Multi-stream based marked point process

Sujun Hong HONG.SUJUN@G.WAKAYAMA-U.JP
Hirotaka Hachiya HHACHIYA@WAKAYAMA-U.AC.JP
Graduate School of System Engineering, Wakayama University

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

When using a point process, a specific form of the model needs to be designed for intensity function, based on physical and mathematical prior knowledge about the data. Recently, a fully trainable deep learning-based approach has been developed for temporal point processes. This approach models a cumulative hazard function (CHF), which is capable of systematic computation of adaptive intensity function in a data-driven manner. However, this approach does not take the attribute information of events into account although many applications of point processes generate with a variety of marked information such as location, magnitude, and depth of seismic activity. To overcome this limitation, we propose a fully trainable marked point process method, modeling decomposed CHFs for time and mark using multi-stream deep neural networks. In addition, we also propose to encode multiple marked information into a single image and extract necessary information adaptively without detailed knowledge about the data. We show the effectiveness of our proposed method through experiments with simulated toy data and real seismic data.

Keywords: point process, time series analysis, multi-stream model, seismic event

1. Introduction

In recent years, tremendous disasters have been caused by the earthquake and heavy rainfall. Unlike ordinary time-series data recorded in a regular time cycle, e.g., stock price and maximum temperature, these natural environmental events usually occur in an irregular time cycle. This irregular time cycle would make the prediction and analysis prohibitively tricky. To treat irregular time-series data, a stochastic process called **point process** has been studied.

An important characteristic of the point process for overcoming the irregular time cycle is to model a decay function, called, **intensity function**, which represents how likely a next event occurring given a certain time and location. So far, several works for designing intensity functions, have been done (Ogata (1998); Omi et al. (2019); Zhu et al. (2019)). A baseline model of intensity function is a mixture of an exponential decay function in time and a Gaussian function in space, i.e., in ETAS (Ogata (1998)), that has been applied for predicting aftershocks. Besides, the model of intensity function has been explored, to capture a non-linear relationship to a past sequence of events, and heterogeneous spatial dependence, by introducing Gaussian diffusion kernels (Zhu et al. (2019)), recurrent neural network (RNN) (Du et al. (2016); Mei and Eisner (2016); Xiao et al. (2017); Li et al. (2018); Upadhyay et al. (2018); Huang et al. (2019)) and transformer (Zuo et al. (2020)).

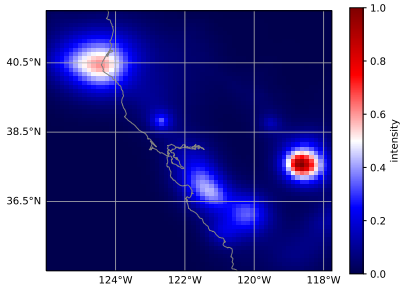


Figure 1: An example of mark-encoded images for seismic events with the magnitude of 3 or greater, occurred in Northern California between 2011 and 2015 (NCEDC (2014)). To create the image, the attribute information mark of an event, i.e., latitude and longitude, magnitude and depth is encoded using Eq. 14, and the sum of all images for the events is normalized to 0 to 1 in a pixel-wise manner.

Recently, a fully trainable deep learning-based model was proposed (Omi et al. (2019)) where a **cumulative hazard function** (CHF) is modeled by the combination of RNN and fully-connected network (FCN), and the intensity function is computed through the back-propagation of networks. This fully trainable model has been applied for the temporal point process and shown to provide better performances.

However, in this model, attribute information of events, called **mark**, is not taken into account so far although many applications of the point process generate marked events, e.g., the location, magnitude and depth of seismic activities. Also, while the marked point process has been studied, mark information has been rather explicitly encoded into the intensity function, based on physical and mathematical prior-knowledge, e.g., in ETAS (Ogata (1998)), Gaussian diffusion kernel and transformer-based models (Zhu et al. (2019); Zuo et al. (2020)).

In this work, we explore a fully trainable marked point process by decomposing the joint CHF of time and mark to reduce its complexity induced by the heterogeneous multi-modality between time and mark. We model the decomposed CHFs, temporal and marked CHFs, using multi-stream deep neural networks—each cumulative hazard is generated based on the intermediate variables containing the information of the past sequence of events and time-mark candidates.

In addition, to realize higher accurate marked point process, we also propose to encode multiple marked information into a single image visually without detailed and strict prior-knowledge. Fig. 1 depicts an example of a mark-encoded image for seismic events with a magnitude of 3 or greater, that occurred in Northern California between 2011 and 2015. This figure was created following Eq. 14 and the mark of seismic events, i.e., its location (latitude and longitude), magnitude, and depth, are visually marked in the image. The figure visually explains seismic events frequently occur in three different locations and the ones in the east tend to have higher magnitudes. Through auto-encoder network with our proposed point process model, called Multi-stream-PP (Multi-stream fully trainable marked

Point Process), we let the model extract necessary information from the image to accurately predict future events.

We show the effectiveness of our proposed methods through experiments with simulated toy data and real seismic data recorded in Northern California. From the experiments, we confirm that our method could predict future events with higher accuracy than the existing methods.

2. Related works

In this section, we review the basis of the marked point process (MPP) and introduce the state-of-the-art temporal point process (TPP) methods.

2.1. Formulation of point process

In an MPP, an event e_i is considered to occur at an irregular time t_i with attribute information, i.e., **mark** $\boldsymbol{m}_i \in \mathbf{R}_{d^{\mathcal{M}}}$ where $d^{\mathcal{M}}$ is the number of attributes. The mark here is assumed to be continuous values, e.g., the location coordinate of an event $\boldsymbol{m}_i = (x_i, y_i)$. Let $H_t = \{(t_i, \boldsymbol{m}_i) | t_i < t\}$ be a past sequence of time-mark pairs, and $\lambda(t, \boldsymbol{m}|H_t)$ be an intensity function conditioned on the past sequence H_t. Then, the log-likelihood of observing N-event sequence $\{(t_i, \boldsymbol{m}_i)\}_{i=1}^N$ in the time period $[0, T]$ is defined (Chen et al. (2021)) as

$$\log p\Big(\{(t_i, \boldsymbol{m}_i)\}_{i=1}^N\Big) = \sum_{i=1}^N \log \lambda(t_i, \boldsymbol{m}_i|H_{t_i}) - \int_0^T \int_{\mathbf{R}_{d^{\mathcal{M}}}} \lambda(t, \boldsymbol{m}|H_{t_i})\, dt\, d\boldsymbol{m} \tag{1}$$

where the second term is to exponentially discount the probability density by the integral of intensities of non-occurrence time and mark. The intensity function is optimized to maximize the log-likelihood but computing the second term of Eq. 1, i.e., the integral of the intensity function, would be in general difficult without an explicit model.

