**	0.6671	0.5212
NMF	3.1884e+004	0.6720	0.5020
FNMF(Gradient-based)	4.9262e+004	0.6852	0.5996
Linear Regression	-	-	1.1886

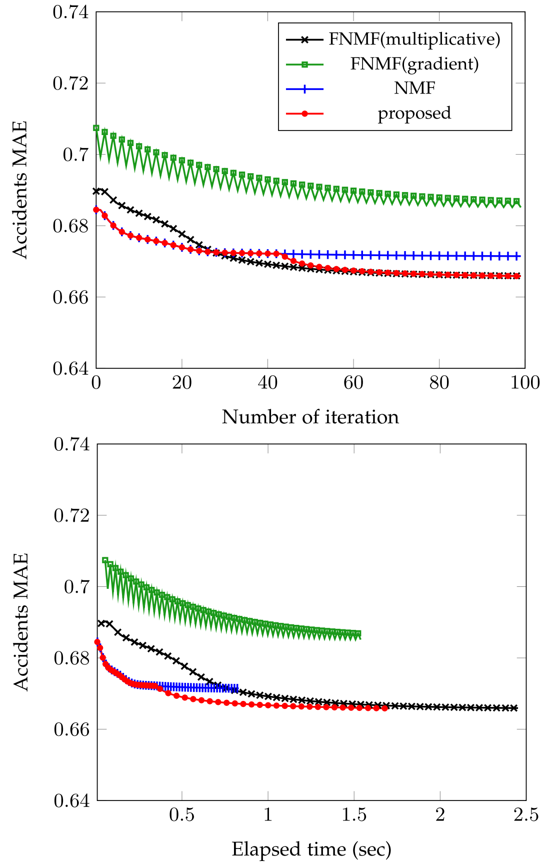


Fig. 5. (upper) Number of iterations vs. Accidents MAE. (lower) Time vs. Accidents MAE.

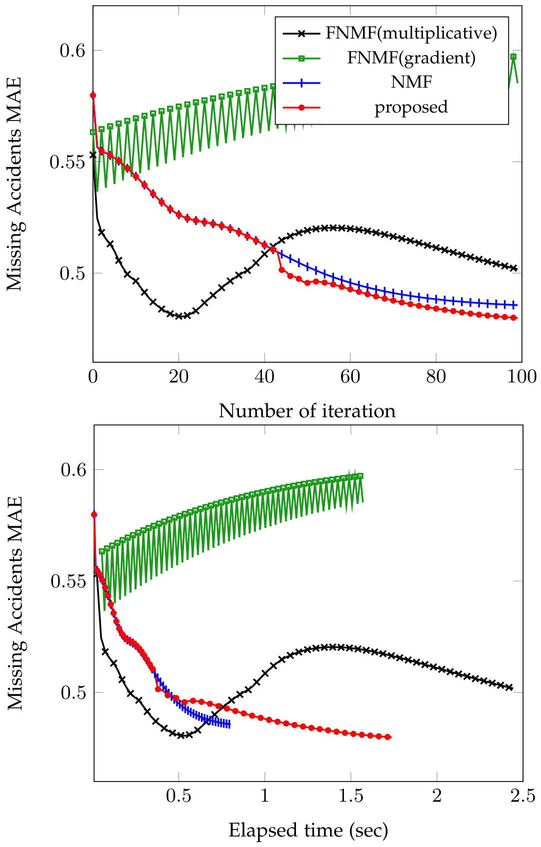


Fig. 6. (upper) Number of iterations vs. Missing Accidents MAE. (lower) Time vs. Missing Accidents MAE.

respectively. All the obtained values are listed in Table II. For the gradient descent, we set the stepsize $\eta = 9.0 \times 10^{-6}$. The calculation was repeated 50 times for each algorithm and the progress of each measure was plotted with respect to a trial that gave the best value for the objective function. For the linear regression, we assumed that each element in the accident data is a dependent variable and the other road and traffic data are predictor variables. Each missing element was replaced by the mean of the observed data in its column. The accident data in one row were predicted by the linear regression model.

