Spatiotemporal Event Forecasting from Incomplete Hyper-local Price Data

ABSTRACT

Hyper-local pricing data, e.g., about foods and commodities, exhibit subtle spatiotemporal variations that can be useful as crucial precursors of future events. Three major challenges in incorporating such pricing data include: i) temporal depencies underlying features; ii) spatiotemporal missing values; and iii) constraints underlying economic phenomena. These challenges hinder traditional event forecasting models from being applied effectively. This paper proposes a novel spatiotemporal event forecasting model that concurrently addresses the above challenges. Specifically, given continuous price data, a new soft time-lagged model is designed to select temporally dependent features. To handle missing values, we propose a data tensor completion method based on price domain knowledge. The parameters of the new model are optimized using a novel algorithm based on the Alternative Direction Methods of Multipliers (ADMM). Extensive experimental evaluations on multiple datasets demonstrate the effectiveness of our proposed approach.

1 INTRODUCTION

Hyper-local pricing data, about goods and commodities, is becoming useful as an economic variable to study the unfolding of large societal events. For example, research [1] on the 2011 Egypt uprisings, the so-called "*Arab Spring*" [2], has demonstrated a correlation between a rise in food prices and social unrest. A rapid rise in raw material prices, in this case agricultural products, coincided with a series of demonstrations, protests, and even civil wars throughout the Arab world. Moreover, there are strong spatio-temporal associations one could detect. For instance, considering the price of coffee, Colombia's largest export, recent research [3] indicated that when the price of this labor-intensive commodity rises, work hours and wages increase and conflict declines in the areas that produce it. In contrast, the collapse in coffee prices from 1997-2003 resulted in 18% more guerrilla attacks, 31% more paramilitary attacks, and 22% more clashes with the authorities in Columbia's coffee growing areas. Therefore, price data about goods and commodities is a powerful indicator linking significant societal events within a spatial and temporal context.

Although price data is collected from organizations such as the World Bank [4], these datasets are often released at only a monthly level and often with a time-lag; as a result they cannot be used as a real-time indicator of evolving events. Nowadays, given the pervasiveness of mobile devices, mobile contributors have become important sources for price data at precise locations. For example, Premise[1] collects thousands of commodity prices in more than 30 countries across six continents, while GasBuddy[2] collects local gas prices from millions of vehicle drivers.

[1] https://www.premise.com/

[2] https://www.gasbuddy.com

Existing spatiotemporal event forecasting methods are mainly focused on datasets containing no or few missing data points and do not consider the characteristics of hyper-local price data, largely because event forecasting based on the price data is a complex problem that faces several important challenges. **1. Temporal price dependence**. Unlike most traditional datasets, pricing information cannot be treated as a set of independent point objects. For example, the average home value in Palo Alto[3] went up from $1.58M in 2013 to $2.48M in 2016, but as it stands, this data does not show the important detail that this rate rose by 30% in the past year alone. This example demonstrates that micro-trends in spatial and temporal profiles important to model as part of the feature extraction process. **2. Missing values in hyper-local datasets**. When prices are collected from local data contributors, missing data are quite common and typically attributed to spatial sparsity and temporal discontinuities. As the missing data often has a strong correlation with its neighbors, simply discarding them is not an ideal strategy as the resulting datasets become too sparse to support meaningful analysis. **3. Price domain knowledge utilization**. Price is not an arbitrary number but a precisely controlled factor determined by markets and economic forces. For example, if the price of a tradable good in one region is markedly higher than in neighboring areas, entrepreneurs will transport the goods from the lower price areas to leverage the price in the higher price areas. Thus, utilizing prior domain knowledge in conjunction with spatially and temporally sparse price data is clearly beneficial.

In order to simultaneously address all these technical challenges, in this paper we propose a novel hyper-local price based event forecasting model (HPEF). The main contributions of our study are summarized below. We:

- **Design a framework for event forecasting based on hyper-local price data.** A flexible framework shown in Figure 1 is proposed for spatiotemporal event forecasting that utilizes hyper-local price data collected by millions of independent data collectors. The proposed framework simultaneously exploits massive spatiotemporal hyper-local price data and addresses the shortage of high ratio missing data from sparse data collectors.
- **Propose a robust model for spatiotemporal missing values with soft time-lagged feature selection.** To model interactions between continuous prices across multiple time periods, we propose a soft time-lagged feature selection regularization based on spatiotemporal price data. To handle interactions among missing values, the model proposed here adopts a spatiotemporal tensor completion framework that is capable of learning the missing values based on temporal and spatial coherence.

[3] A city in San Francisco, CA

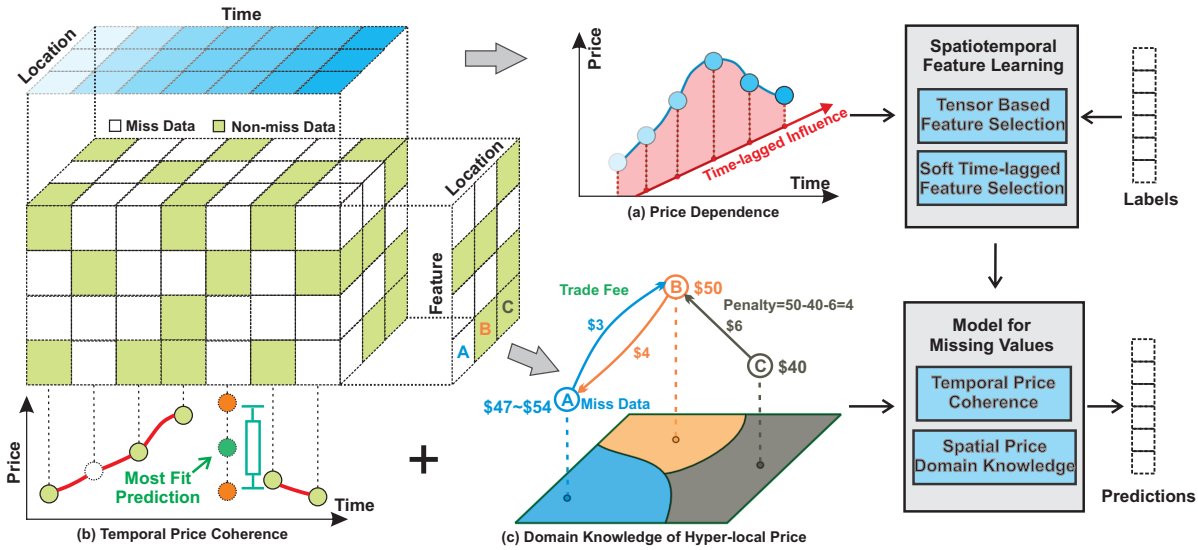


Figure 1: A schematic view of the proposed spatiotemporal event forecasting model for hyper-local price data

- **Develop an efficient algorithm for model parameter optimization.** To optimize the proposed model, a constrained spatiotemporal tensor completion problem combined with coefficient decay weights had to be solved. By introducing auxiliary variables, we have developed an efficient algorithm based on the alternating direction method of multipliers (ADMM) to solve the problem with rapid convergence.
- **Conduct extensive experiments to evaluate the performance of HPEF.** Our proposed method was evaluated using 6 different datasets in two domains: forecasting civil unrest in South America and real estate trends in the United States. The results demonstrate that the proposed approach runs efficiently and consistently outperforms the best of the existing methods in terms of multiple metrics.

The remainder of this paper is organized as follows. Section 2 describes the related work on event forecasting, missing values in high dimensional data, and temporal interaction based feature selection. The problem definition is presented in Section 3. Our proposed HPEF model is described in Section 4 and Section 5 gives details of the new model's optimization algorithm. In Section 6, the experimental results are analyzed and the paper concludes with a summary of our work in Section 7.

2 RELATED WORK

This section introduces related work in several research areas.

