Predicting Abnormal Events in Urban Rail Transit Systems with Multivariate Point Process

Xiaoyun Mo(✉), Mingqian Li, and Mo Li

Nanyang Technological University, Singapore
{xiaoyun001,mingqian001,limo}@ntu.edu.sg

Abstract. Abnormal events in rail systems, including train service delays and disruptions, are pains of the public transit system that have plagued urban cities for many years. The prediction of when and where an abnormal event may occur, can benefit train service providers for taking early actions to mitigate the impact or to eliminate the faults. Prior works rely on rich sources of sensor or log data that require extensive efforts in sensor deployment, data gathering and preparation. In this article, we aim at predicting abnormal events by leveraging only basic information of historical events (*e.g.,* dates, technical causes) that can be easily obtained from existing open records. We propose a non-trivial method which categorizes event pairs based on their basic information, and then characterizes inter-event influence between event pairs via a multivariate Hawkes process. The proposed method overcomes the major hurdle of data sparsity in abnormal events, and retains its efficacy in capturing the underlying dynamics of event sequences. We conduct experiments with a real-world dataset containing Singapore's 5-year abnormal rail events, and compare with a wide range of baseline methods. The results demonstrate the effectiveness of our method.

Keywords: Abnormal event prediction · Multivariate Hawkes process · Data sparsity.

1 Introduction

Mass Rapid Transit (MRT) rail system usually provides the backbone of the public transit system. MRT-related abnormal events including train service delays and disruptions are a crucial problem that has plagued urban cities like Singapore for many years. The occurrence of an abnormal event can impair the journey of thousands to tens of thousands of commuters. The causes of these events vary, but the majority are due to technical faults such as power failures, signal errors, *etc.* On 7th July, 2015, one of the most severe MRT abnormal events in Singapore, which was caused by electrical power trips, crippled two major rail lines in Singapore during evening peak hours and affected up to 413,000 commuters. The operator was fined $5.4 million to take responsibility for this event [22]. Reducing the number of abnormal events, or mitigating their impact on commuters are thus vital tasks for train service providers.

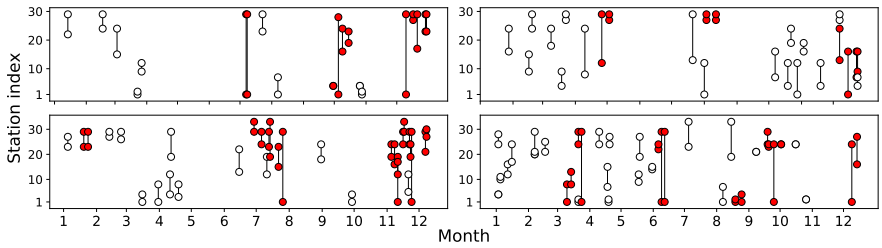

Fig. 1. Distribution of abnormal events from year 2015 (top left) to 2018 (bottom right) on the East-West line of Singapore. Each event is represented by a line segment with two bounding circles, indicating the stretch of abnormal stations. An event is marked in red if it is close to some other event(s) on both temporal (*i.e.,* within 1 week) and spatial (*i.e.,* stretches of stations overlap) scales.

Predictive analysis of MRT abnormal events benefits train service providers and commuters. On one hand, it helps with the predictive maintenance of the MRT system to eliminate hazards proactively, as well as prompt post-event actions to transfer affected commuters. On the other hand, prediction results can enhance public awareness of the operational conditions of the MRT system, and can help them to make decisions about their travel choices. Existing studies related to rail system failures leverage rich sources of data from sensors such as temperature, infrared and strain [8], *etc.*, which are practically hard to execute due to the costly deployment of sensors; as well as data from logs such as maintenance logs, equipment details [11], *etc.*, comprising heterogeneous data sources that require extensive efforts in data gathering, storing and pre-processing.

This paper aims at predicting when (*i.e.,* date) and where (*i.e.,* rail line and stations) will a future abnormal event occur. We leverage historical event data with event attributes including date, the abnormal rail line and stations, as well as the type of technical fault that causes the event. The data are easily accessible from two public channels, namely, official tweets posted by Singapore MRT operators (*i.e.,* SMRT and SBS) and local news feeds (*e.g.,* The Straits Times). We collect data about the abnormal events from January 2015 to December 2019 and perform the study. Fig. 1 shows the spatial-temporal distribution of abnormal events on the East-West line, one of the most popular MRT lines in Singapore. The figure suggests certain locality on both temporal and spatial scales when events take place. It is likely that after one event occurs other events of overlapping stretches may follow, and as a result the sequence of events display a clustered dynamic pattern. This paper makes use of such a pattern, *i.e.,* the excitation influence between events, to model event sequence and therefore forecast future events.

