No Longer Sleeping with a Bomb: A Duet System for Protecting Urban Safety from Dangerous Goods

Jingyuan Wang[†], Chao Chen[†], Junjie Wu[‡*], Zhang Xiong[†]
† School of Computer Science and Engineering, Beihang University, Beijing China
‡ School of Economics and Management, Beihang University, Beijing China.
{jywang, sxyccc, wujj, xiongz}@buaa.edu.cn, *corresponding author

ABSTRACT

Recent years have witnessed the continuous growth of megalopolises worldwide, which makes urban safety a top priority in modern city life. Among various threats, dangerous goods such as gas and hazardous chemicals transported through and around cities have increasingly become the deadly "bomb" we sleep with every day. In both academia and government, tremendous efforts have been dedicated to dealing with dangerous goods transportation (DGT) issues, but further study is still in great need to quantify the problem and explore its intrinsic dynamics in a big data perspective. In this paper, we present a novel system called DGeye, which features a "duet" between DGT trajectory data and human mobility data for risky zones identification. Moreover, DGeye innovatively takes risky patterns as the keystones in DGT management, and builds causality networks among them for pain points identification, attribution and prediction. Experiments on both Beijing and Tianjin cities demonstrate the effectiveness of DGeye. In particular, the report generated by DGeye driven the Beijing government to lay down gas pipelines for the famous Guijie food street.

CCS CONCEPTS

•**Information systems** → **Data mining;** *Geographic information systems;* •**Applied computing** → *Computers in other domains;*

KEYWORDS

Urban Safety, Intelligent Transportation, Pattern Mining, Causal Analysis

1 INTRODUCTION

Nowadays countries and regions all over the world face the challenge of urbanization. The rapid agglomeration of population and industries not only creates monster megalopolises like Beijing, Tokyo and Seoul, but also exposes them to potentially catastrophic risks of various types. For instance, on August 12, 2015, a warehouse storing dangerous goods at the port area of Tianjin exploded, with

Figure 1: Tianjin port explosion on Aug. 12, 2015.

173 people killed and hundreds injured in the blast [1]. In total 304 buildings, 12,428 cars and 7,533 intermodal containers were seriously damaged, and still more surrounding buildings were declared as "structurally unsafe". Local environments and air were also seriously polluted by exploded dangerous goods, which incurs immeasurable loss (see Fig. 1). This painful lesson brings urban safety back to sight as top priority, and indicates the latent but deadly "bomb" we sleep with every day: dangerous goods like gas and hazardous chemicals tanks transported frequently through and around cities.

In the literature, the problem of dangerous goods transportation (DGT) has attracts great attention, with the focuses on transportation route planning [20] and risk analysis [34], both from an operations and optimization view. These studies, though providing constructive managerial insights, usually lack of a micro view of DGT threatens from a big data perspective. Specifically, for a practical application purpose, we first need to determine how to define a dangerous-goods-aware risky zone in a quantitative manner so as to facilitate real-time general monitoring. Also, we should identify the spatio-temporal patterns of DGT and figure out the intrinsic mechanisms behind them for key monitoring and sustainable urban planning. These practical needs indeed motivate our study in this paper, which aims to leverage heterogeneous big data for dealing with DGT issues. Our study can also fall into the research category of urban computing [37], and enrich the dangerous-goods-related studies in this area based on obtained rich DGT trajectory data.

Our main research contributions are summarized as follows. *Firstly*, we present and deploy a novel system called "City Eyes on Dangerous Goods" (DGeye) for real-world DGT risks management (DGTRM). DGeye features a "duet" between DGT trajectory data and human mobile-phone signaling data, and employs a Mahalanobis-distance based measure for risky zones identification. *Secondly*, DGeye innovatively takes risky patterns as the keystones

[1] https://en.wikipedia.org/wiki/2015_Tianjin_explosions

KDD' 17, August 13–17, 2017, Halifax, NS, Canada

DOI: http://dx.doi.org/10.1145/3097983.3097985

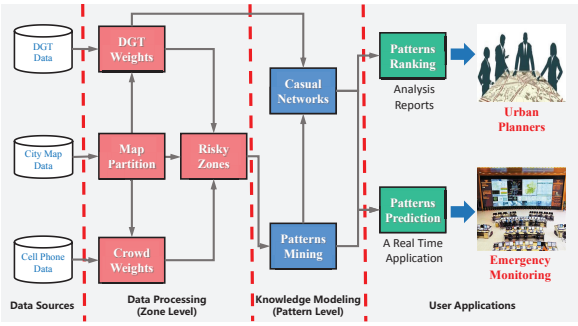


Figure 2: Framework of DGeye.

in DGTRM, and is equipped with an efficient algorithm for maximal patterns mining. A novel trajectory-driven causal network is then built upon these patterns for pattern importance ranking and risks attribution analysis. *Thirdly*, DGeye is capable of risks prediction by adopting an EM-enabled Bayesian network model upon the causal network of risky patterns. Comparative experimental studies with various baselines demonstrate the excellent predictive power of DGeye. *Finally*, DGeye has established itself as a successful deployment in various real-world applications. For instance, as a look back to the Tianjin port explosion disaster, DGeye accurately captures the blast site as one zone inside the first-rank risky pattern. More interestingly, DGeye discloses that the first ranking risky source, which causes 5% downtown areas of Beijing under risk, is the transportation of liquefied gas cylinders to an old famous food street: Guijie. The report generated by DGeye has driven the Beijing government to lay down gas pipelines for Guijie in 2016 [1].

2 THE SYSTEM OVERVIEW

Figure 2 shows the framework of DGeye with four layers. The data source of the system consists of DGT trajectory data, mobile phone signaling data and city map data, which respectively represent the information about dangerous goods, human populations and city geography (see Sect. 3 for details). In the data processing layer, the system partitions a city map into multiple squared zones, and then uses mobile phone and DGT data to calculate the crowd and DGT weights for each zone, respectively. Based on the two kinds of weights, the system detects the urban zones with potential dangerous goods risks in a given time slice of a day as *time-sensitive risky zones* (see Sect. 4 for details).

The knowledge modeling layer is concerned with the pain points of DGT risks within a city. While risky zones are important for real-time monitoring, they are just the "irregular symptoms" of the underlying DGT risks, changing across different time slices and in different days. For an urban management perspective, we would like to unveil the relatively stable patterns, *i.e.*, the pain points, behind the time-variable symptoms, and dig deeply into the causal relations for risk attribution. For this reason, the knowledge modeling layer mainly launches two functions: risky patterns mining and causal network building. The pattern mining function compresses a group of risky zones that are spatially collocated and temporally concurrent into a risky pattern, with the assumption that risky zones in the same pattern might be caused by a same reason (see Sect. 5.1 for details). The causal network building function generates a causal network using risky patterns as vertexes and DGT trajectories passed these patterns as directed edges (see Sect. 5.2 for details). If the destination of DGT trajectories that have passed the pattern p_x is the pattern p_y, we can say the dangerous goods threats in p_x is caused by dangerous goods transportation requirement from p_y. Therefore, we call the network as a causal network.