2.2. Explicit model-based marked point process

ETAS is a widely-used MPP, specifically for predicting aftershocks of earthquakes (Ogata (1998)). The mark information obtained in seismic data includes the location (latitude, longitude) $\boldsymbol{x}$ and magnitude z; that is, $\boldsymbol{m} = (\boldsymbol{x}, z)$. Then, the marked intensity function in ETAS is defined as

$$\lambda(t, \boldsymbol{m}|H_t) = \mu(\boldsymbol{x}) + \sum_{t_i < t} k(z_i) p_{\text{time}}(t - t_i) p_{\text{mark}}(\boldsymbol{x} - \boldsymbol{x}_i | z_i) \tag{2}$$

where $\mu(\boldsymbol{x})$ and $k(z)$ are a base intensity and the expected number of events triggered by a magnitude of z. $p_{\text{time}}(t - t_i)$ and $p_{\text{mark}}(\boldsymbol{x} - \boldsymbol{x}_i|z_i)$ are probability density functions (PDFs) of the deviation of time t and location $\boldsymbol{x}$ from events in the past sequence H_t—the second term is the mixture of PDFs weighted by $k(z)$. To systematically compute the integral of the intensity function, PDFs are designed as an exponential decay and Gaussian functions, respectively, based on the assumption that the time and location are independent of each other.

However, this explicitly designed model would require psychical and mathematical prior-knowledge on the generation process of events, and careful hand-made design of intensity function to obtain higher performance. Thus, the versatility and scalability of the existing MPP models are rather limited; that is, physically unknown attribute information could not be encoded into the intensity function and not used for predicting future events.

2.3. Partially trainable model-based point process

To improve the versatility of the intensity function, recent works (Zhu et al. (2019); Du et al. (2016); Zuo et al. (2020)) introduced partially trainable models, using the mixture of Gaussian diffusion kernels (Zhu et al. (2019)), transformer (Zuo et al. (2020)) and recurrent neural network (RNN) (Du et al. (2016)). A standard recipe in these models is to extract a feature vector $\boldsymbol{h}_i$ from the past sequence H_{t_i} using state-of-the-art deep architectures, e.g., transformer and RNN. Then, $\boldsymbol{h}_i$ is inputted to a handmade non-negative decay function $\phi(\cdot)$, called **hazard function** which is equivalent to the intensity function $\lambda(\cdot)$ as

$$\lambda(t, \boldsymbol{m}|H_{t_i}) = \phi(\tau, \boldsymbol{\sigma}|\boldsymbol{h}_i) \tag{3}$$

where $\tau = t - t_i$ and $\boldsymbol{\sigma} = \boldsymbol{m} - \boldsymbol{m}_i$ are the deviations of time and mark from the last event e_i. Fig. 2 illustrates the relationship between a temporal past sequence $H_{t_i}^{\mathrm{T}} = \{t_i | t_i < t\}$, RNN, and hazard function $\phi(\cdot)$ in a temporal point process (Du et al. (2016)) where the sequence feature vector is extracted using RNN with parameters $\boldsymbol{\alpha}$ as

$$\boldsymbol{h}_i = \mathrm{RNN}_{\boldsymbol{\alpha}}(H_{t_i}^{\mathrm{T}}) \tag{4}$$

Since the hazard function $\phi(\cdot)$ is designed using non-negative decay function, e.g., exponential decay function, the limitation on the versatility would still exist.

2.4. Fully trainable temporal point process

A fully trainable TPP (FT-TPP) model was recently proposed (Omi et al. (2019)) where the cumulative hazard function (CHF) $\Phi(\cdot)$ is modeled by a fully-connected network (FCN) with non-negative-weight constrains as

$$\begin{aligned} \Phi(\tau|\boldsymbol{h}_i) &\equiv \int_0^{\tau} \phi(\eta|\boldsymbol{h}_i) d\eta \\ \widehat{\Phi(\tau|\boldsymbol{h}_i)} &= \mathrm{FCN}_{\boldsymbol{\beta} \geq 0}(\tau, \boldsymbol{h}_i) \end{aligned} \tag{5}$$

where $\mathrm{FCN}_{\boldsymbol{\beta} \geq 0}(\cdot)$ is FCN with non-negative parameters $\boldsymbol{\beta}$.

Given the modeled CHF, the corresponding hazard function $\phi(\cdot)$ can be computed using its derivative as

$$\widehat{\phi(\tau|\boldsymbol{h}_i)} = \frac{\partial}{\partial \tau} \widehat{\Phi(\tau|\boldsymbol{h}_i)} \tag{6}$$

The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of networks are tuned to maximize the log-likelihood with the sequence of N-event, e.g., in Eq. 1 as

$$\begin{aligned} \log p\Big(\{(t_i)\}_{i=1}^N\Big) &= \sum_{i=1}^N \left[\log \frac{\partial}{\partial \tau} \Phi(\tau_i|\boldsymbol{h}_i) - \Phi(\tau_i|\boldsymbol{h}_i)\right] \approx \sum_{i=1}^N \left[\log \frac{\partial}{\partial \tau} \widehat{\Phi(\tau_i|\boldsymbol{h}_i)} - \widehat{\Phi(\tau_i|\boldsymbol{h}_i)}\right] \\ &\equiv \log L(\boldsymbol{\alpha}, \boldsymbol{\beta}|\{t_i\}_{i=1}^N) \end{aligned} \tag{7}$$

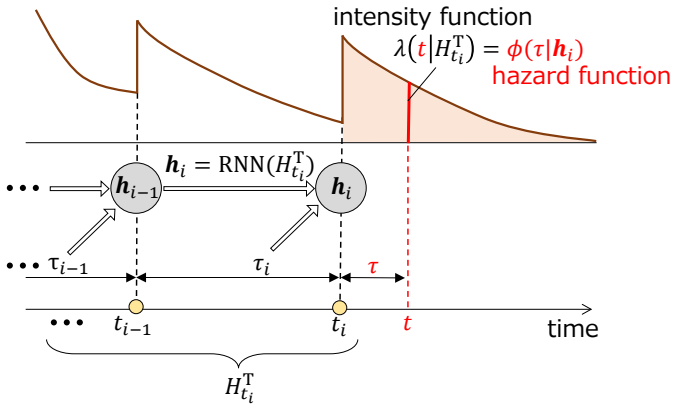


Figure 2: Illustration of RNN-based TPP where a handmade hazard function $\phi(\tau|\boldsymbol{h}_i)$ is computed given a sequence vector $\boldsymbol{h}_i$ extracted from a past sequence $H_{t_i}^{\mathrm{T}}$ by recurrent neural network (RNN).

where the derivative of the CHF model $\widehat{\Phi(\tau_i|\boldsymbol{h}_i)}$ can be efficiently computed through the back-propagation process of networks. This fully trainable model enables the temporal intensity function $\lambda(t)$ adaptive to irregular elapsed-time depending on the past sequence H_t^{T}.