Executing this approach, however, entails special challenges due to a major issue of data sparsity, *i.e.,* the number of abnormal events is extremely limited to capture the sophisticated inter-event influences. Specificaly, the influence decays as the interval between two events' timings increases, and events of different technical causes may be distinctive in the pattern of triggering future events.

Other factors like whether the two events are on the same rail line, or whether their stretches of abnormal stations overlap, may lead to a difference on the magnitude of influence between them. In addition, it is also necessary to quantify by how much an event occurs innately, *i.e.,* occurs as natural arrivals. We need learn from very limited historical events as the training set in order to derive a unified model to numerically quantify the dependencies.

Contributions. This paper proposes a novel method based on multivariate Hawkes process, PAbEve (**P**redicting **Ab**normal **Eve**nt in MRT system), to address the above challenges in predicting MRT abnormal events. Leveraging the information of historical events, which is lightweight and publicly accessible, PAbEve retains its efficacy in modeling the abnormal event sequence, including the timings and locations, and then utilize it to predict the timings and locations of future events. PAbEve captures non-trivial inter-event influences and its parameters are expressive for those influences. We conduct extensive experiments with a real-world dataset containing Singapore's 5-year MRT abnormal events, and evaluate PAbEve in comparison with a wide range of alternative approaches. The results suggest PAbEve outperforms other methods in overall performance.

2 Related Work

Abnormal Event prediction. Predicting abnormal events has attracted extensive attention in recent years. We divide existing works into three categories according to the object being studied. The first category studies on time series of instances in equal-length time steps, and treats those of extreme values as abnormal instances, such as key timings of flu seasons [1], congestion in traffic streams [9], and financial crisis in stock price series [4]. The second category of works attempts to construct indicative features or to find precursors of abnormal events, and use them as predictors for future events. For instances, some works conduct predictive analysis on a rich set of sensor, logging (*e.g.,* maintenance logs) and/or contextual data (*e.g.,* weather), to construct meaningful features for the prediction of railway point failures [11], railway service interruptions [8] and medical equipment failures [21]. Some studies focus on mining media articles (*e.g.,* tweets) to find precursors of social events like protests [3,14,23]. The third category of works directly studies event occurrences and utilize the inter-event correlation to predict future events, such as crime, vehicle collision, *etc* [16]. Generally, the number of abnormal events is limited, resulting in a major challenge of data sparsity. Our work falls into this category. Only a few prior works of abnormal event prediction address the issue of data sparsity [1,15,23]. However, those approaches cannot be applied to our case, as their prediction problems fall into the first/second categories. We are unable to conduct contextual analysis of rich data sources because other relevant information for MRT abnormal events is also limited. Relevant information like the technical cause of event will be instead used as auxiliary covariates in this paper.

Point Process. As a mathematical approach for modeling event sequences, point process has been widely adopted to deal with prediction tasks, such as

the prediction of taxi pickup, crime, neuronal activity, *etc.* Generally, a point process is characterized by a conditional intensity function. Classical point processes, such as Hawkes process [7], formulate the conditional intensity function based on a strong assumption on the dynamic pattern of event sequences. In the past decades, many non-trivial models extend these classical models to 3D spatio-temporal or multivariate space [2,12,19,24]. Recent deep learning techniques incubate state-of-the-art point processes, which are usually able to embed long-term memory of historical events and make very few assumptions on the dynamic pattern of event sequences [5,13,17,25]. Some existing works also propose intensity-free models to develop more general point processes using frameworks such as adversarial learning [10,20].

3 Preliminaries

Temporal point process. A temporal point process is a random process of event occurrence characterized by a *conditional intensity function*, $\lambda(t|\mathcal{H}_t)$, which is the event rate at time $t \in \mathbb{R}$ conditioned on historical events $\mathcal{H}_t$ before t. For convenience, we omit the notation $\mathcal{H}_t$ in the rest of the paper. The functional form of $\lambda(t)$ is usually designed according to the dynamic pattern of event sequences. For example, Hawkes process is a kind of temporal point process that characterizes the self-exciting dynamic pattern, *i.e.,* the occurrence of an event can raise the event rate in the near future.

Multivariate Hawkes process. A multivariate Hawkes process can be regarded as a sequence of correlated Hawkes processes of multiple event types. Formally, for a U-dimensional Hawkes process, the conditional intensity function of the u-th event type, $\lambda_u(t)$, $u = 1, ..U$, is defined as

$$\lambda_u(t) = \mu_u + \sum_{i:t_i<t} \alpha_{uu_i} g(t - t_i) \tag{1}$$

where μ_u is the natural arrival rate (*i.e.,* background rate) of the u-th event type, α_{uu_i} is the trigger coefficient between the u-th and u_i-th event types, and $g(\Delta t)$ is the trigger function that usually decays with the increase of Δt.