The best MAE was achieved by FNMF with multiplicative update, while the proposed method achieved the same level of accuracy with shorter computational time. The performances of these two methods were also comparable in terms of the Accidents MAE value. In terms of the prediction error, the proposed method appears to have the best performance. FNMF with multiplicative update delivered the best prediction accuracy in the middle, but its performance degraded as the updates proceeded. This indicates that learning the model parameter does not necessarily lead to high prediction accuracy. Compared to NMF and linear regression, we can see that the prediction accuracy improves considerably after switching to our model, implying that our model is effective in terms of the prediction accuracy.

D. Selection of the Number of Clusters

The number of clusters was determined by using the proposed method to perform the calculation 50 times for each K in $2 \leq K \leq 10$ and to compute the AIC and BIC defined by equations (6) and (7), respectively. We terminated the algorithm when the rate at which the MSE decreased reached 0.05%. For the initialization, we adopt the method described in Section IV-B. The progress of the average and

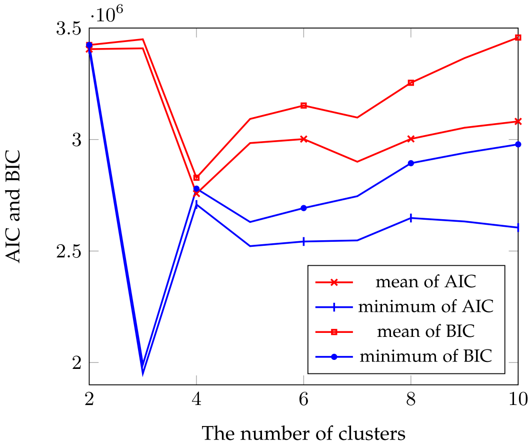


Fig. 7. Minimum and average of AIC and BIC among 50 trials of our method for each value of K.

minimum of AIC and BIC for each value of K is shown in figure 7. In the case of $K = 2$, the value of the minimum and the average for each criterion were nearly the same. In the case of $K = 3$, we obtained the minimum value in the minimum for each criterion while the average value was the largest among all K. In the case of $K = 4$, the values for the minimum and average, which are in the middle between the minimum and average of $K = 3$, are again nearly the same. After a slight decrease in the minimum at $K = 5$, the increase for every K was constant. According to the meaning of the information criteria, we ideally want to compute them with respect to the optimal parameter. Therefore, we adopted the value $K = 3$ that resulted in the smallest value for both AIC and BIC.

V. Cluster Analysis

A. Evaluation of the Features of Each Cluster

In this section, we investigate the characteristics of high-risk locations in each cluster obtained using our framework. We learned FNMF 50 times and chose the result that produced the lowest value for the sum of the squared errors within clusters. We collected the locations with the largest value at $\max_k u_{ik}$ for each cluster k, such that each cluster contains at least 20 locations. The average values of the features for each cluster are shown in Figure 8, by comparing them with the average value of all the rows of X.

TABLE III

Number of Locations and Occupation Rate of Opinions in Each Cluster(%)

Cluster	Number of locations	Narrow	Low Visibility	Speeding	Rushing	Others
#1	28	0.0	0.0	9.4	10.1	7.5
#2	20	0.0	0.0	0.0	0.0	0.0
#3	21	0.0	5.9	4.2	0.0	11.2

Cluster #1, represented by the red line, has a smaller value than the entire average of the traffic flow, which corresponds with the two panels on the right, although for any category of accidents, the number of accidents considerably exceeds the average and other clusters. Thus, we can say that Cluster #1 contains higher-risk locations, which have a low amount of traffic flow and large number of accidents. As it has been considered that the number of accidents is larger when the traffic flow increases in general, Cluster #1 that we obtained implies that the tendency is opposite when we can look at specific accidents in detail. Cluster #2 has large traffic flow values at night time, especially from 7 p.m. to 9 p.m. However, the number of accidents on the upper left panel is lower than the entire average. Cluster #3 is very close to the average in terms of the number of accidents except for accidents involving collisions. On the other hand, Cluster #3 has the largest value among the three clusters representing traffic flows. Locations such as these are considered to be situated on large roads, such as arterial roads and highways. In summary, the characteristics of each cluster can be described as follows

- Cluster #1 (red): relatively small amount of traffic, but large number of accidents with a spike in the number of bicycle accidents.
- Cluster #2 (blue): high density of intersections. Large amount of traffic flow at night time, but a small number of accidents.
- Cluster #3 (green): high traffic flows and number of lanes. Collision accidents are high.