There is a need to incorporate attribute information of events, i.e., the mark $\boldsymbol{m}$, into the model since many applications of the point process generate marked events as well, e.g., location, magnitude, and depth of seismic activity. However, modeling the joint CHF $\Phi(\tau, \boldsymbol{\sigma}|\boldsymbol{h}_i)$ with the fully trainable model is not straightforward due to the heterogeneous multi-modality between time and mark, e.g., with different scales, distributions and resolution requirements, etc.

3. Proposed method

In this work, we extend FT-TPP (see Sec. 2.4) to incorporate marked information without physical and mathematical prior knowledge (discussed in Sec. 2.2). To mitigate the heterogeneous multi-modality between time and mark, we propose Multi-stream fully trainable marked Point Process (Multi-stream-PP), which combines multi-stream networks and FT-TPP.

3.1. Decomposition of joint CHF

Let us assume that time t and mark $\boldsymbol{m}$ of new event are conditionally independent, given intermediate variables $\mathbf{s}_i$ containing the information of the past sequence of time-mark pairs H_{t_i}, and candidates of the deviations on time τ and mark $\boldsymbol{\sigma}$. Then, we decompose the joint CHF given $\mathbf{s}_i$, $\Phi(\tau, \boldsymbol{\sigma}|\mathbf{s}_i)$ into temporal CHF $\Phi_{\mathrm{time}}(\tau|\mathbf{s}_i)$ and marked CHF $\Phi_{\mathrm{mark}}(\boldsymbol{\sigma}|\mathbf{s}_i)$ as

$$\Phi(\tau, \boldsymbol{\sigma}|\mathbf{s}_i) = \Phi_{\mathrm{time}}(\tau|\mathbf{s}_i)\Phi_{\mathrm{mark}}(\boldsymbol{\sigma}|\mathbf{s}_i) \tag{8}$$

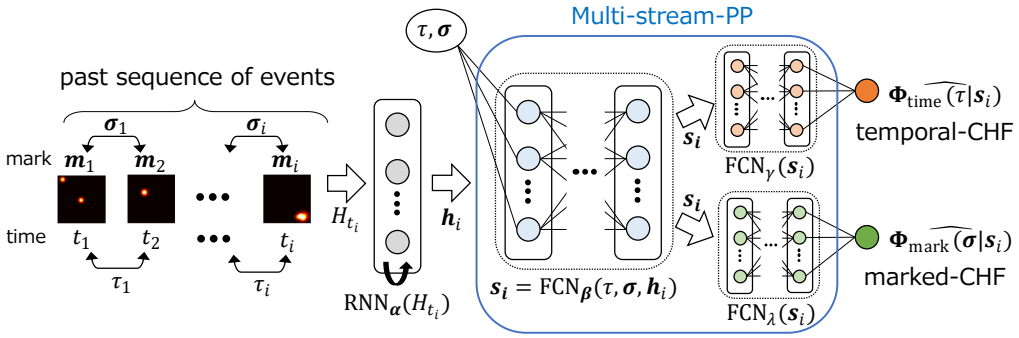


Figure 3: Illustration of architecture of our proposed, multi-stream fully trainable marked point process (Multi-stream-PP) consisting of fusion network $\text{FCN}_{\beta}(\cdot)$, and decomposed multi-stream networks $\text{FCN}_{\gamma}(\cdot)$ and $\text{FCN}_{\lambda}(\cdot)$. The multi-stream networks express decomposed CHFs: temporal-CHF $\Phi_{\text{time}}(\tau|\mathbf{s}_i)$ and marked-CHF $\Phi_{\text{mark}}(\boldsymbol{\sigma}|\mathbf{s}_i)$.

In addition, let us assume that temporal $\phi_{\text{time}}(\tau|\mathbf{s}_i)$ and marked $\phi_{\text{mark}}(\boldsymbol{\sigma}|\mathbf{s}_i)$ hazard functions are obtained by the gradients of individual CHFs w.r.t. elapsed τ and mark $\boldsymbol{\sigma}$ as

$$\frac{\partial}{\partial \tau}\Phi_{\text{time}}(\tau|\mathbf{s}_i) = \frac{\partial \mathbf{s}_i}{\partial \tau}\frac{\partial}{\partial \mathbf{s}_i}\Phi_{\text{time}}(\tau|\mathbf{s}_i) \equiv \phi_{\text{time}}(\tau|\mathbf{s}_i) \tag{9}$$

$$\frac{\partial}{\partial \boldsymbol{\sigma}}\Phi_{\text{mark}}(\boldsymbol{\sigma}|\mathbf{s}_i) = \frac{\partial \mathbf{s}_i}{\partial \boldsymbol{\sigma}}\frac{\partial}{\partial \mathbf{s}_i}\Phi_{\text{mark}}(\boldsymbol{\sigma}|\mathbf{s}_i) \equiv \phi_{\text{mark}}(\boldsymbol{\sigma}|\mathbf{s}_i) \tag{10}$$

Given decomposed hazard and cumulative hazard functions, we can derive the likelihood of observing N-event sequence $\{(t_i, \boldsymbol{m}_i)\}_{i=1}^{N}$ as

$$\log\ p(\{(t_i,\boldsymbol{m}_i)\}_{i=1}^{N}) = \sum_{i=1}^{N}\left[\log \frac{\partial}{\partial \tau}\Phi_{\text{time}}(\tau_i|\mathbf{s}_i) + \log \frac{\partial}{\partial \boldsymbol{\sigma}}\Phi_{\text{mark}}(\boldsymbol{\sigma}_i|\mathbf{s}_i) - \Phi_{\text{time}}(\tau_i|\mathbf{s}_i)\Phi_{\text{mark}}(\boldsymbol{\sigma}_i|\mathbf{s}_i)\right] \tag{11}$$

where we note that there also exist gradients of cross variable, i.e., $\frac{\partial}{\partial \boldsymbol{\sigma}}\Phi_{\text{time}}(\tau|\mathbf{s}_i)$ and $\frac{\partial}{\partial \tau}\Phi_{\text{mark}}(\boldsymbol{\sigma}|\mathbf{s}_i)$ but we do not include them in the likelihood.