4 Methodology

4.1 Problem Definition

In the MRT system, suppose there are R rail lines, and U possible technical faults that cause abnormal events. We use r $(r = 1, ..., R)$ and u $(u = 1, ..., U)$ to denote the indices of rail lines and technical faults, respectively. Each rail line is divided into M equal-size segments so that rail lines are of equal numbers of segments (*i.e.,* "lengths"). We then denote a stretch of stations as $x = [x_-, x_+]$, where x_- and x_+ are the indices of the two bounding segments on a specific rail line, $1 \leq x_- \leq x_+ \leq M$. The list of distinct stretches of the r-th rail line is denoted by X^r, with size $S = |X^r| = \frac{M(M+1)}{2}$ that is identical for all rail lines.

We use s to denote the index of a stretch ($s = 1, ..., S$). Suppose we are given an abnormal event sequence $e_1, ..., e_n$, where $e_i = (t_i, r_i, s_i)$, with $t_i \in \mathbb{Z}$ the time of event in terms of day. The causes of events are denoted by $u_1, ..., u_n$, each of which is the index of technical fault. Given the information of n historical events above, we aim to predict the time, abnormal rail line and stretch of stations of the next event, $e_{n+1} = (t_{n+1}, r_{n+1}, s_{n+1})$.

4.2 Categorization of Event Pairs

We categorize an event pair (e_i, e_j), where $i > j$, hierarchically using the locations and technical faults of both events. The categorization is of three levels. For the first level, we divide event pairs based on their technical faults. For simplicity, we assume that inter-event influence only exists between two events that are caused by the same type of technical fault. According to the official tweets posted by MRT operators, there are 6 main types of technical faults, namely, train fault, track fault, power fault, signal fault, platform fault (mostly the screen door errors), and others. Therefore, for the first level, event pairs are divided into 6 groups of different fault types. For the second level, we distinguish *intra-line* pairs, for which two events occur on the same rail line (*i.e.,* $r_i = r_j$), from *inter-line* pairs, for which two events occur on different rail lines (*i.e.,* $r_i \neq r_j$). The influence between inter-line pair of events is possible as the two rail lines can be run by the same transit operator. For the third level, we further divide event pairs into *overlapping* pairs or *non-overlapping* pairs, according to whether the two events' stretches of abnormal stations overlap (*i.e.,* $X_{s_i}^{r_i} \cap X_{s_j}^{r_j} \neq \emptyset$) or not (*i.e.,* $X_{s_i}^{r_i} \cap X_{s_j}^{r_j} = \emptyset$). Note that the stretches of an inter-line pair can also overlap via interchange stations.

4.3 Multivariate Hawkes Process

We propose a multivariate Hawkes process that can capture the specific inter-event influence of each category of event pairs. We first derive the conditional intensity function, and then present the procedure of prediction.

Conditional intensity function. The occurrence rate of abnormal events on the r-th rail line at the s-th stretch of stations in X^r, is specified by a conditional intensity function defined as

$$\lambda(t, r, s) = \sum_{u=1}^{U} \lambda_u(t, r, s) \tag{2}$$

$$\lambda_u(t, r, s) = \mu_{urs} + \sum_{j:t_j<t, u_j=u} \phi_{u_j}(r, r_j, s, s_j) g(t - t_j) \tag{3}$$

where $\lambda_u(t, r, s)$ is a subordinate conditional intensity function of technical fault u, with $u = 1, ..., U$. μ_{urs} represents the natural arrival rate of events of the type indicated by the subscript indices. The second term of $\lambda_u(t, r, s)$ represents the trigger rates that are brought by events before t. According to the 3-level

categorization described in Section 4.2, we assign each category of event pairs with a distinct trigger coefficient specified by $\phi_u(\cdot)$, and it is defined as

$$\phi_u(r, r', s, s') = \begin{cases} a_u & r \neq r' \; and \; X_s^r \cap X_{s'}^{r'} = \emptyset \\ b_u & r \neq r' \; and \; X_s^r \cap X_{s'}^{r'} \neq \emptyset \\ c_u & r = r' \; and \; X_s^r \cap X_{s'}^{r'} = \emptyset \\ d_u & r = r' \; and \; X_s^r \cap X_{s'}^{r'} \neq \emptyset \end{cases} \tag{4}$$

in which a_u, b_u, c_u and d_u are the trigger coefficients for the u-th fault type, for inter-line non-overlapping, inter-line overlapping, intra-line non-overlapping and intra-line overlapping event pairs, respectively. The trigger function $g(\Delta t)$ is defined in order to weaken the influence as time elapses. Particularly, it is defined in a non-parametric way, *i.e.,* Δt is discretized as $\Delta t = k\delta t$, for $k = 0, ..., K$. The hyper-parameters K and δt control the span and granularity of time intervals, respectively. Then the trigger function $g(\Delta t)$ is specified by a sequence of scalars $[g_k]_{k=1}^K$. When $\Delta t > K\delta t$, $g(\Delta t)$ equals to zero.