Based on the causal networks, in the user application layer, we develop two applications for different types of users. For urban planners, the system generates pattern importance rankings for risk causal analysis (see Sect. 6.1 for details). The ranking gives high priority to the patterns that lead to many other patterns of high importance. According to the ranking list, urban planners can fix the pain points gradually from high priority patterns to low priority ones. For the emergency monitoring application, the system gives accurate state predictions for every patterns, which is of great help to official emergency agencies in allocating emergency resources appropriately and proactively (see Sect. 6.2 for details).

3 DATA SOURCES

The data used in DGeye include: mobile phone signaling data, DGT trajectories, and city maps. In what follows, we provide detailed descriptions to the former two types.

Mobile Phone Signaling Data: Mobile operators build base stations all over a city to offer a "full coverage" signaling service to mobile phone users, and the service records between mobile phones and base stations are called "mobile phone signaling data". A record contains <user ID, station ID, user behavior code, time stamp> fields. The user ID and base station ID are unique identifications for cell phones and base stations. The user behavior code field records communication types between a cell phone and a base station. The time stamp records the occurrence time of a communication. The DGeye system uses the location of the base station that provides signaling services to a mobile phone to approximate the position of the phone user. Given the pretty high penetration rate of mobile phones in metropolises, we can use the amount of mobile phone users to approximate the population in an urban zone.

Dangerous Goods Transporter (DGT) Trajectory: Any DGT[2] in China is mandatorily equipped with a GPS terminal and reports real-time locations to the local government, which are then aggregated into DGT trajectory data. A DGT trajectory record contains < vehicle ID, location, speed, timestamp > fields, where the Vehicle ID field is an unique identification of a DGT, the location and speed fields record the real-time location and speed of a transporter, and the timestamp field records the report time of the record. According to the industry standard, the positions of a transporter are reported every 10 seconds.

4 DATA PROCESSING

4.1 Crowd and DGT Weights

Suppose the DGeye system divides a city map into an $I \times J$ urban zone checkerboard, and the zone in the i-th row and the j-th column is denoted as z_{ij}. Assume the data set contains data of M days, and the system divides one day into N time slices. For the n-th time slice

[2]We use DGT to denote both dangerous goods transportation and transporter interchangeably, which can be distinguished with reference to the context.

on the m-th day, we define a DGT weight d_{ij}^{mn} and a crowd weight c_{ij}^{mn} to measure the number of DGTs and the human population at zone z_{ij}, respectively. We also denote the two weights as d_{ij} and c_{ij} or d_x and c_y for concision when there is no ambiguity.

DGT Weight: The DGeye system extracts DGT weight d_{ij} from the DGT trajectories. For a DGT, when its location l_t is reported at time t, we denote $z(l_t)$ as the zone l_t falls in, and process l_t as follows: If $z(l_t)$ and $z(l_{t-1})$ are spatially adjacent or the same, we count a transporter for $z(l_t)$; otherwise, we count a transporter for every zones on the shortest path from $z(l_{t-1})$ to $z(l_t)$. In this way, we re-sample a transporter if it stays at a zone for a long time. This is reasonable because this kind of places are very likely to be the storage places of dangerous goods. At the end of a time slice, the number of passed DGTs of an urban zone is set as the DGT weight of that zone.

Crowd Weight: For each time slice, the system maintains a binary user-zone matrix $\mathbf{U}$, where a row vector corresponds to a cell phone user and a column corresponds to a urban zone. The element $u_{xy} = 1$ indicates user x appears in zone y during the time slice, and 0 otherwise. At the end of a time slice, the crowd weight of any zone y is calculated as

$$c_y = \sum_x \frac{u_{xy}}{\sum_y u_{xy}}. \tag{1}$$

In this way, if a user visits K zones in a time slice, we only count the user $1/K$ times for each zone.

4.2 Risky Zones Detection

In the risky zones detection function, the system calculates a risk score for each urban zone in a time slice. We say an urban zone is at risk because the zone contains a large population *and* too much dangerous goods. Accordingly, using the product of the crowd weight and the DGT weight as the risk score is appropriate, since we could get a high score only when both the two weights are large enough. There is still an obstacle here — the two weights are not in the same order of magnitudes. For example, the population in Beijing is more than 20 millions, but the DGTs in our Beijing data set is only 3,790. To deal with this, we adopt the Mahalanobis distance for weight scaling. The Mahalanobis distance of two vectors $\mathbf{a}$ and $\mathbf{b}$ in a vector set is defined as

$$D_M(\mathbf{a}, \mathbf{b}) = \sqrt{(\mathbf{a} - \mathbf{b})\Sigma^{-1}(\mathbf{a} - \mathbf{b})^\top}, \tag{2}$$

where Σ is the covariance matrix of the vectors set. The risk score of z_{ij} is then defined as

$$RS_{ij} = D_M\left((d_{ij}, 0)^\top, \mathbf{0}\right) \times D_M\left((0, c_{ij})^\top, \mathbf{0}\right). \tag{3}$$

If RS_{ij} is greater than a threshold, we say zone z_{ij} is at a risky state, otherwise at a low-risk state. In practice, the threshold is set to the 90% upper quantile of all the risk scores.

5 KNOWLEDGE MODELING

5.1 Risky Patterns Mining

A *risky pattern* refers to a set of adjacent urban zones that are at the risky state together frequently in a same time slice of a day. For the n-th time slice on day m, we define a risk matrix $\mathbf{R}^{mn} \in \mathbb{R}^{I \times J}$, where the element $r_{ij}^{mn} = 1$ indicates the risky state of z_{ij} and 0 otherwise.

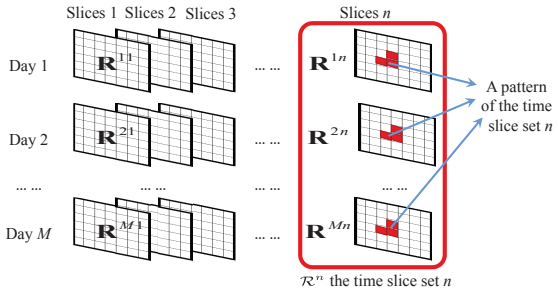


Figure 3: An illustration of risky patterns mining.