Event detection and forecasting. A large body of work has focused on the identification of ongoing events such as earthquakes [5] and disease outbreaks [6], while event forecasting methods predict the incidence of such events in the future. Most event forecasting methods focus on purely temporal events, with no interest in the geographical dimension, such as stock market movements [7] and elections [8]. Few existing approaches are capable of providing true spatiotemporal resolution for the events they predict, although Zhao et al. [9] designed a multitask learning framework that models forecasting tasks in related geo-locations concurrently and Gerber utilized a logistic regression model for spatiotemporal event forecasting [10]. One limitation of these existing studies is that the temporal dimension is considered to be independent of the spatial dimension and any interactions between the two are ignored.

Missing values in high dimensional data. The prevention and management of missing data has been discussed and investigated in earlier work [11]. One research category focuses on estimating missing entries based on the observed values [12], but although these methods work well when missing data are rare, they are less effective when a significant amount of data is missing. Another category of work is to utilize matrix completion to find a matrix with a low-rank for the observed entries, and this has been actively studied in statistical learning [13], information retrieval [14], and signal processing [15]. However, neither of these methods can be applied to high dimensional data. Recently, several methods have extended the matrix completion approach to tensor completion [16][17]: Liu et al. [16] estimated the missing data in video streams via tensor completion generalized from matrix completion methods, while Wang et al. [18] proposed a tensor completion method that preserves temporal consistency in video data to estimate the missing values across frames. However, few of these studies jointly consider both temporal consistency and spatial relationships in high dimensional data.

Feature selection in the presence of time-lagged interactions. Feature selection that considers feature interactions has been attracting research interest for some time. For instance, to enforce specific interaction patterns, Harrell et al. [19] employed a conventional step-wise model selection technique with hierarchical constraint, while Lim and Hastie [20] worked with a hierarchical group-lasso regularization to learn pairwise interactions. Unfortunately, none of these lasso-based approaches work for the time-lagged interactions in price data. Although Park et al. [21] and Suo et al. [22] proposed two different time-lagged regression methods for time series data, both applied lagged weights with hard constraints are too strong an assumption for price data.

3 PROBLEM DEFINITION

In this section, the problem addressed by this research is formalized. The notations used here are summarized in Table 1, in which the terminology of tensors specified in [23] is applied.

Table 1: Math Notations

Notations	Explanations
F, L, T	Feature, location and time interval number
P	Time intervals for temporal weights
K	Dimensions of price data tensor
$\mathcal{W} \in \mathbb{R}^{F\times L\times P}$	Time-lagged weight tensor
$\mathcal{X} \in \mathbb{R}^{F\times L\times P}$	Hyper-local price data tensor
$\mathcal{X}_{(i)} \in \mathbb{R}^{F\times L\times P}$	Unfolded data tensor in i-th dimension
$\mathcal{X}_{(f)} \in \mathbb{R}^{F\times L\times P}$	Unfolded data tensor in feature dimension
$\mathcal{X}_{\{n\}} \in \mathbb{R}^{F\times L\times P}$	Feature tensor in 3 dimensions
$Y \in \mathbb{R}^{L\times P}$	Predicted ground truth matrix
$D \in \mathbb{R}^{L\times L}$	Auxiliary matrix for temporal difference

Let us denote $X = \{x_t\}_t^T$ as a sequence of tuples that contains hyper-local prices collected by local data contributors with spatiotemporal information. Here, $x_t = (f, l, t, p)$ represents a price tuple composed of product type f, geo-location l, timestamp t and price p. For example, a price tuple for flour in the Brazilian market can be represented as (Flour, [38.89,-77.24], 03/20/2014, 6.35). Each price tuple can be geo-coded into an administrative region based on their geographic coordinates such as the city, *Los Angeles*, or neighborhood, *Harlem*, in the New York City borough of Manhattan. The region price of a product in one time interval can be formally defined as follows:

Definition 3.1. **Region Price**. Let us use the set $X_{\{l,t\}}$ to denote prices of product f as $\forall x_k \in X_{\{l,t\}}$ for region l and time interval t. The region price of product f in location l and time interval t can now be defined as $x_{\{f,l,t\}} = \frac{1}{K}\sum_{k=1}^{K} p_{\{f,k\}}$, where $p_{\{f,k\}}$ is the price for the k-th tuple in $X_{\{l,t\}}$ of product f and K is the size of set $X_{\{l,t\}}$.

Organize the region prices in the form of a data tensor $\mathcal{T} \in \mathbb{R}^{F\times L\times T}$ in three dimensions: product, region and time interval. The price of product f in region l and time interval t can now be represented as $\mathcal{T}_{\{f,l,t\}}$. Specifically, the elements $\mathcal{T}_{\{0,l,t\}} = 1$ serve as a dummy feature to provide a compact notation for the bias parameter in the forecasting model. Denoting the position set of missing values as Ω, we can set the missing values in $\mathcal{T}_{\bar{\Omega}} = 0$. However, simply setting missing values as zero will break the price coherence in both spatial and temporal directions. To solve the problem, we create another data tensor variable $\mathcal{X}$ to represent the real price data based on *Law of One Price* [24] which is defined as below.

Definition 3.2. **Law of One Price(LoP)**. LoP is an economic concept which posits that "a good must sell for the same price in all locations". After considering the transportation and transaction fee, the law can be defined as:

$$\frac{\varphi_a}{\varphi_b + c_{b\to a}} = 1 \quad (\varphi_a > \varphi_b)$$

where φ_a and φ_b are the prices at locations a and b, respectively, and $c_{b\to a}$ is the trading fee(including transportation and transaction fees) per unit from location b to a. Therefore, the definition of the new data tensor based on spatiotemporal price prior knowledge is as follows:

Definition 3.3. **Completed Price Tensor**. Given a data sensor $\mathcal{T}$ with missing values, the completed price tensor is the data tensor that has the following properties: 1) $\mathcal{T}_{\bar{\Omega}} = \mathcal{X}_{\bar{\Omega}}$, where $\bar{\Omega}$ is the set of data positions containing data, 2) The prices in $\mathcal{X}$ possess temporal price coherence, and 3) The price at different locations obeys the *LoP* property.

Applying the above definitions, the problem addressed in this paper can be formulated as follows:

Problem Formulation: Given the hyper-local price data tensor $\mathcal{T}$ in spatial and temporal dimensions from time interval t to $t+p$, the goal is to predict the occurrence of future event $Y_{\tau,l}$ for location l at time interval τ. In addition, $\tau = t+P+\delta$, where $\delta > 0$ is the lead time and P is the length of continuous time period considered in our model. Formally, this problem is formulated as learning a mapping function from price tensor data to predict future spatiotemporal events: f: $\mathcal{T}_{\{t:t+P,l\}} \to Y_{\tau,l}$, where f is the forecasting model.

4 MODEL

In this section, we propose a new model to forecast spatiotemporal events based on hyper-local prices. The new tensor based event forecasting model minimizes the following penalized empirical loss:

$$\min_{\mathcal{W},\mathcal{X}} \mathcal{L}(\mathcal{W},\mathcal{X}) + \Omega_t(\mathcal{W}) + \Omega_s(\mathcal{X}) \quad s.t.\ \mathcal{X}_{\bar{\Omega}} = \mathcal{T}_{\bar{\Omega}} \tag{1}$$

where $\mathcal{L}(\mathcal{W},\mathcal{X})$ is the empirical loss function and feature weight tensor $\mathcal{W} \in \mathbb{R}^{F\times L\times P}$ is the parameter of the forecasting model in feature, location and time dimensions, $\Omega_t(\mathcal{W})$ is the regularization term that encodes the interactions between continuous price features, and $\Omega_s(\mathcal{X})$ is the regularization term that ensures the coherence between missing values and existing values.