Parameter learning. The parameters are optimized iteratively using the maximum likelihood estimation. The likelihood of event e_i is defined as

$$L_i = \lambda_{u_i}(t_i, r_i, s_i) \cdot exp\left\{ - \int_{t_{i-1}}^{t_i} \left(\sum_{u=1}^{U} \sum_{r=1}^{R} \sum_{s=1}^{S} \lambda_u(\tau, r, s) \right) d\tau \right\} \tag{5}$$

where the exponential term of Eq. (5) means no event during the time interval (t_{i-1}, t_i) [18]. A lower-bound of the log-likelihood $logL$ of an n-length event sequence is then derived as

$$\begin{aligned} logL \geq & \sum_{i=1}^{n} \left(p_{ii} log \frac{\mu_{u_i r_i s_i}}{p_{ii}} + \sum_{j:t_j<t_i, u_j=u_i} p_{ij} log \frac{\phi_{u_j}(r_i, r_j, s_i, s_j) g(t_i - t_j)}{p_{ij}} \right) \\ & - (t_n - t_0) \sum_{u=1}^{U} \sum_{r=1}^{R} \sum_{s=1}^{S} \mu_{urs} - \sum_{r=1}^{R} \sum_{s=1}^{S} \sum_{j=1}^{n} \left(\phi_{u_j}(r, r_j, s, s_j) \int_{t_j}^{t_n} g(\tau - t_j) d\tau \right) \end{aligned} \tag{6}$$

based on Jensen's inequality ($log(E[X]) \geq E[log(X)]$). The weights p_{ii} and p_{ij}, $j = 1, ..., i-1$ are computed following [24].

$$p_{ii} = \frac{\mu_{u_i r_i s_i}}{\lambda_{u_i}(t_i, r_i, s_i)}, \qquad p_{ij} = \frac{\phi_{u_j}(r_i, r_j, s_i, s_j) g(t_i - t_j)}{\lambda_{u_i}(t_i, r_i, s_i)} \tag{7}$$

Specifically, p_{ii} denotes the probability that event e_i arrives naturally, while p_{ij} denotes the probability that it is rather triggered by a previous event e_j.

Given the lower-bound of $logL$, the analytical solutions of the parameters can be obtained by setting the first derivative of the lower-bound with respect to each parameter to zero and then solving the equations. Specifically, the solutions of μ_{urs}, a_u (and adapted to b_u, c_u and d_u according to Eq. (4)) and g_k are depicted in Eq. (8), (9) and (10), respectively, where $\mathbb{I}[\cdot]$ is the indicator function.

$$\mu_{urs} = \frac{\sum_{i=1}^{n} p_{ii} \mathbb{I}[u_i = u, r_i = r, s_i = s]}{t_n - t_0} \tag{8}$$

Algorithm 1: Iterative algorithm for parameter learning

Input: events $[e_i]_{i=1}^n$, faults $[u_i]_{i=1}^n$, randomly initialized weights p_{ii}'s and p_{ij}'s
Output: values of parameters

repeat
  Update μ_{urs} by Eq. (8) for $u = 1, ..., U$, $r = 1, ..., R$, and $s = 1, ..., S$;
  Update a_u by Eq. (9), similarly for b_u, c_u and d_u, for $u = 1, ..., U$;
  Update g_k by Eq. (10) for $k = 0, ..., K$;
  for $i = 1, ..., n$ *and* $j = 1, ..., i-1$ **do**
    Update p_{ii} by Eq. (7);
    Update p_{ij} by Eq. (7) **if** $t_i - t_j \leq K\delta t$ and $u_i == u_j$ **else** set $p_{ij} = 0$.
until p_{ii}*'s and* p_{ij}*'s converge*;

$$a_u = \frac{\sum_{i=1}^{n} \sum_{j:t_j<t_i, u_j=u_i} p_{ij} \mathbb{I}[u_j = u, r_i \neq r_j, X_{s_i}^{r_i} \cap X_{s_j}^{r_j} = \emptyset]}{\sum_{r=1}^{R} \sum_{s=1}^{S} \sum_{j=1}^{n} \mathbb{I}[u_j = u, r \neq r_j, X_s^r \cap X_{s_j}^{r_j} = \emptyset] \int_{t_j}^{t_n} g(\tau - t_j) d\tau} \tag{9}$$

$$g_k = \frac{\sum_{i=1}^{n} \sum_{j:t_j<t_i, u_j=u_i} p_{ij} \mathbb{I}[k\delta t \leq t_i - t_j < (k+1)\delta t]}{\delta t \sum_{r=1}^{R} \sum_{s=1}^{S} \sum_{j=1}^{n} \phi_{u_j}(r, r_j, s, s_j) \mathbb{I}[k\delta t \leq t_n - t_j]} \tag{10}$$

Provided n historical abnormal events, we first initialize the weights p_{ii}'s and p_{ij}'s by random. After that, a loop is used to iteratively optimize the values of all parameters and weights until convergence, *i.e.,* until the values of parameters (or weights p_{ii}'s and p_{ij}'s) do not change substantially in a single iteration. The iterative algorithm is shown in Algorithm 1.