Algorithm 1 Risky Patterns Mining from $\mathcal{R}^n$

1: Let the pattern candidate set $L_1 = \{\tilde{p} | \tilde{p} = \{z_{ij}\} \wedge \text{supp}(\tilde{p}, n) \geq$ a threshold, $\forall z_{ij}\}$.
2: Let the pattern set as $P^n = L_1$.
3: Let the pattern candidate size counter $cnt = 2$.
4: **repeat**
5: $L_{cnt} = \emptyset$
6: **for** pattern candidates $\tilde{p}_{cnt-1} \in L_{cnt-1}$ **do**
7: **for** $z_b \in \{z_{ij} | z_{ij}$ is adjacent with $\tilde{p}_{cnt-1}\} \cap \{z_{ij} | z_{ij} \in \tilde{p}'_{cnt-1} \wedge \tilde{p}'_{cnt-1} \in \tilde{P}'\}$ **do**
8: **if** $\text{supp}(\tilde{p}_{cnt-1} \cup z_x, n) \geq$ a threshold **then**
9: $\tilde{p}_{cnt} = \{\tilde{p}_{cnt-1} \cup z_x\}, L_{cnt} = L_{cnt} \cup \tilde{p}_{cnt}$
10: $P^n = P^n \cup \tilde{p}_{cnt}, P^n = P^n - \tilde{p}_{cnt-1}$
11: **end if**
12: **end for**
13: **end for**
14: $cnt = cnt + 1$
15: **until** $\{z_{ij} | z_{ij}$ is adjacent with $\tilde{p}_{cnt-1}\} \cap \{z_{ij} | z_{ij} \in \tilde{p}'_{cnt-1} \wedge \tilde{p}'_{cnt-1} \in \tilde{P}'\} = \emptyset$
16: **return** P^n

We further define a time slice set $\mathcal{R}^n = \{\mathbf{R}^{1n}, \mathbf{R}^{2n}, \ldots, \mathbf{R}^{Mn}\}$, which contains the n-th time slices of all M days. When in W matrices of $\mathcal{R}^n$, a set of adjacent zones X are all at risky state, we define the *daily support* of X *w.r.t.* $\mathcal{R}^n$ as

$$\text{supp}(X, n) = W/M. \tag{4}$$

Based on the concept of daily support $\text{supp}(X, n)$, we give the formal definition of the patterns of the time slice set $\mathcal{R}^n$ as follows.

Definition 1 (Risky Patterns of $\mathcal{R}^n$) *A risky pattern of the time slice set $\mathcal{R}^n$ is a set of zones that satisfies: 1) the zones are spatially adjacent; 2) the daily support of the set for $\mathcal{R}^n$ is larger than a given threshold; 3) the set is not a proper subset of any other risky patterns of $\mathcal{R}^n$.*

Figure 3 gives an illustration of risky patterns. The DGeye system mines risky patterns of $\mathcal{R}^n$ using an Apriori-like algorithm [4], with the pseudo-codes given in Algorithm 1. The algorithm maintains a group of pattern candidate sets $\{L_1, L_2, \ldots, L_{cnt-1}, L_{cnt}, \ldots\}$, where L_{cnt} contains all cnt-size zone sets that satisfy the conditions 1) and 2) of Definition 1. The algorithm also maintains a risky pattern set P^n that is used as a return value of the algorithm. In lines 1-2, the algorithm initializes L_1 using all one-size pattern candidates, and uses L_1 to initialize the risky pattern set P^n. In the 6-13 lines,

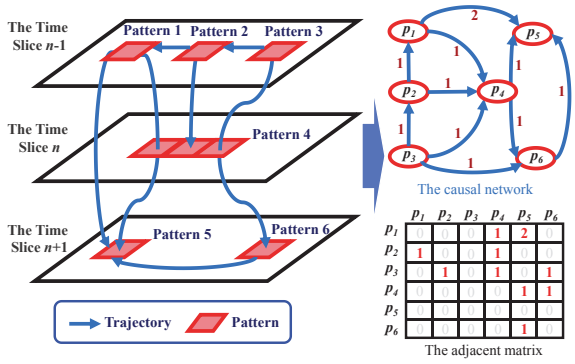


Figure 4: An illustration of causal networks building.

for every pattern candidate $\tilde{p}_{cnt-1}$ in L_{cnt-1}, we enumerate z_x in a zone set, in which all zones are adjacent to $\tilde{p}_{cnt-1}$ to generate a new pattern candidate via $\tilde{p}_{cnt} = \tilde{p}_{cnt-1} \cup \{z_x\}$. In lines 8-11, if the daily support of $\tilde{p}_{cnt}$ is larger than the preset threshold, say 0.8 in our DGeye system, we add $\tilde{p}_{cnt}$ to L_{cnt} and P^n. Because $\tilde{p}_{cnt-1}$ is a proper subset of $\tilde{p}_{cnt}$, according to the condition 3) of Definition 1, we remove $\tilde{p}_{cnt-1}$ from P^n in line 10.

In order to avoid redundant daily support computation, we introduce a "$F_{k-1} \times F_{k-1}$" method [31] to filter z_x in line 7. By defining $\tilde{P}'_{cnt-1}$ as a set of $cnt-1$-size pattern candidates in which all candidates $\tilde{p}'_{cnt-1}$ share $cnt-2$ zones with $\tilde{p}_{cnt-1}$, *i.e.*, $\tilde{P}' = \{\tilde{p}'_{cnt-1} | \tilde{p}'_{cnt-1} \in L_{cnt-1} \wedge |\tilde{p}'_{cnt-1} \cap \tilde{p}_{cnt-1}| = cnt-2\}$, the algorithm is required to select z_x from a pattern candidate in $\tilde{P}'$.

When there is no z_x could be used to increase size of the pattern candidates, the algorithm returns P^n as the risky pattern set of $\mathcal{R}^n$ in line 16. The DGeye system uses Algorithm 1 to mine risky patterns for all time slice sets $\mathcal{R}^n, n = 1, 2, \ldots, N$, and uses their union $P = \bigcup_{n-1}^{N} P^n$ as the final risky pattern set of a city.

5.2 Causal Network Building

Causal network building is another function in the knowledge modeling layer of the DGeye system. In this function, the system uses risky patterns as vertexes and DGT traffics among the patterns as edges to build a weighted directed network. Assuming there are K patterns in the risky pattern set P, the system maintains an adjacent matrix $\mathbf{W} \in \mathbb{R}^{K \times K}$ where the element w_{xy} is the weight for the edge directed from patterns p_x to p_y. When a DGT orderly passes p_x and p_y, we add one to w_{xy}. Note that passing a pattern here means a DGT passes at least one zone in the pattern. If a DGT passes many patterns in sequence, we add an directed edge to any pair of the patterns and count that DGT to all the edge weights. To further illustrate the process of causal network building, we go through a toy example in Figure 4. As shown in the figure, a DGT orderly passes the patterns p_1, p_4, p_5, the weights of the edges $p_1 \rightarrow p_4, p_1 \rightarrow p_5$ and $p_4 \rightarrow p_5$ all should be increased by one.

In the causal network, an edge $p_x \rightarrow p_y$ with a weight greater than one means that some DGTs passes p_x for p_y. In the other words, the reason for p_x being risky is that there are some dangerous goods requirements in p_y. Therefore, we can regard the dangerous goods risk in p_y as a cause of the dangerous goods risk in p_x. That is why we call the network as a *causal network*.

6 USER APPLICATIONS

6.1 Pattern Importance Ranking

The application of the pattern importance ranking is to offer a ranking list of risky patterns based on the causal network for urban safety management. The DGeye system ranks risky patterns following the rule as: a pattern that *i*) causes many patterns and/or *ii*) causes important patterns should have a high importance priority in the ranking list. To this end, we apply a Random Walk with Restart (RWR) model [32] to the causal network to generate ranking scores for risky patterns.