B. Evaluation by External Viewpoints

We used the opinions of pedestrians and drivers, which were collected by Honda Motor Company, Ltd. This information is totally distinct from the model. Hence, it can be regarded as a test of learnability of traffic risk characteristics from an external viewpoint. The opinions were grouped into four categories: the road is narrow, the location on the street has low visibility, there are a large number of speeding vehicles, and there are many pedestrians rushing across the road. We show the ratio of each of these categories of opinions regarding the accidents in each cluster in Table III. We can see that the overall opinion differs for each cluster. Cluster #1 resulted in many opinions about speeding vehicles and rushing pedestrians. This presents high risk in terms of car vs. pedestrian accidents and supports the characteristics determined by our analysis. Cluster #2 did not attract any opinions about possible risks, which is also consistent with our understanding of the cluster. Cluster #3 drew opinions about low visibility, which would be able to lead to collision accidents. This enabled us to conclude that the clusters created by our proposed method successfully captured the characteristics of the high-risk locations, which were highly consistent with the opinions of pedestrians and drivers.

Next, we show the images of locations in Figure 9. These images were captured by Google Street view in Google Maps, and they are images of locations that have the largest and the second-largest values of u_{ik} for each k. The two images on the left correspond to locations in Cluster #1. The upper image shows the location at which most of the accidents occurred in the entire city of Tokyo. This location is at the center of a complex connection of roads with many possible traveling patterns for vehicles, which is consistent with a large number of accidents, even though the amount of traffic is