4.1 Loss Function

The event forecasting error $\mathcal{L}(\mathcal{W},\mathcal{X})$ is defined as the sum of the empirical errors of the prediction values against the labels $Y_{\tau,l}$. For the binary case event forecasting problem, the loss function $\mathcal{L}$ can be the logistic loss [25], as follows:

$$\begin{aligned}\mathcal{L}(\mathcal{W},\mathcal{X}) = -\sum_{t=1}^{T}\sum_{l=1}^{L}[Y_{\tau,l}\log h(\mathcal{W}_l \odot \mathcal{X}_{\{l,t:t+P-1\}}) \\ + (1-Y_{\tau,l})\log(1-h(\mathcal{W}_l \odot \mathcal{X}_{\{l,t:t+P-1\}}))]\end{aligned} \tag{2}$$

where $\mathcal{W}_l$ is the weight matrix at location l. $\mathcal{X}_{\{l,t:t+P-1\}}$ is the data matrix for location l and the time interval range from t to $t+P-1$. The operator $\odot$ is the summation of the Hadamard product of two matrices such that $A \odot B = \sum_{i,j} A_{ij} \cdot B_{ij}$. For the multiple classes event forecasting problem, the softmax [26] loss function can be used.

4.2 Soft Time-lagged Weight Feature Selection

The feature weight regularization terms are shown as follows:

$$\Omega_t(\mathcal{W}) = \lambda_1 \sum_{p=2,l,f}^{P,L,F} \max(|\mathcal{W}_{\{f,l,p-1\}}| - |\mathcal{W}_{\{f,l,p\}}|, 0) + \lambda_2 \|\mathcal{W}_{(f)}\|_1 \tag{3}$$

where $\lambda_{1,2}$ are the parameters for each term. When predicting an outcome at time t based on the prices at the previous P time points, it is natural to assume that the coefficients decay as we move farther away from t as shown in Figure 1(a). However, far from using a hard constraint [22], we propose instead a soft time-lagged weight constraint to penalize the weights that break the assumption. For example, if $|\mathcal{W}_{\{f,l,t-1\}}| > |\mathcal{W}_{\{f,l,t\}}|$, we can use the absolute difference between the two weights as the penalty. Notice that feature weight can be negative if the feature has inverse impact on our predicted event. The second term with *L1* norm ensures the sparsity of feature weights.

4.3 Missing Feature Values in the Presence of Interactions

The missing value regularization terms are as follows:

$$\begin{aligned}\Omega_s(\mathcal{X}) = \theta_1 \sum_{i=1}^{K} \|\mathcal{X}_{(i)}\|_* + \theta_2 \sum_{t=1}^{T'-1} \|\mathcal{X}_{\{t+1\}} - \mathcal{X}_{\{t\}}\|_F^2 \\ + \theta_3 \sum_{f=1}^{F} \sum_{t=1}^{T} \Psi(\mathcal{X}_{\{f,t\}})\end{aligned} \tag{4}$$

where $\mathcal{X}_{(i)}$ represents the unfolded tensor in its i-th dimension, K is the dimension of tensor $\mathcal{X}$, the data tensor time length $T' = T + P$, and $\theta_{1,2,3}$ are the parameters for each term. The spatial coherence function $\Psi(x)$ is defined as:

$$\begin{aligned}\Psi(x) = \sum_{i=1}^{L} \sum_{i,j\in\Gamma}^{L} \max(x_i - x_j - C_{j,i}, 0) \\ + \max(x_j - x_i - C_{i,j}, 0)\end{aligned}$$

where Γ set is defined as $\Gamma = \{i, j | i \neq j, x_j \neq 0\}$, and $C_{i,j}$ represents the trading fee required to shift from location i to j.

The first term in Eq (4) is based on the assumption that rows in an unfolded tensor are not linearly independent; for example, the prices of coffee and coffee beans is likely to be highly related in different locations. In order to find this coherence, we use the nuclear norm $\|\cdot\|_*$, the tightest convex envelop for the rank of matrices, to minimize the rank of the tensor in each dimension. As shown in Figure 1(b), the second term is a smoothing factor to ensure the price change is continuous because the prices of commodities do not change sharply in most circumstances. The third term is to constrain prices in different locations according to the *Law of One Price* [24].

To constraint the price in the data tensor to LoP, we use the function $\Psi(\mathcal{X})$ to penalize those locations that fail to satisfy the law. For example, as shown in Figure 1(c), if $\varphi_b = 50$, and $c_{a\to b} = 3, c_{b\to a} = 4$, our purpose is to constrain φ_a within the range of $[47, 54]$. The penalty for $\varphi_a = 45$ will then be $max(45 - 50 - 4, 0) + max(50 - 45 - 3, 0) = 2$. In contrast, if the price $\varphi_a = 50$, which is within the range, the penalty is zero. Assuming that the trading fee is under Gaussian distribution: $c_{b\to a} \sim \mathcal{N}(\mu_{b\to a}, \Sigma_{b\to a})$, we can use $C_{ji} = \mu_{j\to i}$ as the normal trading fee from j to i in the trading fee matrix C.

Algorithm 1: HPEF ALGORITHM

Input: $\mathcal{T}$, Y
Output: solution $\mathcal{W}$, $\mathcal{X}$
1 Initialize $\rho = 1$, $\mathcal{X} = \mathcal{T}$, $\mathcal{W}$, $\Phi = \mathbf{0}$,
2 Choose $\varepsilon_r > 0$, $\varepsilon_s > 0$
3 **repeat**
4 Update $\mathcal{W}$, $\mathcal{X}$, Φ by Equations (7) ~ (**??**).
5 Update Lagrangian multipliers αs and βs by Equation (23).
6 Update primal and dual residuals r and s.
7 **if** $r > 10s$ **then**
8 $\rho \leftarrow 2\rho$
9 **else if** $10r < s$ **then**
10 $\rho \leftarrow \rho/2$
11 **else**
12 $\rho \leftarrow \rho$
13 **until** $r < \varepsilon_r$ *and* $s < \varepsilon_s$

5 PARAMETER OPTIMIZATION

In this section, an ADMM (Alternating Direction Method of Multipliers) based framework is proposed to solve the parameter optimization problem in Section 4.

To decouple the overlapping terms in Equation 1, we introduce a set of auxiliary variables $\Phi = \{\mathcal{Q}, \mathcal{V}, \mathcal{R}, \mathcal{M}_i, \mathcal{U}, \mathcal{S}\}$ and reformulate the equation as follows:

$$\begin{aligned}&\min_{\mathcal{W},\mathcal{X},\Phi} \mathcal{L}(\mathcal{W}, \mathcal{X}) + \lambda_1 \sum_{p=2,l,f}^{P,L,F} \max(|\mathcal{V}_{\{f,l,p-1\}}| - |\mathcal{V}_{\{f,l,p\}}|, 0) \\ &+ \lambda_2 \|\mathcal{R}_{(f)}\|_1 + \theta_1 \sum_{i=1}^{K} \|\mathcal{M}_{i(i)}\|_* + \theta_2 \sum_{n=1}^{N_x-1} \|\mathcal{U}_{\{n+1\}} - \mathcal{U}_{\{n\}}\|_F^2 \\ &+ \theta_3 \sum_{f=1}^{F} \sum_{n=1}^{N_x} \Psi(\mathcal{S}_{\{f,n\}}) \\ &s.t.\ \mathcal{W} = \mathcal{Q} = \mathcal{R},\ \mathcal{V}_{\{f,p,l\}} = |\mathcal{Q}_{\{f,l,p-1\}}| - |\mathcal{Q}_{\{f,l,p\}}| \\ &\quad \mathcal{X} = \mathcal{U} = \mathcal{S},\ X = M_i,\ i \in [1, K] \quad \mathcal{X}_\Omega = \mathcal{T}_\Omega\end{aligned} \tag{5}$$