Assume there are K patterns in the causal network. We define a ranking score vector $\mathbf{s} = (s_1, s_2, \ldots, s_k, \ldots s_K)^\top$, where s_k is the score of pattern p_k. Given the weight w_{xy} for the edge from p_x to p_y, we define a causal transition probability from p_x to p_y as an out-degree normalized w_{xy}, *i.e.*,

$$g_{xy} = \frac{w_{xy}}{\sum_{k=1}^{K} w_{xk}}. \tag{5}$$

The system iteratively updates the ranking score vector $\mathbf{s}$ using a transition matrix $\mathbf{G}$ composed of g_{xy}. In the $(t+1)$-th iteration round, $\mathbf{s}$ is updated by

$$\mathbf{s}(t+1) = \alpha\, \mathbf{G} \cdot \mathbf{s}(t) + (1-\alpha)\, \mathbf{q}, \tag{6}$$

where $\mathbf{q} = (q_1, q_2, \ldots, q_k, \ldots, q_K)^\top$ is a pattern-size ratio vector. That is, the k-th element of $\mathbf{q}$ is the normalized size of the risky pattern p_x, *i.e.*,

$$q_x = \frac{size(p_x)}{\sum_{k=1}^{K} size(p_k)}. \tag{7}$$

Note that we also $\mathbf{q}$ to initialize $\mathbf{s}$, *i.e.*, let $\mathbf{s}(0) = \mathbf{q}$.

It is easy to show that the above iterations will converge to the following steady state when $t \rightarrow \infty$ [40],

$$\mathbf{s}^* = (1-\alpha)\mathbf{q}(\mathbf{I} - \alpha\mathbf{G})^{-1}, \tag{8}$$

which is finally adopted to rank the importance of risky patterns. The pattern with a greater score has a higher importance priority. A list of ranked patterns is valuable to urban planners and dangerous goods management department of a city. Based on the list, urban planners could undertake an operable plan to clear these risky patterns progressively from the high priority ones to the low ones.

6.2 Risk State Prediction

Another application offered by the DGeye system is the risk state prediction. Urban emergency departments need to monitor the states of risky patterns before all clearance. An accurate prediction of risky patterns' states could also give a proactive guidance to deploy limited urban emergency resources.

We use Bayesian inference for predictive modeling, which adopts the causal network built in Sect. 5.2 as the Bayesian network. As shown in Fig. 5, the patterns that have causal relations with the pattern to be predicted can be categorized into two types, *i.e.*, patterns with observable states in historical times slices, and patterns with unobservable states in the future. For example, suppose we want to predict the pattern states at the time slice n. The states of risky patterns mined from $\mathcal{R}^{\leq n-1}$ are observable but the ones mined from $\mathcal{R}^n$ are unobservable. For convenience, we denote the states of observable patterns as $H = \{h_1, h_2, \ldots, h_x, \ldots, h_{K1}\}$, and

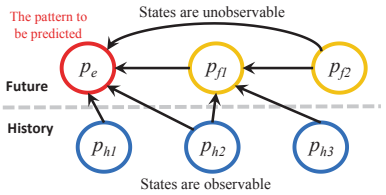


Figure 5: An illustration of the predictive model.

the states of unobservable patterns as $F = \{f_1, f_2, \ldots, f_y, \ldots, f_{K2}\}$. The state of the pattern to be predicted is denoted as e, $e \notin F$.

According to the Bayes' theorem, the posterior probability of e conditioned on H and F is

$$\Pr(e|H, F) = \frac{\Pr(e)\Pr(H, F|e)}{\Pr(H, F)}, \tag{9}$$

which could be approximated using a naïve Bayesian method as

$$\begin{aligned}\Pr(e|H, F) &\propto \Pr(e)\prod_{k=1}^{K1}\Pr(h_k|e)\prod_{k=1}^{K2}\Pr(f_k|e)\\ &\propto \ln(1+\Pr(e)) + \sum_{k=1}^{K1}\ln(1+\Pr(h_k|e)) + \sum_{k=1}^{K2}\ln(1+\Pr(f_k|e)).\end{aligned} \tag{10}$$

However, since the pattern states in F are unobservable, we cannot directly use (10) to calculate the posterior probability of e. Moreover, the impact of edge weights of the causal network are not considered by the posterior probability in (10). To address these, we propose an Expectation-Maximization (EM) algorithm to estimate F and predict e through an iterative way.

Using the causal network as a Bayesian network, the EM algorithm initializes f_x and e using an edge-weighted naïve Bayes model as follows:

$$\begin{aligned}f_x(0) &= \underset{f_x\in\{0,1\}}{\operatorname{argmax}}\ w_{f_x}\ln(1+\Pr(f_x)) + \sum_{k=1}^{K1} w_{(h_k,f_x)}\ln(1+\Pr(h_k|f_x)),\\ e(0) &= \underset{e\in\{0,1\}}{\operatorname{argmax}}\ w_e\ln(1+\Pr(e)) + \sum_{k=1}^{K1} w_{(h_k,e)}\ln(1+\Pr(h_k|e)),\end{aligned} \tag{11}$$

where $w_{(x,y)}$ is the weight of the edge from p_x to p_y in the causal network, w_e and w_{f_x} are the DGT traffic inner the patterns corresponding to e and f_x. In the t-th round of the E-step, we use $e(t-1)$ and $F(t-1)$ estimated in the last round as well as H to update $f_x(t)$ as follows:

$$\begin{aligned}f_x(t) &= \underset{f_x\in\{0,1\}}{\operatorname{argmax}}\ w_{f_x}\ln(1+\Pr(f_x)) + w_{(e,f_x)}\ln(1+\Pr(e(t-1)|f_x))\\ &+ \sum_{k=1}^{K1} w_{(h_k,f_x)}\ln(1+\Pr(h_k|f_x)) + \sum_{k=1}^{K2} w_{(f_k,f_x)}\ln(1+\Pr(f_k(t-1)|f_x)).\end{aligned} \tag{12}$$

In the M-step, we predict e using H and estimated F as

$$\begin{aligned}e(t) = \underset{e\in\{0,1\}}{\operatorname{argmax}}\ & w_e\ln(1+\Pr(e)) + \sum_{k=1}^{K1} w_{(h_k,e)}\ln(1+\Pr(h_k|e))\\ &+ \sum_{k=1}^{K2} w_{(f_k,e)}\ln(1+\Pr(f_k(t)|e)).\end{aligned} \tag{13}$$

When the algorithm reaches a stable state, we use the final $e(t)$ as the state prediction result. We let a pattern state equal to 1 for the risky state, and 0 for the non-risky state. The prior probability and likelihoods are counted from the data set. Compared with traditional Bayesian methods, our predictive model has two advantages: *i*) it exploits the causal dependency among patterns; *ii*) it makes use of causal information in unobservable patterns.