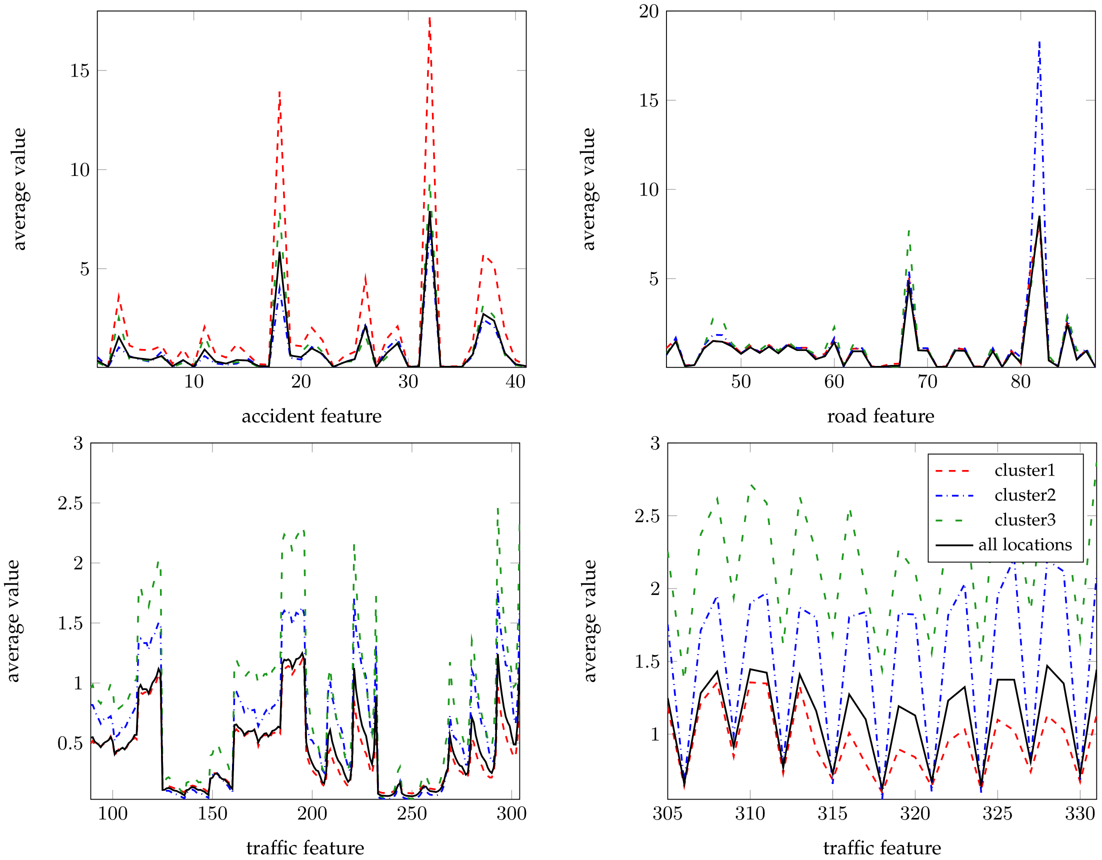


Fig. 8. Tendency of each feature in each cluster. The horizontal axis represents the features and the vertical axis represents the average value of the high-risk locations in each cluster. The upper left panel shows the features from #1 to #41, which correspond to accident statistics. The upper right panel shows the features corresponding to the roadway data, whereas the rest of lower panels show the features corresponding to the traffic flows. The lower left panel presents the flow in each hour. The features from #89 to #196 correspond to the traffic flow from 7 a.m. to 7 p.m. for each type of vehicle for going up and down and the features from #197 to #304 correspond to those from 7 p.m. to 7 a.m. The lower right panel from #305 to #331 shows the total flow in the day time, night time, and throughout the day.

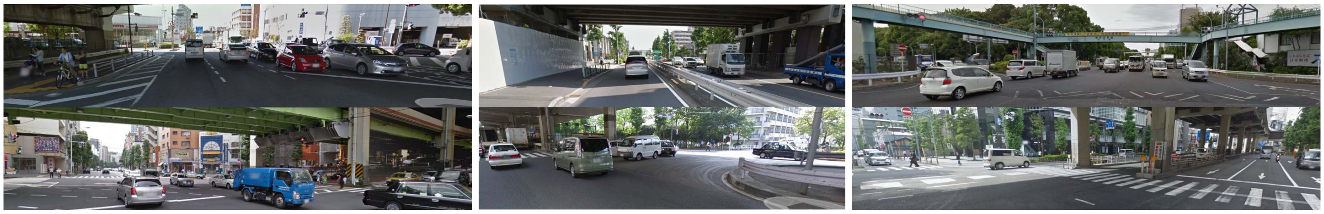


Fig. 9. Images of the high-risk locations that belong to Cluster#1 (left), Cluster#2 (middle), and Cluster #3 (right). The images at the top are those with the largest value of u_{ik}, whereas those at the bottom have the second-largest values.

comparatively low. The images in the middle correspond to locations in Cluster #2. We can see that these roads have guardrails or median strips that would be expected to reduce the number of collision accidents. The images on the right correspond to locations in Cluster #3. By contrast, these roads tend not to have guardrails or median strips. These images also imply that the clusters that were obtained captured important characteristics of traffic risks.

C. Evaluation by Varying the Number of Clusters

This section describes our investigation of the characteristics and transitions between clusters when the number of clusters is increased. We performed our method by setting $K = 5$, at which there is a sub-optimal peak for both AIC and BIC in Figure 7, after which we obtained five clusters. The parameters and condition to belong to each cluster were the same as those in Section V-A.

In Table IV, we show how the high-risk locations in each cluster in the case of $K = 3$ transitioned to the clusters in the case of $K = 5$. It can be seen that the new Cluster #1 and Cluster #3 correspond to the previous Cluster #1 and Cluster #2 in the case of five clusters, respectively, as more than 80% of the locations are now contained by the new corresponding cluster. Apart from this, we can see that Cluster #2 corresponds to Cluster #5. The average values of x_i for each cluster are also shown in Figure 10.