The augmented Lagrangian function of Equation (5) is:

$$\begin{aligned}&L_\rho(\mathcal{W}, \mathcal{X}) = \mathcal{L}(\mathcal{W}, \mathcal{X}) + \langle \alpha_q, \mathcal{W} - \mathcal{Q}\rangle + \frac{\rho}{2}\|\mathcal{W} - \mathcal{Q}\|_F^2 \\ &+ \sum_{p=2,l,f}^{P,L,F} \langle \alpha_{\mathcal{V}\{f,l,p\}}, \mathcal{V}_{\{f,l,p\}} - |\mathcal{Q}_{\{f,l,p-1\}}| + |\mathcal{Q}_{\{f,l,p\}}|\rangle \\ &+ \frac{\rho}{2} \sum_{p=2,l,f}^{P,L,F} \|\mathcal{V}_{\{f,l,p\}} - |\mathcal{Q}_{\{f,l,p-1\}}| + |\mathcal{Q}_{\{f,l,p\}}|\|_F^2 + \langle \alpha_r, \mathcal{W} - \mathcal{R}\rangle \\ &+ \frac{\rho}{2}\|\mathcal{W} - \mathcal{R}\|_F^2 + \sum_{i=1}^{K} \langle \beta_i, \mathcal{X} - \mathcal{M}_i\rangle + \frac{\rho}{2} \sum_{i=1}^{K} \|\mathcal{X} - \mathcal{M}_i\|_F^2 \\ &+ \langle \alpha_u, \mathcal{X} - \mathcal{U}\rangle + \frac{\rho}{2}\|\mathcal{X} - \mathcal{U}\|_F^2 + \langle \alpha_s, \mathcal{X} - \mathcal{S}\rangle + \frac{\rho}{2}\|\mathcal{X} - \mathcal{S}\|_F^2\end{aligned} \tag{6}$$

To solve the objective function in Equation (1) with multiple unknown parameters $\mathcal{W}$ and $\mathcal{X}$, we propose the hyper-local price based event forecasting (HPEF) algorithm shown in Algorithm 1. This algorithm alternately optimizes each of the unknown parameters until convergence is achieved. Lines 4-6 show the updating of each of the unknown parameters and residuals by solving the sub-problems described below. The derivation of S can be found in Appendix B.

5.1 Update $\mathcal{W}$

The weight tensor $\mathcal{W}$ is learned as follows:

$$\begin{aligned}\mathcal{W} = \arg\min_{\mathcal{W}} \mathcal{L}(\mathcal{W}) + \langle \alpha_q, \mathcal{W} - \mathcal{Q}\rangle + \frac{\rho}{2}\|\mathcal{W} - \mathcal{Q}\|_F^2 \\ + \langle \alpha_r, \mathcal{W} - \mathcal{R}\rangle + \frac{\rho}{2}\|\mathcal{W} - \mathcal{R}\|_F^2\end{aligned} \tag{7}$$

which is a generalized logistic regression containing least squares loss functions. Newton method can be performed to solve this problem.

5.2 Update $\mathcal{Q}$

The auxiliary variable $\mathcal{Q}$ is learned as follows:

$$\begin{aligned}\mathcal{Q} = \arg\min_{\mathcal{Q}} \langle \alpha_q, \mathcal{W} - \mathcal{Q}\rangle + \frac{\rho}{2}\|\mathcal{W} - \mathcal{Q}\|_F^2 \\ + \sum_{p=2,l,f}^{P,L,F} \langle \alpha_{\upsilon\{f,l,p\}}, \mathcal{V}_{\{f,l,p\}} - |\mathcal{Q}_{\{f,l,p-1\}}| - |\mathcal{Q}_{\{f,l,p\}}|\rangle \\ + \frac{\rho}{2}\sum_{p=2,l,f}^{P,L,F} \|\mathcal{V}_{\{f,l,p\}} - |\mathcal{Q}_{\{f,l,p-1\}}| - |\mathcal{Q}_{\{f,l,p\}}|\|_F^2\end{aligned} \tag{8}$$

For notation simplicity, we define $\mathcal{Q}_p = \mathcal{Q}_{\{f,l,p\}}$ and $\mathcal{W}_p = \mathcal{W}_{\{f,l,p\}}$ when feature f and location l are fixed. When $p = 1$, the variable $\mathcal{Q}_1$ is learned as follows:

$$\begin{aligned}\mathcal{Q}_1 = \arg\min_{\mathcal{Q}_1} \langle \alpha_{q1}, \mathcal{W}_1 - \mathcal{Q}_1\rangle + \frac{\rho}{2}\|\mathcal{W}_1 - \mathcal{Q}_1\|_F^2 \\ + \langle \alpha_{\upsilon 2}, \mathcal{V}_2 - |\mathcal{Q}_1| + |\mathcal{Q}_2|\rangle + \frac{\rho}{2}\|\mathcal{V}_2 - |\mathcal{Q}_1| - |\mathcal{Q}_2|\|_F^2\end{aligned} \tag{9}$$

The problem can be divided into two sub-problems in case $\mathcal{Q}_1 \geq 0$ or $\mathcal{Q}_1 < 0$. Each sub-problem is a quadratic programming problem and is easy to be solved. When $\mathcal{Q}_1 \geq 0$, its optima solution $\mathcal{Q}_{1+} = \frac{1}{2\rho}(\alpha_{q1} + \alpha_{\upsilon 2}) + \frac{1}{2}(\mathcal{W}_1 + \mathcal{V}_2 + \|\mathcal{Q}_2\|)$ if $\mathcal{Q}_{1+} \geq 0$; otherwise, $\mathcal{Q}_{1+} = 0$. Similarly, $\mathcal{Q}_{1-} = \frac{1}{2\rho}(\alpha_{q1} - \alpha_{\upsilon 2}) + \frac{1}{2}(\mathcal{W}_1 - \mathcal{V}_2 - \|\mathcal{Q}_2\|)$ if $\mathcal{Q}_{1-} < 0$; otherwise, $\mathcal{Q}_{1-} = 0$. The optima solution $\mathcal{Q}_1$ is chosen between $\mathcal{Q}_{1+}$ and $\mathcal{Q}_{1-}$ which has minimum loss.

When $2 \leq p \leq P-1$, the variable $\mathcal{Q}_p$ is learned as follows:

$$\begin{aligned}\mathcal{Q}_p = \arg\min_{\mathcal{Q}_p} \langle \alpha_{qp}, \mathcal{W}_p - \mathcal{Q}_p\rangle + \frac{\rho}{2}\|\mathcal{W}_p - \mathcal{Q}_p\|_F^2 \\ + \langle \alpha_{\upsilon p}, \mathcal{V}_p - |\mathcal{Q}_{p-1}| + |\mathcal{Q}_p|\rangle + \frac{\rho}{2}\|\mathcal{V}_p - |\mathcal{Q}_{p-1}| - |\mathcal{Q}_p|\|_F^2 \\ + \langle \alpha_{\upsilon p+1}, \mathcal{V}_{p+1} - |\mathcal{Q}_p| + |\mathcal{Q}_{p+1}|\rangle + \frac{\rho}{2}\|\mathcal{V}_{p+1} - |\mathcal{Q}_p| - |\mathcal{Q}_{p+1}|\|_F^2\end{aligned} \tag{10}$$

Although $\mathcal{Q}_p$ relates to both its previous and next time periods, its sub-problems are still quadratic programming problems. Similarly, $\mathcal{Q}_{p+} = \frac{1}{3\rho}(\alpha_{qp} - \alpha_{\upsilon p} + \alpha_{\upsilon p+1}) + \frac{1}{3}(\mathcal{W}_p - \mathcal{V}_p + \|\mathcal{Q}_{p-1}\| + \mathcal{V}_{p+1} + \|\mathcal{Q}_{p+1}\|)$, $\mathcal{Q}_{p-} = \frac{1}{3\rho}(\alpha_{qp} + \alpha_{\upsilon p} - \alpha_{\upsilon p+1}) + \frac{1}{3}(\mathcal{W}_p + \mathcal{V}_p - \|\mathcal{Q}_{p-1}\| - \mathcal{V}_{p+1} - \|\mathcal{Q}_{p+1}\|)$, and $\mathcal{Q}_p$ is chosen between $\mathcal{Q}_{p+}$ and $\mathcal{Q}_{p-}$ which has the minimum loss.