7 EXPERIMENTS AND APPLICATIONS

7.1 Experimental Setup

We apply the DGeye system to two big cities of China: Beijing [3] and Tianjin [4]. Beijing is the capital of China with a 20 million population, and Tianjin is a municipality directed by the central government with a 15 million population. The urban safety of the two cities is of the utmost importance undoubtedly. The data sets used in the experiments were collected from January 1 to March 31 in 2015 for Beijing, and from January 1 to February 31 in 2015 for Tianjin. In the experiments, the system divides one day into 24 time slices, *i.e.*, one hour per slice, and divide the maps of the two cities into $500m \times 500m$ urban zones. The covered area of the two cities contain 80×160 zones, respectively.

7.2 Risky Zones Detection

We here verify the risky zones detection function of the system. Figures 6(a) and 6(b) show the spatial distributions of crowd weights and DGT weights in Beijing at the 10:00 time slice on one day in Jan. 2015, and Fig. 6(c) depicts the distribution of the risky zones in Beijing at the same time slice for comparison. The colors indicate the weights and risky score of each zone — the redder, the higher.

As shown in Figs. 6(a) and 6(b), the population of Beijing are mostly distributed in the downtown area, but high DGT weight zones are mainly distributed on an outer beltway surrounding Beijing, *i.e.*, the 5th ring road [5]. As a result, it is interesting to see from fig. 6(c) that many of high-score risky zones detected by the system are not overlapped with the high DGT weight zones, *e.g.*, the red areas inside the 2nd ring covering an entertainment district of Beijing: Dongzhimen and Dongsi. This indeed illustrates why DGeye considers both human population and DGTs in a "duet" way.

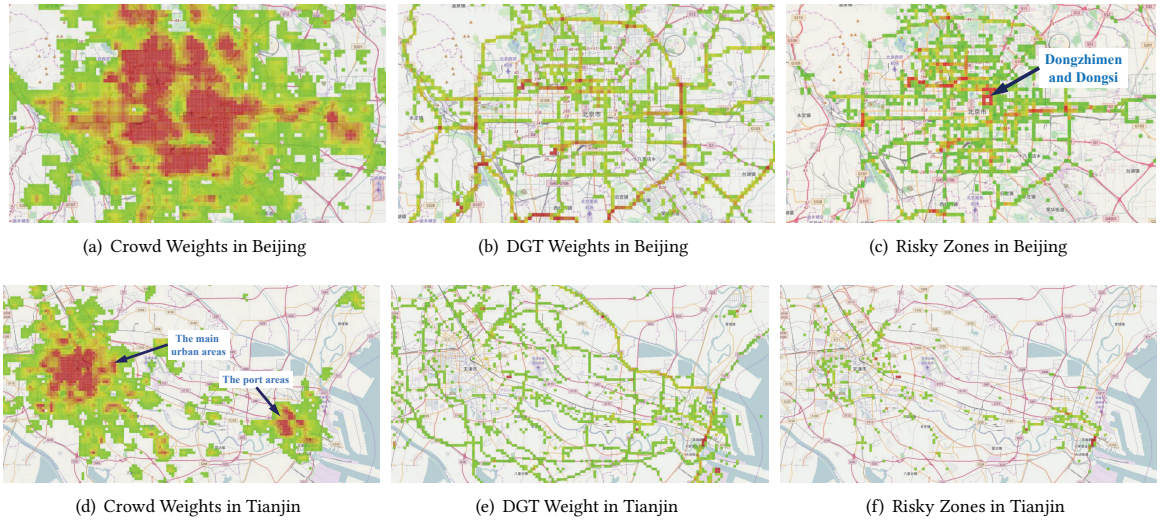
Figures 7(a) to 7(c) exhibit the case of Tianjin at the same time slice as Beijing. As shown in the figures, the population of Tianjin concentrates in two areas, the main urban area and the port area. The DGTs, however, are mainly distributed on beltways and expressways that connect the port area with the main urban area. As to risky zones in Fig. 7(c), again we can find the inconsistency with high DGT weight zones — the high score risky zones are mainly distributed over urban-rural fringe of the main urban area and the downtown of the port area. This is also the result of taking human population into consideration in DGeye.

Figure 8 shows the proportions of risky zones to all urban areas of Beijing and Tianjin for the 24 time slices. Note that all these values are averaged on all days in the data set. It is interesting to see that the temporal distributions of risky zone proportions for the two cities are similar to each other, which indeed coincide more with the rhythms of human activities rather than that of DGTs. Nevertheless, the difference does exist: the emergence of risky zones in the morning peak seems more severe for Beijing.

7.3 Risky Pattern Mining

This subsection demonstrates the risky pattern mining function of the DGeye system. Figure 9 gives temporal distributions of pattern amounts in Beijing and Tianjin, where different colors indicate

[3] https://en.wikipedia.org/wiki/Beijing
[4] https://en.wikipedia.org/wiki/Tianjin
[5] https://en.wikipedia.org/wiki/5th_Ring_Road_(Beijing)



(a) Crowd Weights in Beijing (b) DGT Weights in Beijing (c) Risky Zones in Beijing

(d) Crowd Weights in Tianjin (e) DGT Weight in Tianjin (f) Risky Zones in Tianjin

Figure 6: Spatial distributions of crowd weights, DGT weights, and risky zones in Beijing and Tianjin.

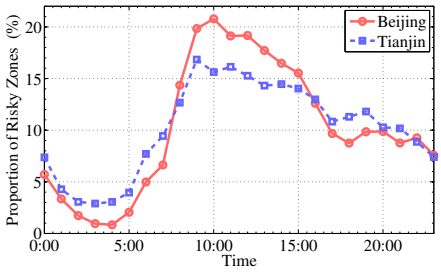
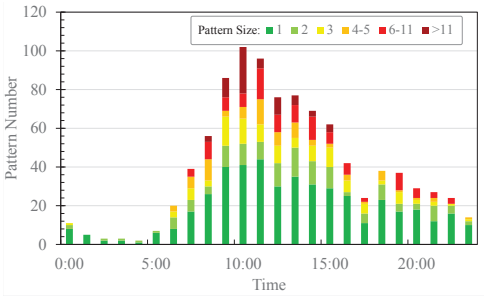
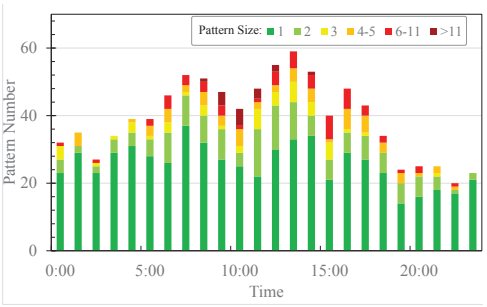


Figure 7: Temporal distributions of risky-zone proportions.