Prediction. Given the conditional intensity functions and historical abnormal events, the probability density function of some $t \in (t_n, +\infty), r, s$ being the time, rail line, and stretch of the next event is given as

$$f(t, r, s) = \lambda(t, r, s) \cdot exp\left\{ - \int_{t_n}^{t} \lambda(\tau) d\tau \right\} \tag{11}$$

We predict the timing of event e_{n+1} by taking its expectation as

$$\hat{t}_{n+1} = E[t|\mathcal{H}_{t_n}] = \frac{\int_{t_n}^{t_n+T} t \cdot \left(\sum_{r=1}^{R} \sum_{s=1}^{S} f(t, r, s) \right) dt}{\int_{t_n}^{t_n+T} \left(\sum_{r=1}^{R} \sum_{s=1}^{S} f(t, r, s) \right) dt} \tag{12}$$

where T is a sufficiently large time duration (*e.g.,* $T = 150$). After that, the abnormal rail line as well as the stretch of stations are predicted as follows:

$$\hat{r}_{n+1} = \underset{r}{\operatorname{argmax}} f(\hat{t}_{n+1}, r) = \underset{r}{\operatorname{argmax}} \sum_{s=1}^{S} f(\hat{t}_{n+1}, r, s) \tag{13}$$

$$\hat{x}_{n+1} = E[x|\mathcal{H}_{t_n}, \hat{t}_{n+1}, \hat{r}_{n+1}] = \frac{\sum_{s=1}^{S} X_s^{\hat{r}_{n+1}} \cdot f(\hat{t}_{n+1}, \hat{r}_{n+1}, s)}{\sum_{s=1}^{S} f(\hat{t}_{n+1}, \hat{r}_{n+1}, s)} \tag{14}$$

in which $\hat{x}_{n+1} = [\hat{x}_-, \hat{x}_+]$, with $\hat{x}_-$ and $\hat{x}_+$ the indices of the two bounding segments that specify a stretch of stations on the $\hat{r}_{n+1}$-th rail line. All the integrals in the equations above are approximated using summation.

Table 1. Distribution of events by fault.

Fault type	train	track	power	signal	platform	others
Num. of events	86	81	19	27	22	19

5 Experiments

5.1 Experimental Setup

Dataset. We collect MRT abnormal events from January 2015 to December 2019, from two open sources, *i.e.,* official tweets posted by operators and local news feeds[1]. The provided information includes date, approximate time of the day, rail line, cause, and the stretch of affected stations. We set the causes of a few events with no cause specified as the "others" type of technical fault. After filtering out isolated incidents with system irrelevant causes (*e.g.,* passenger's fall, animal invasion), we finally obtain 254 events, the distribution of which by fault type is shown in Table 1. We sort the events by date and use the first 75% for training and the last 25% for testing. There is no validation set due to the scarcity of observed events. But we provide sensitivity analysis which shows clear trends of the impact of hyper-parameters on the performance.

Baselines. We compare PAbEve with 9 baseline methods, where 5 are for timing prediction only (*i.e.,* NextDay, Auto-regressive, Hawkes parametric, Hawkes non-parametric and NNPP), and the rest 4 for both timing and location prediction (*i.e.,* Poisson loc, MMEL loc, MMEL fault+loc and RMTPP loc).

- NextDay: a naive baseline which uses the next day of the most recent event as the prediction result, *e.g.,* $\hat{t}_{n+1} = t_n + 1$.
- Poisson loc: a homogeneous Poisson process with the conditional intensity function $\lambda(t, r, s) = \frac{\sum_{i=1}^{n} \mathbb{I}[r_i=r, s_i=s]}{t_n - t_0}$ that is constant over time.
- Auto-regressive [6]: which assumes the most recent l inter-event intervals are linearly correlated. We select l as 6.
- Hawkes parametric (Hawkes p) [7]: a temporal Hawkes process with trigger function g defined parametrically as $g(t - t') = e^{-\beta(t-t')}$.
- Hawkes non-parametric (Hawkes n/p): a temporal Hawkes process with trigger function g estimated non-parametrically, *i.e.,* $g = \{g(k\delta t)|k = 0, 1, ...\}$.
- MMEL loc [24]: a multivariate Hawkes process with the u'-th dimensional conditional intensity function $\lambda_{u'}(t)$ given in Eq. (1), where $\{u' = (r, s)|r = 1, ..., R; s = 1, ..., S\}$. Hyper-parameters D is set as 1 and α as 0 for simplicity without losing generality.
- MMEL loc+fault [24]: which uses the same settings as MMEL loc, but with $\{u' = (r, s, v)|r = 1, ..., R; s = 1, ..., S; v = 1, ..., U\}$.
- RMTPP loc [5]: a neural marked temporal point process with its marks being the items in $\{u = (r, s)|r = 1, ..., R; s = 1, ..., S\}$.