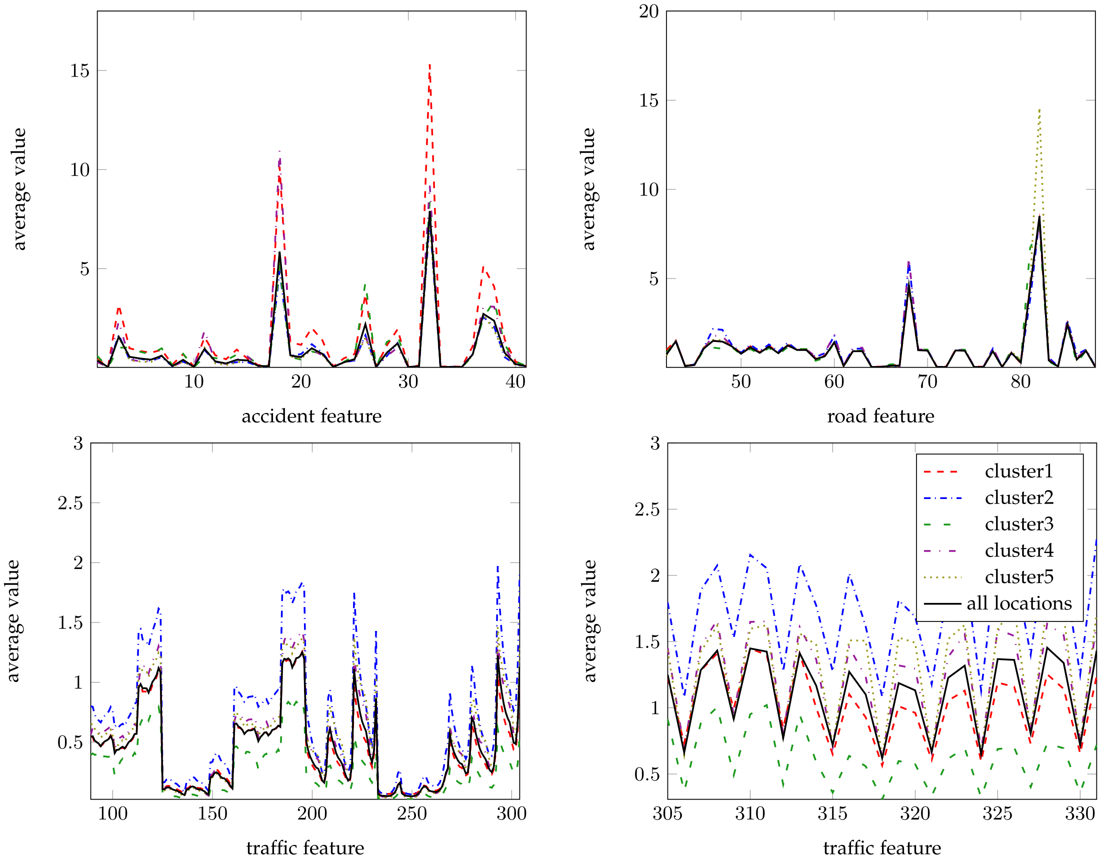


Fig. 10. Tendency of each feature in each cluster in the case of $K = 5$. The horizontal and vertical axes are the same as those in Figure 8.

TABLE IV
TRANSITIONS BETWEEN $K = 3$ AND $K = 5$

	#1	#2	#3	#4	#5	Others	Total
#1	24	0	1	3	0	0	28
#2	0	1	0	0	19	0	20
#3	0	17	0	4	0	0	21
Others	48	192	19	47	65	474	845
Total	72	210	20	54	84	474	914

Cluster #3 and Cluster #4 can be seen as newly produced clusters. Cluster #3 contains the lowest number of high-risk locations. The density of intersections in this cluster is high and the number of accidents when turning a corner is higher than the average. Although only 20 locations belong to this cluster, a large number of social opinions were posted for these locations. Cluster #4 includes some locations that previously belonged to Cluster #1 and Cluster #3 in the case where $K = 3$ and the number of collision accidents is higher than the average. Therefore, this cluster can be considered to contain another group of high-risk locations. Many social opinions regarding speeding vehicles and rushing pedestrians are also posted. In summary, we can say that we observed consistent clusters as before and newly produced clusters, which imply the existence of various risk factors when the number of clusters is increased.

D. Quantitative Evaluation of Cluster Features

This section describes the quantitative evaluation and characterization of each cluster. We extracted a set of conditions, represented as a stochastic decision list, that represents the membership of each cluster. The stochastic decision list [23] is a set of conditions on features X that characterizes a variable of interest $Y \in \{0, 1\}$. Here, features refers to interpretable features, such as accident features, road features, and traffic features, whereas the variable of interest denotes whether a location is assigned to each cluster. According to the minimum description length principle, the model is better when the code length of the model is shorter. Therefore, we define the code length of each stochastic decision list and choose one for each cluster, which gives the shortest code length among a fixed number of candidates.