When $p = P$, the variable $\mathcal{Q}_P$ is learned as follows:

$$\begin{aligned}\mathcal{Q}_P = \arg\min_{\mathcal{Q}_P} \langle \alpha_{qP}, \mathcal{W}_P - \mathcal{Q}_P\rangle + \frac{\rho}{2}\|\mathcal{W}_P - \mathcal{Q}_P\|_F^2 \\ + \langle \alpha_{\upsilon P}, \mathcal{V}_P - |\mathcal{Q}_{P-1}| + |\mathcal{Q}_P|\rangle + \frac{\rho}{2}\|\mathcal{V}_P - |\mathcal{Q}_{P-1}| - |\mathcal{Q}_P|\|_F^2\end{aligned} \tag{11}$$

$\mathcal{Q}_P$ only relates to its previous period and its sub-problems are still quadratic programming problems. Similarly, $\mathcal{Q}_{P+} = \frac{1}{2\rho}(\alpha_{qP} - \alpha_{\upsilon P}) + \frac{1}{2}(\mathcal{W}_P - \mathcal{V}_P + \|\mathcal{Q}_{P-1}\|)$, $\mathcal{Q}_{P-} = \frac{1}{2\rho}(\alpha_{qP} + \alpha_{\upsilon P}) + \frac{1}{2}(\mathcal{W}_P + \mathcal{V}_P - \|\mathcal{Q}_{P-1}\|)$, and $\mathcal{Q}_P$ is selected between $\mathcal{Q}_{P+}$ and $\mathcal{Q}_{P-}$ which has the minimum loss.

5.3 Update $\mathcal{V}$

The auxiliary variable $\mathcal{V}$ is learned as follows:

$$\begin{aligned}\mathcal{V} = \arg\min_{\mathcal{V}} \sum_{p=2,l,f}^{P,L,F} \langle \alpha_{\upsilon\{f,l,p\}}, \mathcal{V}_{\{f,l,p\}} - |\mathcal{Q}_{\{f,l,p-1\}}| - |\mathcal{Q}_{\{f,l,p\}}|\rangle \\ + \frac{\rho}{2}\sum_{p=2,l,f}^{P,L,F} \|\mathcal{V}_{\{f,l,p\}} - |\mathcal{Q}_{\{f,l,p-1\}}| - |\mathcal{Q}_{\{f,l,p\}}|\|_F^2 + \max(\mathcal{V}_{\{f,l,p\}}, 0)\end{aligned} \tag{12}$$

For notation simplicity, we define $\mathcal{V}_p = \mathcal{V}_{\{f,l,p\}}$ and $\mathcal{Q}_p = \mathcal{Q}_{\{f,l,p\}}$ when feature f and location l are fixed. The problem can be divided into the combination of subproblems for each $\mathcal{V}_p (p \geq 2)$:

$$\begin{aligned}\mathcal{V}_p = \arg\min_{\mathcal{V}_p} \langle \alpha_{\upsilon p}, \mathcal{V}_p - |\mathcal{Q}_{p-1}| + |\mathcal{Q}_p|\rangle \\ + \frac{\rho}{2}(\alpha_{\upsilon p}, \mathcal{V}_p - |\mathcal{Q}_{p-1}| + |\mathcal{Q}_p|) + \lambda_1 \max(\mathcal{V}_p, 0)\end{aligned} \tag{13}$$

We define the function g and f as:

$$\begin{aligned}g(\mathcal{V}_p) = \lambda_1 \max(\mathcal{V}_p, 0) \\ f(\mathcal{V}_p) = \langle \alpha_{\upsilon p}, \mathcal{V}_p - |\mathcal{Q}_{p-1}| + |\mathcal{Q}_p|\rangle + \frac{\rho}{2}(\alpha_{\upsilon p}, \mathcal{V}_p - |\mathcal{Q}_{p-1}| + |\mathcal{Q}_p|)\end{aligned} \tag{14}$$

The problem can be considered as the following iterative procedure known as ISTA[27]:

$$\mathcal{V}_p^{k+1} = \arg\min_{\mathcal{V}_p} g(\mathcal{V}_p) + \frac{1}{2\eta}\|\mathcal{V}_p - (\mathcal{V}_p^k - \eta \bigtriangledown f(\mathcal{V}_p^k)\|_2^2 \tag{15}$$

where k is the k-th iteration and η is the step size. In our case where $g(\mathcal{V}_p) = \lambda_1 \max(\mathcal{V}_p, 0)$, the one-dimensional problem can be solved by the following theorem:

THEOREM 5.1. *For the iterative shrinkage-thresholding problem of $g(x) = \max(x, 0)$, which is defined as follows:*

$$\min_x \left\{ \lambda \max(x, 0) + \frac{1}{2\eta}(x - x_0)^2 \right\}$$

where $x_0 = x^k - \eta \bigtriangledown f(x^k)$. The shrinkage operator of this problem equals to:

$$S_{\lambda\eta}(x_0) = \begin{cases} x_0 - \lambda\eta, & \text{if } x_0 > \lambda\eta \\ x_0, & \text{if } x_0 < 0 \\ 0, & \text{otherwise} \end{cases}$$

PROOF. if $x > 0$, the problem converts to $\min_x \lambda x + \frac{1}{2\eta}(x - x_0)^2$, its analytical solution is $x = x_0 - \lambda\eta$. When $x < 0$, the problem is $\min_x \frac{1}{2\eta}(x - x_0)^2$, whose analytical solution is $x = x_0$. Otherwise, $x = 0$. □

The fast iterative shrinkage-thresholding version can be found in Appendix A.

5.4 Update $\mathcal{R}$

The auxiliary variable $\mathcal{R}$ is learned as follows:

$$\mathcal{R} = \arg\min_{\mathcal{R}} \lambda_2 \|\mathcal{R}_{(f)}\|_1 + \langle \alpha_r, \mathcal{W} - \mathcal{R}\rangle + \frac{\rho}{2}\|\mathcal{W} - \mathcal{R}\|_F^2 \tag{16}$$

The problem can be solved by the soft-thresholding operator of L1 norm. Its analytical solution is:

$$\mathcal{R} = \text{fold}_f[\mathbf{S}_{\lambda_1/\rho}(\mathcal{W}_{(f)} + \frac{\alpha_r}{\rho})] \tag{17}$$

where $\mathbf{S}_\lambda(a) = sgn(a)(|a| - \lambda)_+$ and fold_f is the matrix describing tensor folding in its feature dimension.

5.5 Update $\mathcal{X}$

The data tensor $\mathcal{X}$ is learned as follows:

$$\begin{aligned}\mathcal{X} = \arg\min_{\mathcal{X}} \mathcal{L}(\mathcal{X}) + \sum_{i=1}^{K} \langle \beta_i, \mathcal{X} - \mathcal{M}_i \rangle + \frac{\rho}{2} \sum_{i=1}^{K} \|\mathcal{X} - \mathcal{M}_i\|_F^2 \\ + \langle \alpha_u, \mathcal{X} - \mathcal{U} \rangle + \frac{\rho}{2} \|\mathcal{X} - \mathcal{U}\|_F^2 + \langle \alpha_s, \mathcal{X} - \mathcal{S} \rangle + \frac{\rho}{2} \|\mathcal{X} - \mathcal{S}\|_F^2 \\ s.t. \mathcal{X}_\Omega = \mathcal{T}_\Omega\end{aligned} \tag{18}$$

Similar to the weight tensor $\mathcal{W}$, this is also a generalized logistic regression with a least squares loss function. $\mathcal{X}_{\bar{\Omega}}$ can be solved by Newton method, and all the remaining values in $\mathcal{X}_\Omega$ are equal to $\mathcal{T}_\Omega$.