[1] An example tweet on 17th February, 2015, is "11:27:37 [EWL] Due to a train fault at Jurong East, there will be no train service from Lakeside to Clementi on the east bound...". The event data is accessible via https://github.com/PAbEve/data

Table 2. Prediction performance of evaluated approaches.

	NextDay	**Auto-reg.**	**Hawkes p**	**Hawkes n/p**	**NNPP**
MAE	8.156	7.131	6.634	6.797	74.878
	Poisson loc	**MMEL loc**	**MMEL fault+loc**	**RMTPP loc**	**PAbEve**
MAE	6.562	6.869	9.588	7.438	**6.156**
Hit rate	0.453	0.503	0.478	0.547	**0.578**
CosSim	0.437	0.485	0.458	0.473	**0.562**

– NNPP [17]: a fully neural temporal point process which models the cumulative conditional intensity function and obtain the conditional intensity function via its derivative.

Metrics. We use 3 kinds of metrics to evaluate the performance of compared methods on m test events, including (1) **MAE**, which is the mean absolute error in days between the predicted and ground-truth times, *i.e.,* $\frac{1}{m}\sum_{l=1}^{m}|\hat{t}_{n+l}-t_{n+l}|$; (2) **Hit rate**, which is the proportion of test events where the predicted and ground-truth rail lines are the same, *i.e.,* $\frac{1}{m}\sum_{l=1}^{m}\mathbb{I}[\hat{r}_{n+l}=r_{n+l}]$; (3) **CosSim**, which is the mean cosine similarity between the predicted and ground-truth stretches, *i.e.,* $\frac{1}{m}\sum_{l=1}^{m}\frac{\boldsymbol{x}_{n+l}\cdot\hat{\boldsymbol{x}}_{n+l}}{|\boldsymbol{x}_{n+l}||\hat{\boldsymbol{x}}_{n+l}|}\mathbb{I}[r_{n+l}=\hat{r}_{n+l}]$. For MAE, smaller values are better, while for Hit rate and CosSim, larger values are preferred.

5.2 Experimental Results

We run each evaluated method for 10 rounds, and take the average of 10 rounds for each metric. We set the hyper-parameters as $\delta t = 1$, $K = 25$ and $M = 10$. The results of prediction are summarized in Table 2.

Results of timing prediction. In terms of MAE, we may draw the following conclusions. First, parametric methods are neither superior nor inferior to semi-parametric methods according to the experiment results. Semi-parametric method refers to those with a part of the model (*e.g.,* the trigger function) designed in a non-parametric way, including all Hawkes-based evaluated methods except Hawkes p. We see among the top 5 performing methods, there are semi-parametric methods PAbEve with MAE 6.156, Hawkes n/p with MAE 6.797 and MMEL loc with MAE 6.869, and also parametric methods Poisson loc with MAE 6.562 and Hawkes p with MAE 6.634. Second, among the parametric methods, however, only light-weight methods, *i.e.,* Poisson loc, Hawkes p and Autoregressive, can compete with semi-parametric methods. Heavy-weight methods which may not be able to express the sparse data may perform arbitrarily bad, such as NNPP which yield a MAE of 74.878. Particularly, the naive Poisson loc outperforms all the other evaluated methods except PAbEve, which indicates that those non-trivial methods may degrade when trained on insufficient data. Third, among the semi-parametric Hawkes-based methods, we see Hawkes n/p of 6.797 outperforms MMEL loc of 6.869, and MMEL loc outperforms MMEL fault+loc of 9.588. We suspect that increasing the number of inputs (*e.g.,* rail line, technical fault) is probable to worsen the performance, as the number of parameters to learn are increased as well. Overall PAbEve outperforms the others

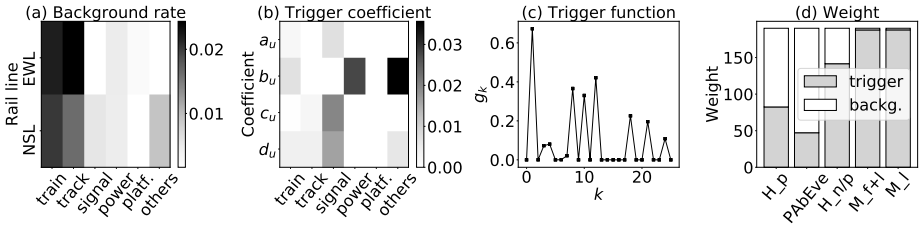

Fig. 2. Visualization of (a) background rates, (b) trigger coefficients, (c) trigger function, and (d) sum of training events' p_{ii}'s ("backg.") and that of p_{ij}'s ("trigger").

as it properly incorporates all aspects of information via the dedicated design of inter-event influences.