(a) Beijing



(b) Tianjin

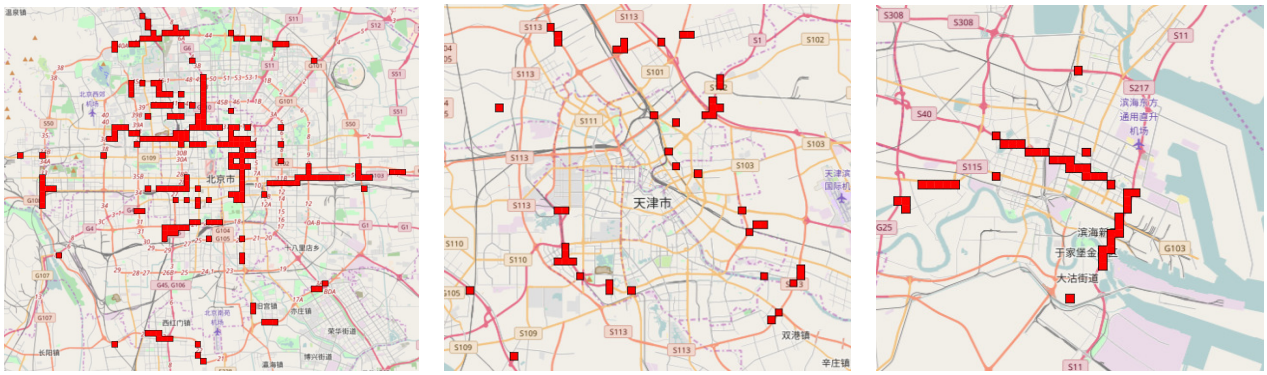
Figure 8: Temporal distributions of risky patterns.

patterns of different sizes. As can be seen, the temporal distributions of patterns in Beijing and Tianjin are very different, which is in sharp contrast to that of risky zones in Fig. 8. This well illustrates why we need risky patterns given risky zones already; that is, patterns indicate relatively stable rules and zones depict instant phenomenon. Two another observations are worth noting. First, the temporal distribution of patterns has an obvious tide in Beijing, but fluctuates much more smoothly in Tianjin. Second, compared with Tianjin, Beijing has obviously more big-size patterns.

The reason for the above differences lies in the diverse requirements of dangerous goods of the two cities. Dangerous goods in Beijing are consumed by citizens in daily life, such as gasoline requirements of gas stations and liquefied gas of restaurants. Therefore, the temporal distribution of risky patterns has a similar rhythm with people's daily life. Moreover, since dangerous goods must be delivered to gas stations and restaurants in downtown area of Beijing every day, DGTs have to drive through many high-population zones in adjacent, which results in many big-size risky patterns in Beijing. Figure 10(a) is a risky patterns' map of Beijing at the 10:00 time slice. As can be seen, many big-size risky patterns are located in the downtown area.

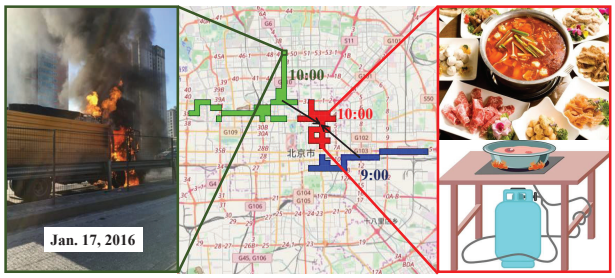
Unlike Beijing, dangerous goods requirement in Tianjin is driven by chemical materials import and export in the Tianjin port. Therefore, the correlation between pattern amount and city life rhythm is very weak. Figures 10(b) and 10(c) show the maps of risky patterns in Tianjin main urban area as well as the port area for the 10:00 time slice. As shown in the figures, most of big-size patterns are in the port area, and the main urban area is relatively safe.

The above differences in pattern distributions suggest different DGT monitoring strategies for Beijing and Tianjin. The Beijing government should pay more attention to DGTs in many areas of the downtown in the middle of a day. Oppositely, the Tianjin government could only monitor some particular areas in the port area but for a whole day.

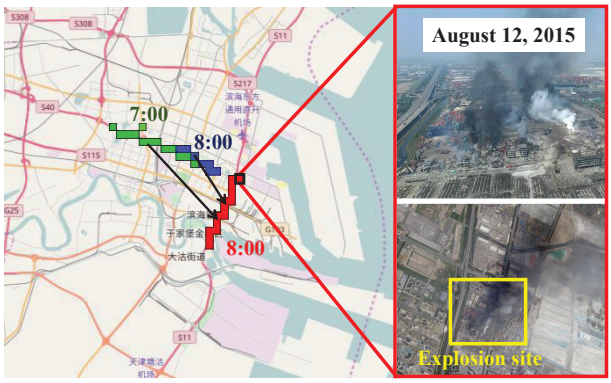


(a) Downtown Area of Beijing (b) Main Urban Area of Tianjin (c) Port Area of Tianjin

Figure 9: Distributions of risky patterns in Beijing and Tianjin.



(a) The Beijing Showcase



(b) The Tianjin Showcase

Figure 10: Applications of pattern importance ranking.

7.4 Importance Ranking

We here give two showcases of the pattern importance ranking application of DGeye. Figure 11 shows the first ranking patterns (the red blocks in the maps) in Beijing and Tianjin, and the green and blue blocks are the first and second ranking patterns, respectively, that are caused by the red patterns.

As shown in Figure 11(a), the first ranking pattern in Beijing is located at the Dongzhimen and Dongsi district, which is a famous entertainment district of Beijing [6]. Especially, the Dongzhimen area has an extremely well known food street, Guijie [7]. A major cuisine offered by restaurants in Guijie is "hot pot" [8], which is a kind of interesting cuisine that cooks raw foods in a simmering metal pot at the center of dining tables. A hot pot table usually equips with a mini gas stove that connects to a liquid gas cylinder, which forms an enormous demand of gas cylinders transported by DGTs to Guijie every day. As shown in Fig. 11(a), the blue and green patterns cover three main roads heading to the red pattern from the north, east and west, respectively. This implies that DGTs transport liquid gas cylinders from suburbs to the red pattern through these three roads, and thus expose the zones in the blue and green patterns to the threaten of explosion. On January 17, 2016, a truck fully loaded with liquid gas cylinders was on fire at the road covered by the green pattern [2]. According to the causality modeled by the causal network, there are 150 risky zones that are directly caused by the red pattern, which cover about 5% downtown areas of Beijing.

As shown in Fig. 11(b), the first ranking pattern in Tianjin is located in the port area, covering a north-south road aside a wharf. The first and second ranking patterns caused by the red pattern cover an east-west road across residential areas of the port. Obviously, the purpose of DGTs driving through the green and blue patterns is going to the wharf aside the red pattern to load or unload dangerous goods. Based on above analysis, we can discover an urban planning defect in the port area of Tianjin: the depots of dangerous goods in the wharf are too close to residential areas, and the government should not let DGTs drive across residential areas. This fatal defect actually has triggered an irreparable tragedy: the Tianjin port blast of dangerous goods on August 12, 2015, which was happened right at the intersection of the roads covered by the red pattern and the green/blue patterns!

7.5 Patten State Prediction

In this subsection, we evaluate the performance of the patten state prediction function in the DGeye system. The benchmarks include: *i*) the Prior model (Pr), which only uses prior probability $\Pr(e)$ to predict pattern states. Because the daily support threshold of risky patterns is set to 0.8, the Prior model always predicts patterns at the risky state. *ii*) The Likelihood model (LL), which uses the likelihood $\sum_{k=1}^{K1} w_{(h_j,e)} \Pr(h_j|e)$ to predict pattern states. The likelihood model is used to evaluate the predictive efficiency of observable patterns.