5.6 Update $\mathcal{M}_i$

Each auxiliary variable $\mathcal{M}_i$ is learned as follows:

$$\mathcal{M}_i = \arg\min_{\mathcal{M}_i} \theta_1 \sum_{i=1}^{K} \|\mathcal{M}_{i(i)}\|_* + \langle \beta_i, \mathcal{X} - \mathcal{M}_i \rangle + \frac{\rho}{2} \|\mathcal{X} - \mathcal{M}_i\|_F^2 \tag{19}$$

The optimal solution to $\mathcal{M}_i$ can be obtained by the soft thresholding method [16]. Specifically, the analytical solution is:

$$\mathcal{M}_i = \text{fold}_i[\mathbf{D}_{\frac{\theta_1}{\rho}}(X_(i) + \frac{1}{\rho}\beta_{i(i)})] \tag{20}$$

where $D_\tau(\cdot) = U\Sigma_+V^T$ and the i-th diagonal element of Σ_+ is $\max(0, \sigma_i - \tau)$. Suppose that the singular vector decomposition of matrix $\cdot$ is $U\Sigma_+V^*$, and denote the i-th diagonal element of Σ_+ by σ_i.

5.7 Update $\mathcal{U}$

The auxiliary variable $\mathcal{U}$ is learned as follows:

$$\begin{aligned}\mathcal{U} = \arg\min_{\mathcal{U}} \theta_2 \sum_{n=1}^{N_x-1} \|\mathcal{U}_{\{n+1\}} - \mathcal{U}_{\{n\}}\|_F^2 \\ + \langle \alpha_u, \mathcal{X} - \mathcal{U} \rangle + \frac{\rho}{2} \|\mathcal{X} - \mathcal{U}\|_F^2\end{aligned} \tag{21}$$

To simplify the derivation, we introduce an auxiliary matrix $D_{l,n}$ to represent $\|\mathcal{U}_{\{n+1\}} - \mathcal{U}_{\{n\}}\|_F^2$ as $\|\mathcal{U}_{(f)}D_{l,n}\|_F^2$, where $\mathcal{U}_{(f)}$ is the unfolded matrix in the feature dimension. The analytical solution of $\mathcal{U}$ is $\text{fold}_f[(\alpha_u + \rho\mathcal{X}_{(f)})(2\theta_2 \sum_{n=1}^{N_x-1} D_n D_n^T + \rho * I)^{-1}]$, where $D_{\{l,n\}}$ is formally defined as:

$$D_{n,l_{(ij)}} = \begin{cases} -1, & \text{if } i = j + L(n-1) \\ 1, & \text{if } i = j + Ln \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

5.8 Lagrangian Multipliers and Stop Condition

The Lagrangian multiplier α_r is updated as follows:

$$\begin{aligned}&\alpha_q \leftarrow \alpha_q + \rho(\mathcal{W} - \mathcal{Q}),\ \alpha_r \leftarrow \alpha_r + \rho(\mathcal{W} - \mathcal{R}) \\ &\alpha_{v\{f,l,p\}} \leftarrow \alpha_{v\{f,l,p\}} + \rho(|\mathcal{Q}_{\{f,l,p-1\}}| - |\mathcal{Q}_{\{f,l,p\}}|) \\ &\beta_i \leftarrow \beta_i + \rho(\mathcal{X} - \mathcal{M}_i),\ \alpha_u \leftarrow \alpha_u + \rho(\mathcal{X} - \mathcal{U}) \\ &\alpha_s \leftarrow \alpha_s + \rho(\mathcal{X} - \mathcal{S})\end{aligned} \tag{23}$$

The stop condition is determined by primal and dual residuals of the $(k+1)$th iteration, which are calculated based on the following theorem. The parameters in the theorem labeled with superscript k (e.g., $\mathcal{W}^k$) represent to its corresponding value in the k^{th} iteration.

THEOREM 5.2. *The primal residual and dual residual of the algorithm are as follows:*

- *Primal residual of objective function.*

$$\begin{aligned}r = &\|\mathcal{W} - \mathcal{Q}\|_F + \|\mathcal{W} - \mathcal{R}\|_F + \|\mathcal{X} - \mathcal{U}\|_F \\ &+ \sum_{p=2,l,f}^{P,L,F} \|\mathcal{V}_{\{f,l,p\}} - |\mathcal{Q}_{\{f,l,p-1\}}| - |\mathcal{Q}_{\{f,l,p\}}|\|_F \\ &+ \sum_{i=1}^{K} \|\mathcal{X} - \mathcal{M}_i|\|_F + \|\mathcal{X} - \mathcal{S}\|_F\end{aligned} \tag{24}$$

- *Dual residual of objective function*

$$\begin{aligned}s = &\rho(\|\mathcal{Q}^k - \mathcal{Q}^{k+1}\|_F + \|\mathcal{V}^k - \mathcal{V}^{k+1}\|_F + \|\mathcal{R}^k - \mathcal{R}^{k+1}\|_F \\ &\|\sum_{i=1}^{K}(\mathcal{M}_i^k - \mathcal{M}_i^{k+1}) + \mathcal{U}^k - \mathcal{U}^{k+1} + \mathcal{S}^k - \mathcal{S}^{k+1}\|_F)\end{aligned} \tag{25}$$

The proof of Theorem 5.2 can be found in Appendix C.

6 EXPERIMENTAL EVALUATION

In this section, the performance of the proposed model HPEF is evaluated using 6 real-world datasets from different domains. First, the experimental setup is introduced in Section 6.1, after which the effectiveness and efficiency of our model is evaluated against several existing methods for a number of different data missing scenarios in Section 6.2. All the experiments were conducted on a 64-bit machine with Intel(R) core(TM) quad-core processor (i7CPU@ 3.60GHz) and 32.0GB memory.

6.1 Experimental Setup

6.1.1 Datasets and Labels. In this paper, 6 different datasets from different domains were used for the experimental evaluations, as shown in Table 2. Among these, two datasets composed of commodity price data were used to forecast civil unrest events for two different countries in South America. The commodity price data were provided by Premise[4], who collected their price data via mobile phones from their network of data contributors located in 30 countries. Data for the period from September 1, 2013 to May 4, 2014 were used for training, while June 1, 2014 to February 1, 2015 data was used for the performance evaluation. The resulting event forecast were validated against a labeled civil unrest event set, referred to as the gold standard report (GSR), which was exclusively provided by MITRE [28]. The GSR is a collection of civil unrest news reports from the most influential newspaper outlets in Latin

[4]https://www.premise.com/

America [29]. An example of a labeled GSR event is given by the tuple: (STATE="Para", COUNTRY="Brazil", DATE="12/2014").

The other four datasets were collected to track real estate values in the United States based on short-term rental prices. The rental prices were collected from the company Airbnb[5] for four cities in the United States, as shown in Table 2. Residential properties were divided into 46 different categories according to their numbers of bedrooms, bathrooms and their other facilities, which served as the features for our model. Within each city, its different neighborhoods were used as spatial locations. For example, 16 neighborhoods were selected in New York as shown in Table 2. The data for the first 60% of the time period covered was used for training, while the remaining 40% was used for the performance evaluation. The forecasting results for the trends in house values were validated against the actual real estate values reported by Zillow[6] for the same period. An example of a house value change event is: (Neighborhood="Glendale", City="Los Angeles", Date="07/2016", Event="Raise/Down").