Results of location prediction. Location prediction consists of the prediction of abnormal rail line and the stretch of stations. We evaluate 5 methods that can conduct location prediction, and PAbEve outperforms the others. For the prediction of abnormal rail lines, RMTPP loc performs close to PAbEve's, but it uses unadjustable prediction for all events (*i.e.,* standard deviation is zero), and so is Poisson loc. Comparison between MMEL loc and MMEL fault+loc shows increasing the number of inputs may worsen the performance. For the prediction of abnormal stretches, PAbEve again outperforms all others. Among them RMTPP loc predicts trivially using the entire rail line. Both RMTPP loc and Poisson loc predict using the same value for all events.

Model interpretation. To interpret PAbEve, we visualize the estimated background rates, trigger coefficients, trigger function and weights. The results are shown in Fig. 2. From Fig. 2(a) and (b), we see events of the highest background rates are those on both rail lines caused by train fault or track fault, and the category of inter-line overlapping event pairs caused by "others" fault has the most significant inter-event influences. For the trigger function, as shown in Fig. 2(c), it fluctuates between 0 and 0.67, which is dissimilar to exponential or power law functions and this may simply be resulted from the data sparsity issue. Finally, for each of the 5 Hawkes-based evaluated methods, we investigate the probabilities of an event being natural arrival (represented by the sum of p_{ii}'s) or triggered event (represented by the sum of p_{ij}'s). PAbEve is the one with the largest ratio of background versus trigger (*i.e.,* about 3 to 1).

Results of sensitivity tests. We explore the impact of hyper-parameters δt, K and M, on the prediction performance. Each result is averaged over 10 rounds. The results are depicted in Fig. 3. We test the impact of δt by setting K to 25, and changing δt from 1 to 10 with PAbEve. The MAEs shown in Fig. 3(a) depict a clear trend that increasing δt will probably worsen the performance. When $\delta t \leq 2$, PAbEve outperform all other evaluated methods. Similarly, we test the impact of K by setting δt to 1, and changing K from 5 to 100 at an interval of 5 with PAbEve. The MAEs depicted in Fig. 3(b) range from 6.156 to 8.056. When $K \leq 65$, PAbEve outperform all other methods with the MAEs oscillating between 6.156 and 6.538. Finally, for the number of rail line segments, M, as shown in Fig. 3(c), there are only mild changes of MAEs for most methods when

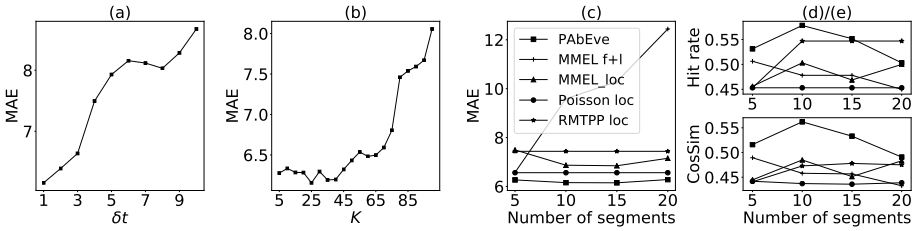

Fig. 3. Sensitivity tests. (a) MAEs under δt from 1 to 10 ($K = 25$); (b) MAEs under K from 5 to 100 ($\delta t = 1$); (c) MAEs, (d) Hit rates and (e) CosSims, of all location-available methods under M from 5 to 20 ($\delta t = 1$, $K = 25$).

M changes, except MMEL fault+loc for which the errors increase significantly. For location prediction results shown in Fig. 3(d) and (e), as M increases, there is no specific trend for the hit rate of rail line or the similarity measure of stretch.

6 Conclusion

We present a novel solution to predicting when and where will a future MRT abnormal event occur, based on historical abnormal events. We first categorize event pairs based on basic contextual information, and then design a multivariate Hawkes process to model the sparse sequence of abnormal events. The proposed PAbEve approach retains its efficacy when being trained on extremely limited training events. Experimental results using real-world data from open sources demonstrate the superiority of PAbEve over other alternative solutions.

Acknowledgments

This work is supported by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), and Singapore MOE AcRF Tier 1 RG18/20.