3.2. Multi-stream fully trainable marked point process (Multi-stream-PP)

The decomposed expression of likelihood enables us to model temporal and marked hazard functions individually and to mitigate the complexity caused by the heterogeneous multi-modality between time τ and mark $\boldsymbol{\sigma}$.

More specifically, as illustrated in Fig. 3, we can model temporal and mark CHFs using fusion and multi-stream deep neural networks. In the fusion network $\text{FCN}_{\beta}(\cdot)$, the intermediate variables $\mathbf{s}_i$ is extracted by taking the correlation among the time-mark candidates τ and $\boldsymbol{\sigma}$, and the past sequence of events H_{t_i} into account.

$$\mathbf{s}_i = \text{FCN}_{\boldsymbol{\beta}\geq 0}\Big(\tau, \boldsymbol{\sigma}, \text{RNN}_{\boldsymbol{\alpha}}(H_{t_i})\Big) \tag{12}$$

In the multi-stream, individual CHFs are approximated in a fine-grained manner as follows,

$$\begin{aligned}\widehat{\Phi_{\text{time}}(\tau|\mathbf{s}_i)} &= \text{FCN}_{\boldsymbol{\gamma}\geq 0}(\mathbf{s}_i)\\ \widehat{\Phi_{\text{mark}}(\boldsymbol{\sigma}|\mathbf{s}_i)} &= \text{FCN}_{\boldsymbol{\lambda}\geq 0}(\mathbf{s}_i)\end{aligned} \tag{13}$$

where parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ of streams are also non-negative. All parameters in the entire architecture, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ are tuned so as to maximize the log-likelihood (Eq. 11).

3.3. Mark-encoded vector

In order to realize a fully-trainable marked point process with high accuracy, it is important to embed information about an event e_i appropriately in the mark vector $\boldsymbol{m}_i$. In this work, we consider the following two embedding methods.

3.3.1. Direct mark-encoded vector (direct-mark)

The mark vector $\boldsymbol{m}_i$ simply consists of attribute values. For example, in a seismic event with location (latitude, longitude) $\boldsymbol{x}$, magnitude z and depth d, a mark vector is expressed by $\boldsymbol{m} = (\boldsymbol{x}, z, d)$.

3.3.2. Image-based mark-encoded vector (imaginary-mark)

Multiple marked information of an event e_i is encoded into a single image using the shape, size, location, pixel-value and channel-axis information. For example, in the seismic event, the mark could be encoded into a single-channel image by

$$I_i(x, y) = z_i \exp\left\{-\sqrt{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2 + d_i^2}\right\} \tag{14}$$

where the location coordinate $\boldsymbol{x}_i = (x_i, y_i)$ of latitude x_i and longitude y_i is encoded as the pixel coordinate, and the magnitude z_i and depth d_i are encoded as the size and intensity on the image. Fig. 1 depicts an example of a mark-encoded image for seismic events with the magnitude of 3 or greater occurred in Northern California between 2011 and 2015 (NCEDC (2014)) [1]. The figure visually explains seismic events occur frequently in three different locations and the ones in the east tend to have higher magnitudes.

Then, we extract a mark vector $\boldsymbol{m}_i$ from a mark-encoded image I_i using an encoder-network as

$$\boldsymbol{m}_i = \text{Enc}_{\boldsymbol{\theta}}(I_i) \tag{15}$$

where $\text{Enc}_{\boldsymbol{\theta}}(\cdot)$ is the encoder-network with parameters $\boldsymbol{\theta}$, consisting of multiple convolutional layers pre-trained to reproduce images with the decoder-network, i.e., auto-encoder.

1. Note that in Fig. 1, multiple mark-encoded images are summed and normalized for the visualization purpose.

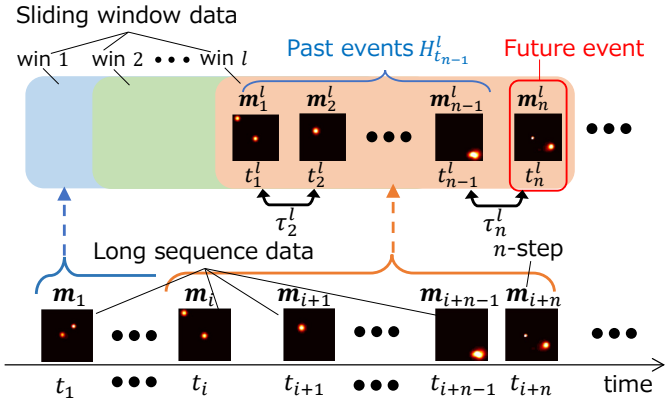


Figure 4: Process of splitting the long sequence data into sliding windows for experimental evaluation. Each of sliding windows contains $n-1$-step past events and n-th step future event.

3.4. Future event prediction

Similarly to existing point process models, our proposed Multi-stream-PP provides the likelihood for the given candidate of future events but not directly outputs the future event itself; thus, the candidate-based search is required for the prediction. For this purpose, we prepare candidates of elapsed-time $\{\tau_{(c)}\}$ and mark-encoded vector $\{\boldsymbol{m}_{(c)}\}$ based on training data.

With the candidates, we predict a next event $(\widehat{t_{i+1}}, \widehat{\boldsymbol{m}_{i+1}})$ alternatively following three steps:

1. Select the maximum-likelihood elapsed-time $\widehat{\tau_{i+1}}$ from the candidate $\{\tau_{(c)}\}$ while the mark-encoded vector $\widehat{\boldsymbol{m}_{i+1}}$ is temporarily set to the past observed one $\boldsymbol{m}_{i-2}$.

2. Select the maximum-likelihood mark-encoded vector $\widehat{\boldsymbol{m}_{i+1}}$ from the candidate $\{\boldsymbol{m}_{(c)}\}$ given the selected elapsed-time $\widehat{\tau_{i+1}}$ at step 1.

3. Select again the maximum-likelihood elapsed-time $\widehat{\tau_{i+1}}$ from the candidate $\{\tau_{(c)}\}$ given the selected marked-encoded vector $\widehat{\boldsymbol{m}_{i+1}}$ at step 2.

Step 2 and 3 are repeated until convergence or with a certain number of iterations.