The code length of the stochastic decision list is represented by the summation of the code length of the data and that of the model. The code length of the model in which Y is correctly selected is short, but the code length of the complicated model is long. Therefore, a simpler and well-explaining model gives a shorter code length overall.

The model of the stochastic decision list is expressed as

$$\begin{aligned}
&\text{If } X_1 \geq a_1 \text{ then } Y = 1 \text{ with probability } \theta_1^{(1)}\\
&\text{Else if } X_2 \geq a_2 \text{ then } Y = 1 \text{ with probability } \theta_1^{(2)}\\
&\qquad\vdots\\
&\text{Else if } X_{K'} \geq a_{K'} \text{ then } Y = 1 \text{ with probability } \theta_1^{(K)}.
\end{aligned}$$

In each condition, not only $\geq$ but also $\leq$ is allowed. Now, considering a set of data, $x_{ij}, y_j \;\; (i = 1, \ldots, K, j = 1, \ldots, N)$, we define the number of data satisfying the model's condition

TABLE V

SELECTED FEATURES FOR CLUSTERS

Cluster	Accident feature	Road feature	Traffic feature
#1	the number of accidents while turning to the right in the night is larger than 0.937	None	None
#2	the number of accidents by passenger car is smaller than the mean	the density of the injunction without a traffic signal is larger than 16.23	None
#3	None	the number of car lanes is larger than 7.02	Day/Night/One-day traffic flow is larger than the minimum

for $Y = 1$ and actually $Y = 1$ as n_{11}, and the number of data satisfying the model's condition for $Y = 1$ but actually $Y = 0$ as n_{10}. n_{00} and n_{01} are defined in the same manner. Then, we define the code length of data as

$$(n_{10}+n_{11})H\left(\frac{n_{11}}{n_{10}+n_{11}}\right)+(n_{00}+n_{01})H\left(\frac{n_{01}}{n_{00}+n_{01}}\right)+\frac{1}{2}\left(\log_2\frac{(n_{10}+n_{11})\pi}{2}+\log_2\frac{(n_{00}+n_{01})\pi}{2}\right),$$

where $H(x) = -x\log_2(x) - (1-x)\log_2(1-x)$ and $H(0) = H(1) = 0$.

Next, we define the code length of the model. The model consists of a set of conditions and each condition includes one feature, one sign of inequality, and one threshold. The code lengths are assigned to each of them in the condition, and the sum of these code lengths is the code length of the model. The variable in each condition can be represented by $\log_2 K'$ bits. The direction of inequality, $\geq$ or $\leq$, can be represented by 1 bit. The code length required to represent the thresholds depends on how the thresholds are set. For instance, if we split the X_i-axis five-fold and restrict a threshold to be one of them, $\log_2 4 = 2$ bits are required to represent the threshold. Finally, we compare the total code length, the sum of code lengths for data and a model, and select the best model.

We have a completed vector $\hat{x}_i$ and cluster c_i for each location i from the result of the matrix factorization clustering. We determined candidates of conditions for each cluster as follows from the mean vector of each cluster. In terms of accident feature, one of all the rows or the sum of rows 1-16 (total number of accidents) are candidates of condition variables. In terms of road feature, columns that are remarkably larger than the mean of all data (such as number of traffic lanes or intersection density) are candidates. In terms of traffic feature, total day time traffic flows, night time traffic flows, or all-day traffic flows, or their combinations are candidates. Thresholds of conditions for each feature are determined as follows. If the minimum of the cluster is larger than the mean of all locations, we set the minimum to be the threshold. If the maximum of the cluster is smaller than the mean of all locations, we set the maximum to be the threshold. Otherwise, the mean of all locations is set to be the threshold.

The final form of the stochastic decision list is represented as combinations of the conditions explained above. For instance, the condition such as "the number of collision accidents is larger than the minimum, the number of road lines is larger than the mean, day time and one-day traffic flows are larger than the mean" is constructed as a candidate. Therefore, the number of models we consider is $(41+1+1)\times 2^A \times 2^3$. Here, A represents the number of candidate features ($A = 0$ in Cluster #1 and $A = 2$ in Cluster #2, #3). We calculate the code length for each model and select the best one.