References

1. Adhikari, B., Xu, X., Ramakrishnan, N., Prakash, B.A.: Epideep: Exploiting embeddings for epidemic forecasting. In: Proceedings of the 25th ACM SIGKDD. pp. 577–586 (2019)
2. Apostolopoulou, I., Linderman, S., Miller, K., Dubrawski, A.: Mutually regressive point processes. In: Advances in Neural Information Processing Systems. pp. 5115–5126 (2019)
3. Deng, S., Rangwala, H., Ning, Y.: Learning dynamic context graphs for predicting social events. In: Proceedings of the 25th ACM SIGKDD. pp. 1007–1016 (2019)
4. Ding, D., Zhang, M., Pan, X., Yang, M., He, X.: Modeling extreme events in time series prediction. In: Proceedings of the 25th ACM SIGKDD. pp. 1114–1122 (2019)
5. Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., Song, L.: Recurrent marked temporal point processes: Embedding event history to vector. In: Proceedings of the 22nd ACM SIGKDD. pp. 1555–1564 (2016)

6. Engle, R.F., Russell, J.R.: Autoregressive conditional duration: a new model for irregularly spaced transaction data. Econometrica pp. 1127–1162 (1998)
7. Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika **58**(1), 83–90 (1971)
8. Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D., Hampapur, A.: Improving rail network velocity: A machine learning approach to predictive maintenance. Transportation Research Part C: Emerging Technologies **45**, 17–26 (2014)
9. Li, J., Han, Z., Cheng, H., Su, J., Wang, P., Zhang, J., Pan, L.: Predicting path failure in time-evolving graphs. In: Proceedings of the 25th ACM SIGKDD. pp. 1279–1289 (2019)
10. Li, S., Xiao, S., Zhu, S., Du, N., Xie, Y., Song, L.: Learning temporal point processes via reinforcement learning. arXiv preprint arXiv:1811.05016 (2018)
11. Li, Z., Zhang, J., Wu, Q., Gong, Y., Yi, J., Kirsch, C.: Sample adaptive multiple kernel learning for failure prediction of railway points. In: Proceedings of the 25th ACM SIGKDD. pp. 2848–2856 (2019)
12. Marsan, D., Lengline, O.: Extending earthquakes' reach through cascading. Science **319**(5866), 1076–1079 (2008)
13. Mei, H., Eisner, J.M.: The neural hawkes process: A neurally self-modulating multivariate point process. Advances in Neural Information Processing Systems **30**, 6754–6764 (2017)
14. Ning, Y., Muthiah, S., Rangwala, H., Ramakrishnan, N.: Modeling precursors for event forecasting via nested multi-instance learning. In: Proceedings of the 22nd ACM SIGKDD. pp. 1095–1104 (2016)
15. Ning, Y., Tao, R., Reddy, C.K., Rangwala, H., Starz, J.C., Ramakrishnan, N.: Staple: Spatio-temporal precursor learning for event forecasting. In: Proceedings of the 2018 SIAM International Conference on Data Mining. pp. 99–107 (2018)
16. Okawa, M., Iwata, T., Kurashima, T., Tanaka, Y., Toda, H., Ueda, N.: Deep mixture point processes: Spatio-temporal event prediction with rich contextual information. In: Proceedings of the 25th ACM SIGKDD. pp. 373–383 (2019)
17. Omi, T., Ueda, N., Aihara, K.: Fully neural network based model for general temporal point processes. arXiv preprint arXiv:1905.09690 (2019)
18. Rasmussen, J.G.: Lecture notes: Temporal point processes and the conditional intensity function. arXiv preprint arXiv:1806.00221 (2018)
19. Reinhart, A., et al.: A review of self-exciting spatio-temporal point processes and their applications. Statistical Science **33**(3), 299–318 (2018)
20. Shchur, O., Biloš, M., Günnemann, S.: Intensity-free learning of temporal point processes. arXiv preprint arXiv:1909.12127 (2019)
21. Sipos, R., Fradkin, D., Moerchen, F., Wang, Z.: Log-based predictive maintenance. In: Proceedings of the 20th ACM SIGKDD. pp. 1867–1876 (2014)
22. Tan, C.: Smrt fined record $5.4 million for july 7 breakdown (jan 2016), https://www.straitstimes.com/singapore/transport/smrt-fined-record-54-million-for-july-7-breakdown
23. Zhao, L., Sun, Q., Ye, J., Chen, F., Lu, C.T., Ramakrishnan, N.: Multi-task learning for spatio-temporal event forecasting. In: Proceedings of the 21th ACM SIGKDD. pp. 1503–1512 (2015)
24. Zhou, K., Zha, H., Song, L.: Learning triggering kernels for multi-dimensional hawkes processes. In: International Conference on Machine Learning. pp. 1301–1309 (2013)
25. Zuo, S., Jiang, H., Li, Z., Zhao, T., Zha, H.: Transformer hawkes process. In: International Conference on Machine Learning. pp. 11692–11702. PMLR (2020)