4. Experimental evaluation

In this section, we evaluate our proposed methods, Multi-stream-PP with direct-mark and imaginary-mark, through experiments with simulated toy and real seismic data. We evaluate the performance of our proposed method with two marked point process models—ETAS (Ogata (1998)) and simple marked extension of FT-TPP (Omi et al. (2019)).

4.1. Sliding-window evaluation data

Both toy and real seismic data consist one long sequence respectively. To conduct training and evaluation, we split the long sequence data into sliding windows with n-step, as shown in Fig. 4. Let the last n-th event $(t_n^l, \boldsymbol{m}_n^l)$ be the future event for prediction, and the sequential events $H_{t_{n-1}}^l$ before the last be the history of past events for input.

In both toy and real seismic data, the last 800 events are used for evaluation; 769 windows, each of which contains 31 events, e.g., $n = 31$ and $L = 769$, are created following Sec. 4.1. The window data before the last 800th event, are used for training models. From training data, 1000 candidates of elapsed-time $\{\tau_{(c)}\}$ are created by equally splitting the range between the minimum and maximum ones. Similarly, 900 candidates of mark-encoded vector $\{\boldsymbol{m}_{(c)}\}$ are created using k-means clustering.

4.2. Evaluation metrics

To evaluate the performance of future event prediction, we use mean absolute error (MAE) for the elapsed-time, and mean squared error (MSE) for mark-encoded vector. More specifically, as for MAE, the top-10 likelihood candidates of elapsed-time $\tau_{(c)}$ are selected from each window and the error between the median of the candidates and the true value τ_n^l is computed as

$$\mathrm{MAE} = \frac{1}{L}\sum_{l=1}^{L}\left|\tau_n^l - \mathrm{med}(\{\tau_{(c)}^l\}_{p(\tau_{(c)}^l)\geq p_{10}})\right| \tag{16}$$

where L is the number of windows in the evaluation data and p_{10} is the top-10th likelihood.

As for the MSE, the predicted mark $\widehat{\boldsymbol{m}_{i+1}}$ is selected from candidates using the process in Sec. 3.4, and the distance to the true value $\boldsymbol{m}_n^l$ is computed as

$$\mathrm{MSE} = \frac{1}{L}\sum_{l=1}^{L}\left\|\boldsymbol{m}_n^l - \widehat{\boldsymbol{m}_{i+1}}\right\|^2 \tag{17}$$

4.3. Setting of existing methods

The existing methods, ETAS and simple marked extension of FT-TPP are used for the performance comparison. As for ETAS, we define each function in Eq. 2 as

$$\begin{aligned} k(z) &= A\exp\Big(\alpha(z-z_0)\Big) \\ p_{\mathrm{time}}(\tau) &= \frac{p-1}{c}\Big(1+\frac{\tau}{c}\Big)^{-p} \\ p_{\mathrm{mark}}(\boldsymbol{m}-\boldsymbol{m}_i|z) &= \frac{1}{2\pi\sigma(z)}\exp\left\{-\frac{(\boldsymbol{m}-\boldsymbol{m}_i)^2}{2\sigma(z)}\right\} \\ \sigma(z) &= D\exp\Big(\gamma(z-z_0)\Big) \end{aligned}$$

where z_0 is a reference magnitude and parameters $(\mu, A, c, \alpha, p, D, \gamma)$ are tuned using the maximum likelihood method (Jalilian (2019)).

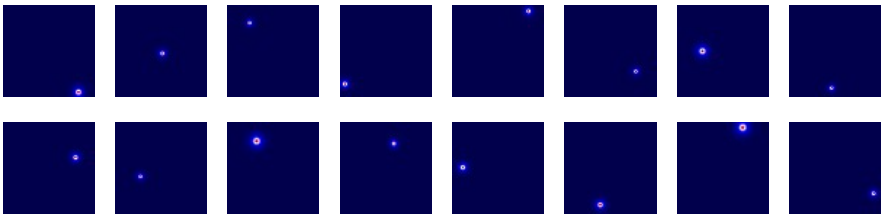


Figure 5: Examples of candidates of mark-encoded image $I_i(x, y)$ in simulated toy data. 900 candidate are created using k-means clustering from attribute vectors in training data, and then converted to marked-encoded vectors directly (direct-mark) or via image (imaginary-mark).

In simple marked extension of FT-TPP, $\boldsymbol{m}$ in Eq. 1 is set as $\boldsymbol{m} = (\boldsymbol{x}, z, d)$ where pairs of elapsed-time τ_i and mark deviation $\boldsymbol{\sigma}_i$ are inputted to RNN and the hidden $\boldsymbol{h}_i$ is extracted. The log-likelihood function of FT-MPP with joint-hazard-function as

$$\log\ p(\{(t_i,\boldsymbol{m}_i)\}_{i=1}^N) = \sum_{i=1}^N \left[\log \frac{\partial}{\partial \tau}\Phi(\tau, \boldsymbol{\sigma}|\boldsymbol{h}_i)\nabla_{\phi_{\text{time}}} + \log \frac{\partial}{\partial \boldsymbol{\sigma}}\Phi(\tau, \boldsymbol{\sigma}|\boldsymbol{h}_i)\nabla_{\phi_{\text{mark}}} - \Phi(\tau, \boldsymbol{\sigma}|\boldsymbol{h}_i)\right] \tag{18}$$

where $\nabla_{\phi_{\text{time}}}$ and $\nabla_{\phi_{\text{mark}}}$ are the time t and mark $\boldsymbol{m}$ scale-adjustment factors required for total differentiation.

4.4. Evaluation with simulated toy data

We use the simulation data generated following a Hawks process in time and a Gaussian mixture distribution in space (Zhu et al. (2019)). An attribute vector of an event e_i is expressed as $\boldsymbol{m}_i = (\boldsymbol{x}_i, z_i)$ where the intensity z_i is generated by computing based on the past events at each coordinates $\boldsymbol{x}_i$ using a pre-defined Gaussian mixture distribution. Then, mark-encoded image $I_i(x, y)$ is created using Eq. 14 without the depth d_i as shown in Fig. 5.

The results of the evaluation of the proposed and existing methods are depicted in Table 1. The table shows that our proposed methods would predict future events with higher accuracy in comparison with existing methods. As for more detailed analysis for the prediction of elapsed-time, Fig. 6 depicts the comparison between true τ_n^l and predicted elapsed-time $\widehat{\tau_n^l}$ at each event e_i. The figures show that the existing methods predict relatively smaller elapsed-time values for entire events. Meanwhile, our proposed methods tend to capture even medium elapsed time, although it would be still difficult to capture large elapsed time due to the highly imbalanced data where the majority of data is with small (nearly zero) elapsed time [2]. One of our future works will be to devise loss functions to improve the prediction for large elapsed time, e.g., by using importance weighting.