Table 2: Characteristics of datasets used (CU=civil unrest; RE=real estate)

Dataset	Domain	Locations	Time Period	Missing Data
Argentina	CU	4	09/2013 - 02/2015	73.35%
Brazil	CU	10	09/2013 - 02/2015	68.82%
New York	RE	16	01/2015 - 07/2016	49.61%
Los Angeles	RE	13	05/2015 - 08/2016	56.19%
New Orleans	RE	4	06/2015 - 06/2016	65.32%
San Francisco	RE	5	05/2015 - 07/2016	68.18%

6.1.2 Parameter Settings and Metrics. There are 5 tunable parameters in the proposed HPEF model, which could be divided into two groups, namely feature weight parameters $\lambda_{1,2}$ and data completion parameters $\theta_{1,2,3}$. Based on a 10-fold cross validation on the training set, these were set as $\lambda_1 = 0.2$, $\lambda_2 = 0.1$, $\theta_1 = 0.2$, $\theta_2 = 0.3$, and $\theta_3 = 0.3$.

In the experiment, the event forecasting task was to predict whether or not there would be an event during the next time step for a specific location. For the civil unrest datasets, the time step was set at one week and the location as a city. For the real estate dataset, the time step was set at one month and the location as a neighborhood. To validate the prediction performance, different metrics were adopted: the True Positive Ratio (TPR) designates the percentage of positive predictions that successfully match the events that truly happen, while the False Positive Ratio (FPR) denotes the percentage of positive predictions that are actually false alarms. A receiver operating characteristic (ROC) curve was also utilized to evaluate the forecasting performance as its discrimination threshold for each predictive model was varied. Finally, an area under ROC curve (AUC) was also applied as a comprehensive measure of forecasting performance.

6.1.3 Comparison Methods. The following methods are included in the performance comparison presented here: 1) Logistic regression (LR) [25]. For each location, LR utilizes a logit function to map the price observations into future event occurrences; 2) LASSO [30]. Different LASSO models are built for corresponding locations. The regularization parameter is set as 0.2 based on a 10-fold cross validation on the training set; 3) Multitask Learning(MTL) [9]. In the multi-task model, each task represents the forecast for each location. The regularization parameters $\lambda_1 = 0.02$ and $\lambda_2 = 0.005$ are set based on a 10-fold cross-validation; 4) Ordered LASSO (OL)[22]. In ordered LASSO, the time-lagged order constraint is applied on the continuous price data. The regularization parameter $\lambda = 0.1$ is set based on cross-validation; 5) LASSO-TC and 6) OL-TC are LASSO and OL methods based on the tensor completion [16] data with low-rank constraint, respectively.

[5]https://www.airbnb.com
[6]http://www.zillow.com/

6.2 Performance

In this section, the effectiveness in terms of the AUC and ROC curves and the runtime efficiency are analyzed for all six of the comparison methods and the results compared to those obtained using HPEF.

6.2.1 Event forecasting performance on AUC. Table 3 compares the effectiveness and robustness achieved by the different methods when forecasting events with different missing data ratios. The AUC measure has been adopted to quantify the performance and the original percentages of missing data are underscored. The missing data ratio was then manually increased by randomly reducing the number of price data points in both the spatial and temporal dimensions.

The results shown in Table 3 demonstrate that the methods that take into account the temporal dependence of price and spatiotemporal missing values performed better. Specifically, the performance of HPEF outperformed the other methods for nearly all of the different missing data ratios. LASSO-TC and OL-TC also performed competitively with high missing ratios in three datasets. Looking across the different missing data ratios, LASSO-TC and OL-TC achieved better robustness against missing values than their original versions, LASSO and OL, that do not consider missing values. For example, the performance of LASSO dropped an average 25% when the missing data ratio increased by 20% for most datasets. In contrast, HPEF was able to handle the missing value problem in multiple data sources, dropping on average less than 10% when the missing data ratio increased by more than 20%. MTL was also not particularly sensitive to the change in missing values, largely due to its ability to handle the lack of data by sharing the information across different tasks. In all, HPEF outperformed all the other methods in 4 out of the 6 datasets for all the different missing data ratios by 7% on average, and achieved the second best performance on the other 2 datasets. This is because HPEF is capable of handling the two crucial challenges, namely temporal price dependence and high missing value ratios effectively.

6.2.2 Efficiency on running time. The right hand column of each dataset in Table 3 shows the training time efficiency comparison for HPEF and the six competing methods. The efficiency evaluation results for the other missing data ratios follow a similar pattern to that shown in Table 3 and are not provided due to space limitations. The running times on the test set for all the comparison methods were effectively instant (i.e., less than 0.1 millisecond for each prediction), so these are also not provided here. According to Table 3, the

Table 3: Event forecasting performance based on area under the curve (AUC) of ROC

	Argentina(CU)				Brazil(CU)				New York(RE)			
	73%	85%	95%	RT(ms)	73%	85%	95%	RT(ms)	50%	75%	90%	RT(ms)
LR	0.577	0.541	0.375	**3.02**	0.505	0.474	0.499	**4.20**	0.338	0.394	0.403	**8.75**
LASSO	0.577	0.553	0.366	3.83	0.588	0.562	0.491	35.84	0.456	0.326	0.354	66.46
MTL	0.579	0.568	0.410	5.10	0.528	0.498	0.506	6.10	0.343	0.373	0.361	19.79
OL	0.634	0.601	0.376	151.81	0.550	0.524	0.416	217.33	0.380	0.359	0.365	97.29
LASSO-TC	0.569	0.578	0.451	10.51	0.586	0.605	0.501	25.47	0.449	0.354	0.384	66.67
OL-TC	0.651	0.612	0.431	153.31	0.601	0.611	0.540	215.31	0.417	0.433	**0.443**	115.21
HPEF	**0.723**	**0.621**	**0.589**	237.26	**0.653**	**0.645**	**0.613**	388.68	**0.491**	**0.452**	0.425	265.23
	Los Angeles(RE)				**New Orleans(RE)**				**San Francisco(RE)**			
	56%	80%	90%	RT(ms)	65%	80%	95%	RT(ms)	68%	80%	95%	RT(ms)
LR	0.577	0.479	0.441	**7.42**	0.582	0.481	0.449	**8.58**	0.530	0.555	0.424	**5.53**
LASSO	0.554	0.486	0.390	66.92	0.583	0.491	0.476	35.17	0.564	0.558	0.435	35.38
MTL	0.557	0.520	0.444	9.97	0.577	0.543	0.572	16.50	0.543	0.543	0.521	6.75
OL	0.523	0.506	0.426	20.00	0.574	0.506	0.504	282.33	0.466	0.504	0.501	206.13
LASSO-TC	0.511	0.463	0.526	68.21	0.635	**0.632**	0.617	22.08	0.572	0.530	0.446	38.88
OL-TC	0.520	0.516	0.529	19.74	0.567	0.563	0.495	265.58	0.520	0.521	0.510	93.38
HPEF	**0.571**	**0.566**	**0.558**	132.45	**0.696**	0.628	**0.598**	374.13	**0.632**	**0.607**	**0.586**	265.81

running time (RT) of the LR method was around 6.3 milliseconds per prediction, outperforming all the other methods. OL and HPEF both required hundreds of milliseconds on each prediction. However, the running times achieved by these methods were all well below 11 hours for a 2-year-long training sets for both the week and month-wise event forecasting tasks, making this eminently practical for real-world applications.

6.2.3 Event forecasting performance on ROC curves. Figure 2 illustrates the event forecasting performance ROC curves for 6 datasets in two domains, namely civil unrest and real estates. For the 4 real estate datasets shown in Figures 2(d)-(f), HPEF performs the best overall, with ROC curves covering the largest area above the axis. The ROC curves for HPEF are consistently above those of the other methods when FPR is less than 0.4 in datasets including New York, New Orleans, and San Francisco. For the Brazilian dataset, OL-TC performs best when FPR is smaller than 0.5, while HPEF outperforms the other methods when $FPR > 0.5$. OL also achieves quite competitive performances for the Argentina dataset by an apparent margin for a high FPR. MTL generally achieves a limited performance, but its performance is robust against missing data, as can be seen in Table 3.