First, we consider Cluster #1. In Cluster #1, the numbers of accidents are larger than the mean, but no feature of the road data can be found. Traffic flows are smaller than the mean. Then, for $(41 + 1 + 1) \times 2^3 = 344$ conditions that "one of the accident features is larger than the threshold, one of the combinations of three traffic flows is smaller than the thresholds," we compute the code length. In Cluster #2, the number of accidents is small and traffic flow is large. Two features about injunction density are remarkable in the road data. Therefore, for $(41+1+1)\times 2^2\times 2^3 = 1376$ conditions that "one of the accident features is smaller than the threshold, one of the combinations of two features about injunction density is larger than the threshold, and one of the combinations of three traffic flows is larger than the threshold," we compute the code length. In Cluster #3, the number of collision accidents is remarkable and wide roadways and many car lanes can be seen. Traffic flows are as many as in Cluster #2. Therefore, for $(41 + 1 + 1) \times 2^2 \times 2^3 = 1376$ conditions that "one of the accident features is larger than the threshold, one of the combinations of the roadway width and the number of car lanes is larger than the threshold, and one of the combinations of three traffic flows is larger than the threshold," we compute the code length.

In Table V, we show the obtained results. For Cluster #1, the condition "the number of accidents while turning to the right in the night is larger than 0.937" is selected. This condition is satisfied by two locations (in no cluster) except locations in Cluster #1. We can see "the number of accidents while turning to the right in the night" is a typical feature of Cluster #1. On the other hand, the small traffic flows are not selected as distinct features. For Cluster #2, the condition "the number of accidents by passenger car is smaller than the mean and the density of injunctions without a traffic signal is larger than 16.23" is selected. The injunction density is valued but the traffic features are not selected. For Cluster #3, the condition "one of day/night/one-day traffic flows is smaller than the minimum (121543, 64929 and 184110 respectively), and the number of car lanes is larger than 7.02". Features on accidents are not selected, but the number of car lanes and large traffic flows are selected. These conditions appeared to be equivalent in our data, and the combination of these two conditions only yields a larger code length of the model. These results correspond with the features we assess from the mean vector of each cluster qualitatively.

VI. Conclusion

In this paper, we established a novel framework for traffic risk mining that is designed to simultaneously predict the number of accidents and cluster high-risk locations. Our method is based on an algorithm that uses multiplicative updates of a variant of FMF, which is also considered an extension of NMF.

We identified two clusters that represent latent risks and that divide high-risk locations. One of these clusters contained a group of comparatively large roadways that present a high collision risk between vehicles. The other cluster contains a group of locations with low traffic flows but which have a high number of accidents. This implies that we succeeded in extracting locations with comparatively higher risk from among a large number of locations.

From the results of clustering, we could create a ranking of risky locations. Locations could be sorted by the degree to which they belong to the risky cluster, and comparing the number of accidents occurring at the location could facilitate the estimation of the risk. We also established methods of evaluation and characterization of the obtained clusters. In summary, our work can be regarded as the first step toward a new research area of traffic risk mining.

References

[1] Honda Motor Company Ltd. *Safety Map*. Accessed: May 5, 2018 http://safetymap.jp/

[2] Honda Motor Company Ltd. *Honda Sustainability Report*. Accessed: May 5, 2018. [Online]. Available: http://world.honda.com/sustainability/report/

[3] E. Bayam, J. Liebowitz, and W. Agresti, “Older drivers and accidents: A meta analysis and data mining application on traffic accident data,” *Expert Syst. Appl.*, vol. 29, no. 3, pp. 598–629, 2005.

[4] T. Beshah and S. Hill, “Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia,” in *Proc. AAAI Artif. Intell. Develop. (AI-D)*, 2010.

[5] L. Chang and W. Chen, “Data mining of tree-based models to analyze freeway accident frequency,” *J. Safety Res.*, vol. 36, no. 4, pp. 365–375, 2005.

[6] T. Chen, Z. Zheng, Q. Lu, W. Zhang, and Y. Yu. (2011). “Feature-based matrix factorization.” [Online]. Available: https://arxiv.org/abs/1109.2271

[7] M. Chong, A. Abraham, and M. Paprzycki, “Traffic accident analysis using machine learning paradigms,” *Informatica*, vol. 29, no. 1, pp. 89–98, 2005.

[8] I. S. Dhillon and S. Sra, “Generalized nonnegative matrix approximations with Bregman divergences,” in *Proc. Adv. NIPS*, 2005, pp. 283–290.