2. In the simulated toy data used in the experiment, the ratio of events of elapsed time ≥ 1, is only about 10%, i.e., 4,239 out of 40,000 for the training and 101 out of 800 for the test. Since elapsed time ≥ 1 is the minority in the training, in general, it would be difficult to predict large elapsed times.

Table 1: Performance comparison using mean absolute error (MAE) for temporal prediction and mean squared error (MSE) for mark prediction on simulated toy data. The method with the best score is indicated in bold.

Methods	MAE	MSE
ETAS (Ogata (1998))	0.642	2.081
FT-TPP (Omi et al. (2019))	0.419	-
FT-MPP with joint-hazard-function	0.429	1.281
Proposed Multi-stream-PP with direct-mark	**0.392**	1.279
Proposed Multi-stream-PP with imaginary-mark	0.399	**0.949**

Table 2: Performance comparison using mean absolute error (MAE) for temporal prediction and mean squared error (MSE) for mark prediction on real seismic data. The method with the best score is indicated in bold.

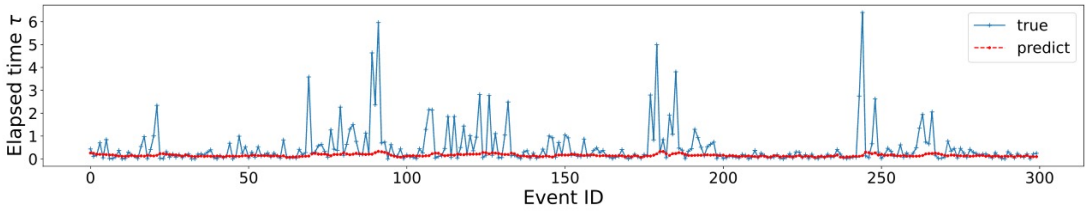
Methods	MAE	MSE
ETAS (Ogata (1998))	44.473	7.084
FT-TPP (Omi et al. (2019))	41.693	-
FT-MPP with joint-hazard-function	44.311	6.204
Proposed Multi-stream-PP with direct-mark	**39.531**	4.804
Proposed Multi-stream-PP with imaginary-mark	41.195	**3.946**

Secondly, the predicted mark-encoded images are depicted in Fig. 7 in which the first and second rows are for the true I_n^l and predicted $\widehat{I_n^l}$ mark-encoded images respectively. The figures show that our proposed method, Multi-stream-PP with imaginary-mark would select the candidates with similar locations and shapes.

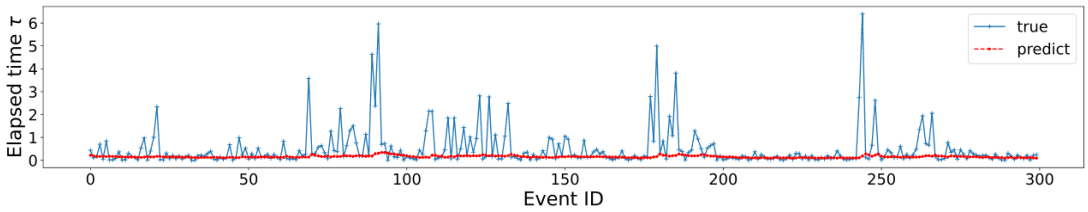
4.5. Evaluation with real seismic data

The Northern California Earthquake Data Center (NCEDC) provides public time-series data (NCEDC (2014)) observed by a strong motion sensor, GPS, and other geophysical sensors. We extract 13,607 seismic event records from years 1978 to 2019, each of which is a magnitude larger than 3.0. As for the elapsed time τ_i, we compute the time difference between two consecutive seismic events, e.g., e_i and e_{i+1}. As for the marked information $\boldsymbol{m}_i$, we use the geographic location (latitude and longitude) $\boldsymbol{x}_i$, magnitude z_i, and depth d_i for a seismic event. From the marked information $\boldsymbol{m}_i$, the intensity of each pixel is calculated using Eq. 14 as shown in Fig. 1.

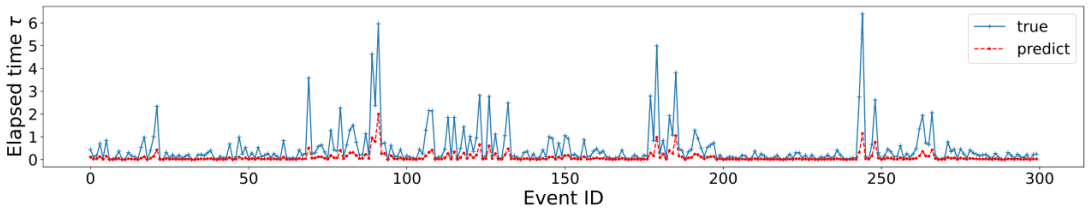
The results of the evaluation of the proposed and existing methods are depicted in Table 2. These results show that our proposed methods could predict future events with higher accuracy compared to existing methods. As for more detailed analysis for the prediction of elapsed-time τ_n^l, Fig. 8 depicts the comparison between true τ_n^l and predicted $\widehat{\tau_n^l}$ elapsed-time. The figures show that the existing methods tend to stay around the average value or to predict almost zero, while our methods tend to follow the true value relatively well.



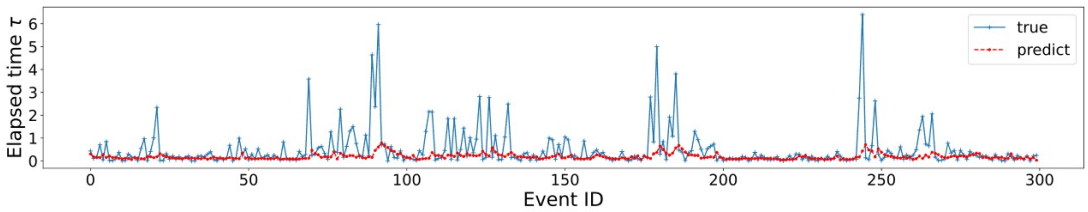
(a) FT-TPP



(b) FT-MPP with joint-hazard-function



(c) Proposed Multi-stream-PP with direct-mark



(d) Proposed Multi-stream-PP with imaginary-mark

Figure 6: Comparison between true τ_n^l and predicted $\widehat{\tau_n^l}$ elapsed-time values in simulated toy data for each method. Blue and red lines are the true and predicted ones respectively.