[6] https://en.wikipedia.org/wiki/Dongzhimen
[7] https://www.travelchinaguide.com/attraction/beijing/guijie-street.htm
[8] https://en.wikipedia.org/wiki/Hot_pot

Table 1: Prediction performance comparison.

Beijing							
	EM	NB	Pr	LL	LR	SVM	ANN
Pr-risky	**0.89**	0.83	0.80	0.84	0.80	0.80	0.79
Re-risky	0.93	0.94	**1.00**	0.66	0.90	0.92	0.88
F1-risky	**0.91**	0.88	0.89	0.74	0.85	0.85	0.83
Pr-all	**0.79**	0.74	0.65	0.72	0.70	0.71	0.70
Re-all	**0.82**	0.79	0.80	0.62	0.75	0.75	0.74
F1-all	**0.80**	0.76	0.72	0.67	0.73	0.73	0.72
Tianjin							
	EM	NB	Pr	LL	LR	SVM	ANN
Pr-risky	**0.93**	0.85	0.76	0.92	0.77	0.76	0.76
Re-risky	0.82	0.87	**1.00**	0.43	**1.00**	**1.00**	**1.00**
F1-risky	**0.87**	0.86	0.86	0.59	**0.87**	0.86	0.86
Pr-all	**0.89**	0.78	0.58	0.78	0.59	0.58	0.58
Re-all	**0.82**	0.79	0.76	0.54	0.77	0.76	0.76
F1-all	**0.85**	0.79	0.66	0.64	0.67	0.66	0.66

iii) The Naïve Bayes model (NB), which uses the initialization model of the EM algorithm in (11) to predict pattern states. *iv*) The Logistic Regression model (LR), Support Vector Machine (SVM) model, and Artificial Neural Networks (ANN) model. The three models use the states of observable patterns as inputs, which are used to evaluate the performance of classical models in our prediction scenarios. The data set of Beijing contains trajectories and mobile phone records of 90 days, and that of Tianjin contains trajectories and mobile phone records of 60 days. We use data of the first 2/3 days as training sets and the remaining 1/3 days as test sets.

The prediction results are listed in Table 1, where precision (Pre), recall (Re) and F1 scores (F1) for the risky state and all states are used as evaluation measures. As can be seen, in both Beijing and Tianjin data sets, the proposed model (EM) achieves the best performances compared with all baselines in terms of all measures except for the recall of the risky state — the prior model always predicts patterns as risky and therefore its risky state recall is 100%. Another observation is the relatively poor performances of LR, SVM and ANN. This might be ascribed to the scarcity of training samples; that is, we can only sample the state of a pattern one time in one day, and hence only have 60 and 40 training samples for every patterns in Beijing and Tianjin, respectively. In this kind of scenario, Bayesian methods seem more effective than completely supervised classification models.

To sum up, we have: *i*) the causality relations in the causal network is very effective for pattern state prediction; *ii*) the information of unobservable states exploited by the proposed EM algorithm could improve the prediction performance; *iii*) the Bayesian method adopted by our model is very suitable to the prediction scenario with small training sets.

7.6 Policy Applications

The Beijing showcase introduced in Sect. 7.4 has been reported to the Beijing government as a report of the system. The Beijing government started a gas pipeline laying project in the Guijie food street of the Dongzhimen and Dongsi district in September 2016. As reported by the news [1], Beijing "Guijie" is about to bid farewell to the gas-cylinder era in 2017.

8 RELATED WORKS

Dangerous goods transportation becomes a very hot topic in hazardous materials management and intelligent transportation system (ITS) areas. In order to control societal risks caused by dangerous goods, some DGT monitoring systems, such as MITRA [26] and GOOD ROUTE [5], are deployed. Most of them focus on monitoring and collecting locations of DGTs only but omit the important human activities for "duet playing". In academia, ITS researchers focus on DGT route planning [20] and transportation systems designing such as rail way DGT [33]. In hazardous materials management, researchers focus on DGT risk definition [9, 28] and analysis [12, 34]. Most of these works study DGT from an operations and optimization view, and have a basic assumption: if a plan is well designed and executed, DGT risks will be under control. In practice, however, many uncertainties could disturb the deployment of plans. Data driven approaches are becoming more desired to detect and analyze risks of dangerous goods in real-world applications.

Spatial pattern mining is a key function of DGeye. Traditional frequent pattern mining algorithms, such as Apriori [4] and FP-growth [14], discover frequent patterns from a transaction set. Many spatial clustering algorithms, such as CLARANS [23], DBSCAN [11], and ST-DBSCAN [7], generate spatial patterns from a spatial distance view [13]. The collocation [17] and spatio-temporal sequential patterns mining [18] algorithms detect frequent collocations and/or concurrences from spatio-temporal data sets. The risky pattern mining method in DGeye is a union of the above algorithms, which mines frequent concurrences of spatial patterns for collocated dangerous goods and populations. Therefore, we adopt a zone-pattern two step mining scheme based on Mahalanobis distance and an Apriori-like algorithm.

Another key function of DGeye is transportation causal analysis. In this area, most studies focus on analyzing causality between transportation and economic indicators from a macro view, such as using the Granger test to analyze causality between transportation and GDP [6] or regional economic growth [19, 22] in different countries and regions [3, 8, 27] . In the micro level, Ref. [21] proposes an outlier tree based causality discovery algorithm for spatio-temporal interactions in urban traffic data, and Ref. [10] proposes a two-step framework for inferring the root cause of anomalies in urban traffic data. Few works have analyzed causality among patterns/events of dangerous goods transportation.

Our study can also fall into the research category of urban computing [37]. In this area, research works related to our study include: data-driven urban analysis [16, 38], urban anomaly detection [24, 25, 39], urban public security [29, 30], citizens behaviors prediction [15, 35, 36], and still many. To our best knowledge, our work is among the earliest studies in urban computing area that try to snuff out the threats from dangerous goods.

9 CONCLUSIONS

In this paper, we present a novel system call DGeye for urban dangerous goods management. DGeye features in leveraging both DGT trajectory data and human activity data for timely risk monitoring as well as proactive risk mitigation. Specifically, DGeye first adopts Mahalalnobis distance for scaling and defines risky zones in a quantitative manner for real-time monitoring. The keystone of

DGeye , however, lies in risky patterns that reveal the rhythms of the risks in a city and are mined by a carefully designed Apriori-like algorithm. A causal network is then built by taking patterns as vertexes and trajectories as directed edges for risk ranking, attribution and prediction, which makes DGeye an ideal decision support system for urban planning and emergency management. DGeye has proven itself in successfully recognizing the hidden explosion risks in Guijie food street of Beijing and in port area of Tianjin. In particular, the report from DGeye has driven the Beijing government to lay down a gas pipeline in Guijie and bid farewell to the gas cylinder transportation history.