[9] J. Han, J.-G. Lee, H. Gonzalez, and X. Li, “Mining massive RFID, trajectory, and traffic data sets,” in *Proc. IEEE ICDM Contest, TomTom Traffic Prediction Intell. GPS Navigat.*, 2008, p. 2.

[10] S. Hirai and K. Yamanishi, “Efficient computation of normalized maximum likelihood codes for Gaussian mixture models with its application to clustering,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2011, pp. 1031–1035.

[11] K. Ishiguro and K. Takeuchi, “Extracting essential structure from data,” *NTT Tech. Rev.*, vol. 10, no. 11, pp. 1–6, Nov. 2012.

[12] J. Kim and H. Park, “Sparse nonnegative matrix factorization for clustering,” Georgia Inst. Technol., Atlanta, GA, USA, Tech. Rep. GT-CSE-08-01, 2008.

[13] S. Krishnaveni and M. Hemalantha, “A perspective analysis of traffic accident using data mining techniques,” *Int. J. Comput. Appl.*, vol. 23, no. 7, pp. 40–48, Jun. 2011.

[14] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[15] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Proc. Adv. NIPS*, 2001, pp. 556–562.

[16] X. Li, J. Han, J. Lee, and H. Gonzalez, “Traffic density-based discovery of hot routes in road networks,” in *Advances in Spatial and Temporal Databases. SSTD* (Lecture Notes in Computer Science), vol. 4605. Springer, 2007, pp. 441–459.

[17] K. Moriya, S. Matsushima, and K. Yamanishi, “Traffic risk mining from heterogeneous road statistics,” in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal.*, Oct. 2015, pp. 1–10.

[18] M. Nakano, J. Le Roux, H. Kameoka, T. Nakamura, N. Ono, and S. Sagayama, “Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model,” in *Proc. WASPAA*, 2011, pp. 325–328.

[19] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, “Inferring gas consumption and pollution emission of vehicles throughout a city,” in *Proc. KDD*, 2014, pp. 1027–1036.

[20] H. Shinnou and M. Sasaki, “Refinement of document clustering by using NMF,” in *Proc. PACLIC*, 2007, pp. 430–439.

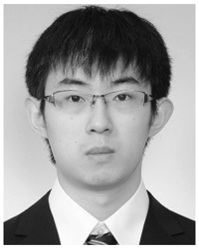
[21] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *Proc. WASPAA*, 2003, pp. 177–180.

[22] W. Xu, X. Liu, and Y. Gong, “Document clustering based on non-negative matrix factorization,” in *Proc. SIGIR*, 2003, pp. 267–273.

[23] K. Yamanishi, “A learning criterion for stochastic rules,” *Mach. Learn.*, vol. 9, no. 2, pp. 165–203, 1992.

[24] S. Zhang, W. Wang, J. Ford, and F. Makedon, “Learning from incomplete ratings using non-negative matrix factorization,” in *Proc. 6th SIAM Conf. Data Mining (SDM)*, 2006, pp. 549–553.

[25] W. J. Lawton and E. A. Sylvestre, “Self modeling curve resolution,” *Technometrics*, vol. 13, no. 3, pp. 617–633, 1971.



Koichi Moriya received the B.E. degree from The University of Tokyo in 2015, where he is currently pursuing the master’s degree with the Department of Mathematical Informatics, Graduate School of Information Science and Technology. His current research interests include succinct data structures, machine learning, and their optimization algorithms.

Shin Matsushima received the Ph.D. degree in information science and technology from The University of Tokyo in 2013. He is currently an Assistant Professor with the Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo. His current research interests include machine learning, data mining, and their optimization algorithms.

Kenji Yamanishi received the M.E. and Dr.Eng. degrees from The University of Tokyo, Japan, in 1987 and 1992, respectively. He was with NEC Corporation from 1987 to 2008. He was a Visiting Scientist with the NEC Research Institute, USA, from 1992 to 1995. Since 2009, he has been leading the Information-Theoretic Machine Learning and Data Mining Group, The University of Tokyo, where he is currently a Professor with the Department of Mathematical Informatics. His interests include information-theoretic machine learning, statistical model selection, and data mining with applications to healthcare and traffic mining.