Secondly, Fig. 9 depicts the true mark-encoded images I_n^l and the ones $\widehat{I_n^l}$ predicted by our proposed method, Multi-stream-PP with imaginary-mark, show that our proposed method selects candidate that coordinates and shapes are close to the true values. Overall, our proposed methods could predict future events with more accurately than the existing methods.

Lastly, as for more detailed analysis for the prediction of mark $\boldsymbol{m}_n^l$, Fig. 10 depicts the examples of true $\boldsymbol{m}_n^l$ and predicted $\widehat{\boldsymbol{m}_n^l}$ marked vectors in the test seismic data. The figure (a) and (b) show that the predicted mark vectors for 10 events with small MSEs are relatively

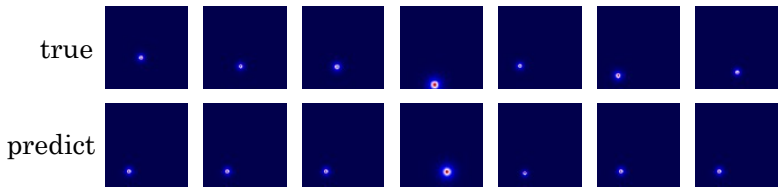


Figure 7: Comparison between true mark-encoded images I_n^l and the ones $\widehat{I}_n^l$ predicted by the our proposed method, Multi-stream-PP with imaginary-mark in simulation data. To create predicted mark-encoded images, the selected mark-encoded vector $\widehat{\boldsymbol{m}_n^l}$ is converted to the image using Eq. 14.

close to the true ones in terms of the location and size of paired symbols, indicating that our proposed method, Multi-stream-PP could handle the combination of multiple marked information well. In addition, the figure (c) and (d) show that the predicted ones for 10 events with largest magnitude or depth are also close to the true ones in the magnitude and depth, i.e., the size of symbols, but are relatively far from true ones in the location. This would indicate that the prediction performances are degraded for the minor combination of mark values, e.g., with large magnitude and depth, since the candidates created by k-means from training data tend to be sparse. Thus, one of our future works will be to introduce dynamic exploration of time-mark candidates.

5. Conclusion

In this work, we proposed a multi-stream based fully trainable marked point process (Multi-stream-PP) where the complex joint cumulative hazard function (CHF) is decomposed, and temporal and marked CHFs are modeled by multi-stream deep neural networks. In addition, we proposed to encode multiple mark information into a single image to lead the model adaptively extract necessary features in order to predict subsequent events. We showed the effectiveness of our proposed methods, called Multi-stream-PP, with existing methods using simulated toy data and Northern California earthquake data. Further analysis of the proposed method with multiple real data would be future work.

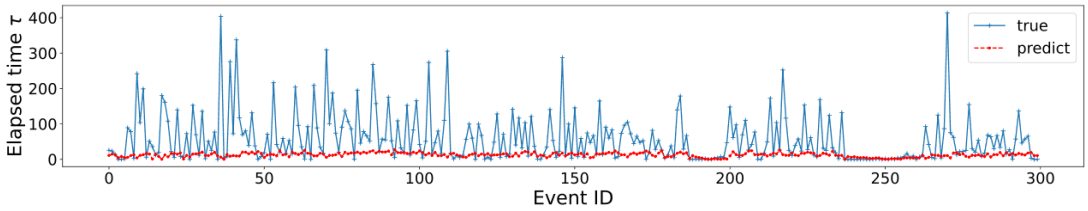
Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP20K11863.

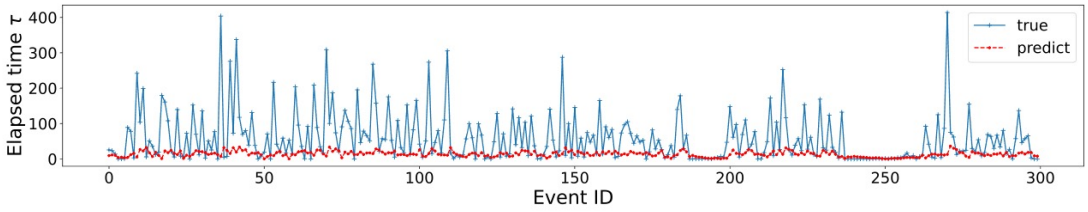
References

Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel. Neural spatio-temporal point processes. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

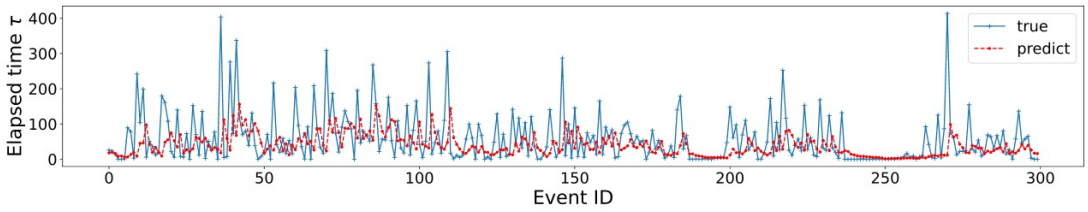
Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vec-