ACKNOWLEDGMENTS

Jingyuan Wang, Chao Chen and Zhang Xiong were supported in part by the National Natural Science Foundation of China (Grants 61572059, 61202426, 61332018) and the State's Key Project of Research and Development Plan of China (Grant 2016YFC1000307). Junjie Wu was supported in part by the National Natural Science Foundation of China (Grants 71531001, U1636210, 71490723) and the National High Technology Research and Development Program of China (Grant SS2014AA012303).

REFERENCES

[1] 2016. Beijing GUI Street implementation of the five comprehensive renovation project to enhance the overall environment. http://48h.news/2016/09/20/beijing-gui-street-implementation-of-the-five-comprehensive-renovation-project-to-enhance-the-overall-environment/. (2016).
[2] 2016. Firefighters run on foot for rescue due to traffic jam. http://english.sina.com/china/2016/0117/882483.html. (2016).
[3] Houda Achour and Mounir Belloumi. 2016. Investigating the causal relationship between transport infrastructure, transport energy consumption and economic growth in Tunisia. *Renewable and Sustainable Energy Reviews* 56 (2016), 988–998.
[4] Rakesh Agrawal, Ramakrishnan Srikant, and others. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. 487–499.
[5] I Annex. 2009. Description of Work, the Sixth Framework Programme, Priority 2-IST, Information Society Technologies. *ÆGIS project, CN* 224348 (2009).
[6] Mehmet Aldonat Beyzatlar, Müge Karacal, and Hakan Yetkiner. 2014. Granger-causality between transportation and GDP: A panel data approach. *Transportation Research Part A: Policy and Practice* 63 (2014), 43–55.
[7] Derya Birant and Alp Kut. 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering* 60, 1 (2007), 208–221.
[8] Juan Gabriel Brida, Martín Alberto Rodríguez-Brindis, and Sandra Zapata-Aguirre. 2016. Causality between economic growth and air transport expansion: empirical evidence from Mexico. *World Review of Intermodal Transportation Research* 6, 1 (2016), 1–15.
[9] Roberto Bubbico, Sergio Di Cave, and Barbara Mazzarotta. 2004. Risk analysis for road and rail transport of hazardous materials: a simplified approach. *Journal of Loss Prevention in the Process Industries* 17, 6 (2004), 477–482.
[10] Sanjay Chawla, Yu Zheng, and Jiafeng Hu. 2012. Inferring the root cause in road traffic anomalies. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 141–150.
[11] Martin Ester, Hans Peter Kriegel, Jrg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *International Conference Knowledge Discovery and Data Mining*. 226–231.
[12] B Fabiano, F Curro, AP Reverberi, and R Pastorino. 2005. Dangerous good transportation by road: from risk analysis to emergency planning. *Journal of Loss Prevention in the Process Industries* 18, 4 (2005), 403–413.
[13] Tony H Grubesic, Ran Wei, and Alan T Murray. 2014. Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. *Annals of the Association of American Geographers* 104, 6 (2014), 1134–1156.
[14] Jiawei Han, Jian Pei, and Yiwen Yin. 2000. Mining frequent patterns without candidate generation. In *ACM Sigmod Record*, Vol. 29. ACM, 1–12.
[15] MX Hoang, Y Zheng, and AK Singh. 2016. Forecasting citywide crowd flows based on big data. *ACM SIGSPATIAL* (2016).
[16] Liang Hong, Yu Zheng, Duncan Yung, Jingbo Shang, and Lei Zou. 2015. Detecting urban black holes based on human mobility data. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 35.
[17] Yan Huang, Shashi Shekhar, and Hui Xiong. 2004. Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and Data Engineering* 16, 12 (2004), 1472–1485.
[18] Yan Huang, Liqin Zhang, and Pusheng Zhang. 2008. A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and data engineering* 20, 4 (2008), 433–448.
[19] Michael Iacono and David Levinson. 2016. Mutual causality in road network growth and economic development. *Transport Policy* 45 (2016), 209–217.
[20] Bahar Y Kara and Vedat Verter. 2004. Designing a road network for hazardous materials transportation. *Transportation Science* 38, 2 (2004), 188–196.
[21] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. 2011. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1010–1018.
[22] Kirsi Mukkala and Hannu Tervo. 2013. Air transportation and regional growth: which way does the causality run? *Environment and Planning A* 45, 6 (2013), 1508–1520.
[23] Raymond T. Ng and Jiawei Han. 2002. CLARANS: A method for clustering objects for spatial data mining. *IEEE transactions on knowledge and data engineering* 14, 5 (2002), 1003–1016.
[24] Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. 2013. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 344–353.
[25] Linsey Xiaolin Pang, Sanjay Chawla, Wei Liu, and Yu Zheng. 2011. On mining anomalous patterns in road traffic streams. In *International Conference on Advanced Data Mining and Applications*. Springer, 237–251.
[26] E Planas, E Pastor, F Presutto, and J Tixier. 2008. Results of the MITRA project: Monitoring and intervention for the transportation of dangerous goods. *Journal of hazardous materials* 152, 2 (2008), 516–526.
[27] Rudra P Pradhan and Tapan P Bagchi. 2013. Effect of transportation infrastructure on economic growth in India: the VECM approach. *Research in Transportation Economics* 38, 1 (2013), 139–148.
[28] Grant Purdy. 1993. Risk analysis of the transportation of dangerous goods by road and rail. *Journal of Hazardous materials* 33, 2 (1993), 229–259.
[29] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, Teerayut Horanont, Satoshi Ueyama, and Ryosuke Shibasaki. 2013. Modeling and probabilistic reasoning of population evacuation during large-scale disaster. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1231–1239.
[30] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, and Ryosuke Shibasaki. 2014. Prediction of human emergency behavior and their mobility following large-scale disaster. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 5–14.
[31] Pang-Ning Tan and others. 2006. *Introduction to data mining*. Pearson Education India.
[32] Hanghang Tong, Christos Faloutsos, and Jia Y Pan. 2006. Fast Random Walk with Restart and Its Applications. In *IEEE Sixth International Conference on Data Mining, 2006. ICDM'06*. IEEE Computer Society, 613–622.
[33] Manish Verma. 2011. Railroad transportation of dangerous goods: A conditional exposure approach to minimize transport risk. *Transportation research part C: emerging technologies* 19, 5 (2011), 790–802.
[34] Manish Verma and Vedat Verter. 2007. Railroad transportation of dangerous goods: Population exposure to airborne toxins. *Computers & operations research* 34, 5 (2007), 1287–1303.
[35] Junbo Zhang, Yu Zheng, and Dekang Qi. 2016. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction. *arXiv preprint arXiv:1610.00081* (2016).
[36] Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, and Xiuwen Yi. 2016. DNN-based prediction model for spatio-temporal data. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 92.
[37] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 3 (2014), 38.
[38] Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. 2011. Urban computing with taxicabs. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 89–98.
[39] Yu Zheng, Huichu Zhang, and Yong Yu. 2015. Detecting collective anomalies from multiple spatio-temporal datasets across different domains. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2.
[40] Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, and others. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, Vol. 3. 912–919.