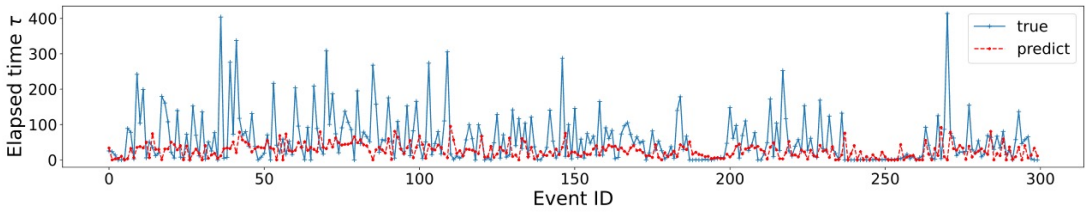
(*a*) FT-TPP



(*b*) FT-MPP with joint-hazard-function



(*c*) Proposed Multi-stream-PP with direct-mark



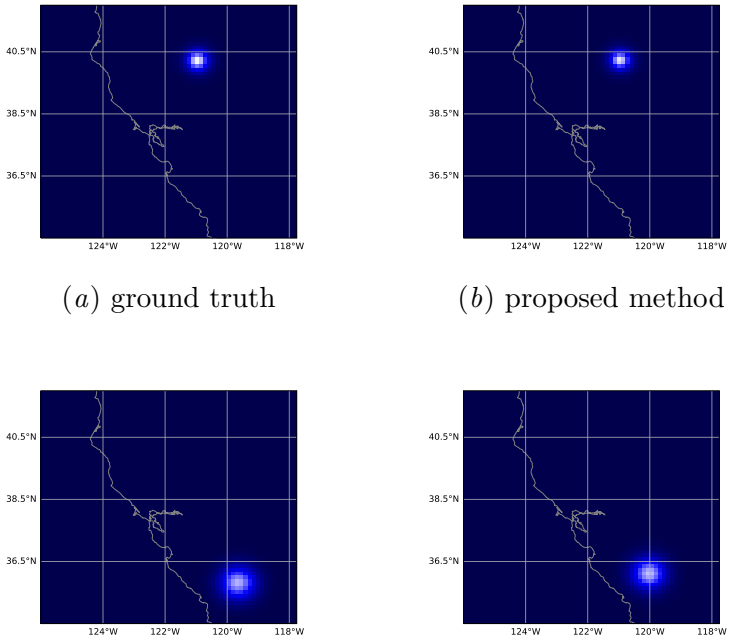
(*d*) Proposed Multi-stream-PP with imaginary-mark

Figure 8: Comparison between true τ_n^l and predicted $\widehat{\tau_n^l}$ elapsed-time values in real seismic data for each method. Blue and red lines are the true and predicted ones respectively.

tor. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1555–1564, 2016.

Hengguan Huang, Hao Wang, and Brian Mak. Recurrent poisson process unit for speech recognition. In *Proceedings the 32nd AAAI Conference on Artificial Intelligence*, volume 33, pages 6538–6545, 2019.

Abdollah Jalilian. ETAS: An R package for fitting the space-time ETAS model to earthquake data. *Journal of Statistical Software, Code Snippets*, 88(1):1–39, 2019. doi: 10.18637/jss.



(a) ground truth (b) proposed method

(c) ground truth (d) proposed method

Figure 9: Comparison between true In^l and the one $\widehat{In^l}$ predicted by our proposed method, Multi-stream-PP with imaginary-mark, for the real seismic event on, May 30, 2013: (a) and (b), and June 25, 2014: (c) and (d).

v088.c01.

Shuang Li, Shuai Xiao, Shixiang Zhu, Nan Du, Yao Xie, and Le Song. Learning temporal point processes via reinforcement learning. *arXiv preprint arXiv:1811.05016*, 2018.

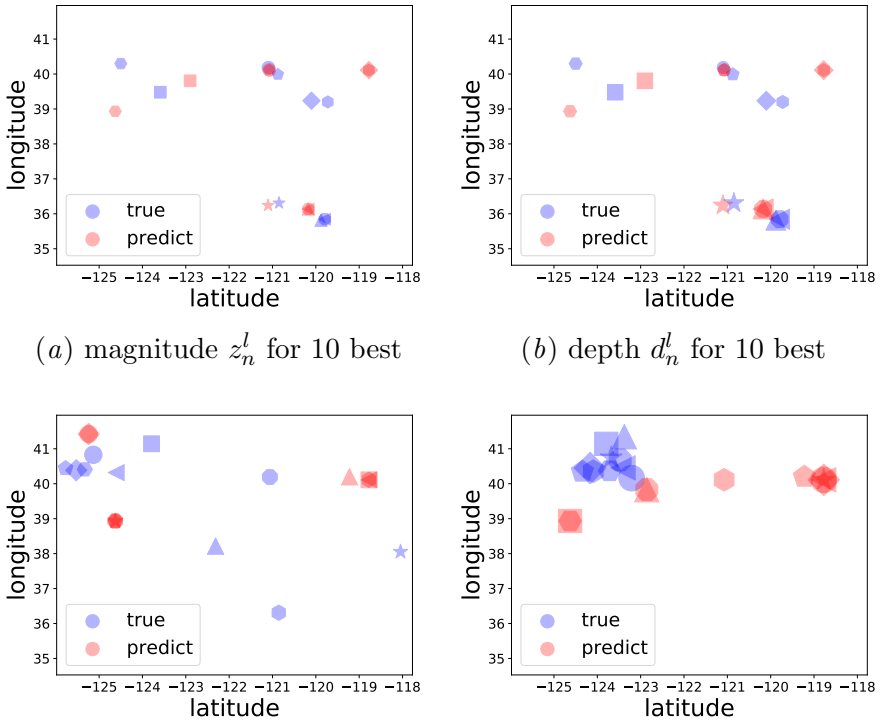
Hongyuan Mei and Jason Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. *arXiv preprint arXiv:1612.09328*, 2016.

UC Berkeley Seismological Laboratory. NCEDC. Northern california earthquake catalog and phase data, 2014.

Yosihiko Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998.

Takahiro Omi, Kazuyuki Aihara, et al. Fully neural network based model for general temporal point processes. In *Advances in Neural Information Processing Systems*, pages 2122–2132, 2019.

Utkarsh Upadhyay, Abir De, and Manuel Gomez-Rodriguez. Deep reinforcement learning of marked temporal point processes. *arXiv preprint arXiv:1805.09360*, 2018.



(a) magnitude z_n^l for 10 best

(b) depth d_n^l for 10 best

(c) magnitude z_n^l for 10 largest

(d) depth d_n^l for 10 largest

Figure 10: Examples of true marked vectors and predicted ones by our proposed method, Multi-stream-PP, in the test seismic data. A pair of blue and red points with the same symbol expresses the ground truth $\boldsymbol{m}_n^l$ and its corresponding prediction $\widehat{\boldsymbol{m}_n^l}$ for the event e_n^l with the coordinate $\boldsymbol{x}_n^l$, magnitude z_n^l and depth d_n^l. The values of coordinate and magnitude/depth are depicted with the position and size of symbols. (a) and (b): examples of true and predicted location and magnitude/depth for the 10 events with smallest MSEs. (c) and (d): examples of ones for 10 events with largest magnitude/depth.

Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen M Chu. Modeling the intensity function of point process via recurrent neural networks. In *proceedings of 31st AAAI conference on Artificial Intelligence*, pages 1597–1603, 2017.

Shixiang Zhu, Shuang Li, and Yao Xie. Interpretable generative neural spatio-temporal point processes. *arXiv preprint arXiv:1906.05467*, 2019.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. *arXiv preprint arXiv:2002.09291*, 2020.