7 CONCLUSION

In this paper, a novel spatiotemporal event forecasting model based on hyper-local price data has been proposed to characterize temporal price dependence, accommodate spatiotemporal missing values, and effectively apply economic domain knowledge. To achieve these goals, we designed a soft time-lagged feature selection model combined with price data completion based on economic domain knowledge. An efficient algorithm for parameter optimization is proposed to solve the optimization problem. Extensive experiments on 6 real-world datasets with multiple data sources demonstrated that the proposed model outperforms other comparable methods for different ratios of missing values. One of our current directions of future work is to extend these methods to develop a continuous-valued indicator of the societal phenomena of interest leveraging the underlying hyper-local data.

A FISTA OF V UPDATE

For the fast iterative shrinkage-thresholding algorithm (FISTA)[27] of $g(x) = \max(x, 0)$ can be solved as follows.

$$
\begin{aligned}
y_{s+1} &= \arg\min_x \lambda \max(x,0) + \frac{\beta}{2}(x - x'_s)^2 \\
x_{s+1} &= (1-\gamma_s)y_{s+1} + \gamma_s y_s \\
where \quad \lambda_0 = 0, \lambda_s &= \frac{1+\sqrt{1+4\lambda_{s-1}^2}}{2} \\
\gamma_s &= \frac{1-\lambda_s}{\lambda_{s+1}}, x'_s = x_s - \eta \bigtriangledown f(x_s)
\end{aligned}
$$

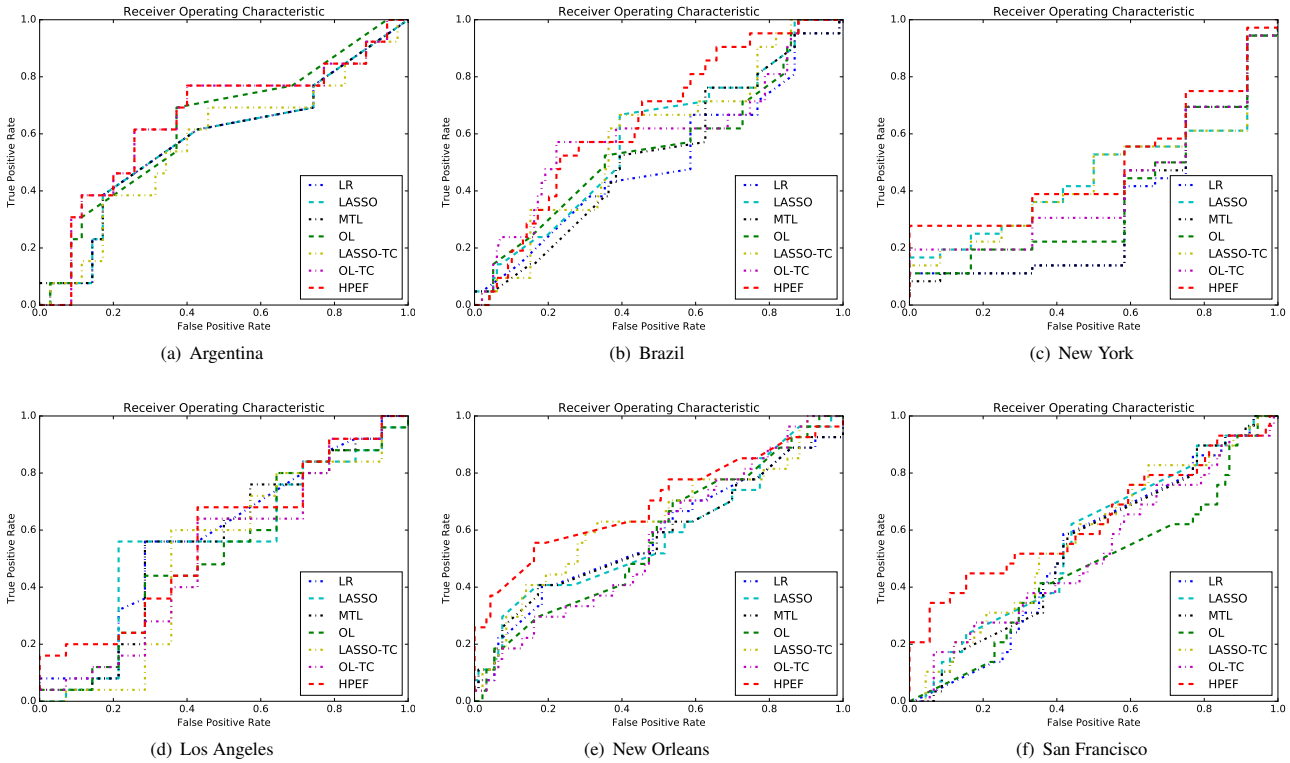
where s is the iteration number. Its shrinkage operator is:

$$
\begin{aligned}
y_{s+1} &= \begin{cases} x'_s - \frac{\lambda}{\beta}, & \text{if } x'_s > \frac{\lambda}{\beta} \\ x'_s, & \text{if } x'_s < 0 \\ 0, & \text{otherwise} \end{cases} \\
x_{s+1} &= (1-\gamma_s)y_{s+1} + \gamma_s y_s
\end{aligned}
$$

B DERIVATION OF $\mathcal{S}$

The auxiliary variable $\mathcal{S}$ is learned as follows:

$$
\begin{aligned}
\mathcal{S} = \arg\min_{\mathcal{S}} \theta_3 \sum_{f=1}^{F} \sum_{t=1}^{T} \Psi(\mathcal{X}_{\{f,t\}}) \\
+ \langle \alpha_s, \mathcal{X} - \mathcal{S} \rangle + \frac{\rho}{2} \|\mathcal{X} - \mathcal{S}\|_F^2
\end{aligned} \tag{26}
$$



(a) Argentina (b) Brazil (c) New York

(d) Los Angeles (e) New Orleans (f) San Francisco

Figure 2: Receiver operating characteristic (ROC) curves for the performance on different datasets

The spatial coherence function $\Psi(x)$ is:

$$\Psi(x) = \sum_{i=1}^{L} \sum_{i,j\in\Gamma}^{L} \max(x_i - x_j - C_{j,i}, 0) + \max(x_j - x_i - C_{i,j}, 0)$$

where Γ set is defined as $\Gamma = \{i, j | i \neq j, s_j \neq 0\}$. For notation simplicity, we define $s_i = S_{\{f,i,n\}}$, $c_{ji} = C_{\{j,i\}}$, $\alpha_{s_i} = \alpha_{s\{f,i,n\}}$, $x_i = \mathcal{X}_{\{f,i,n\}}$. So for each s_i, its augmented Lagrangian function is:

$$\mathcal{L}_{s_i} = \theta_3 \sum_{i,j\in\Gamma}^{L} \{\max(s_i - s_j - c_{ji}, 0) + \max(s_j - s_i - c_{ij}, 0)\} + \langle \alpha_{s_i}, x_i - s_i \rangle + \frac{\rho}{2} \|x_i - s_i\|_F^2$$

Introducing two auxiliary sets $\Gamma_{in} = \{i, j | s_i - s_j - c_{ji} \geq 0, s_j \neq 0, i \neq j\}$ and $\Gamma_{out} = \{i, j | s_j - s_i - c_{ij} \geq 0, s_j \neq 0, i \neq j\}$, the objective function can be rewritten as:

$$\begin{aligned} \mathcal{L}_{s_i} =& \theta_3 \sum_{i,j\in\Gamma_{in}}^{L} (s_i - s_j - c_{ji}) + \theta_3 \sum_{i,j\in\Gamma_{out}}^{L} (s_j - s_i - c_{ij}) \\ &+ \langle \alpha_{s_i}, x_i - s_i \rangle + \frac{\rho}{2} \|x_i - s_i\|_F^2 \\ =& \frac{\rho}{2} s_i^2 + (\theta_3 \|\Gamma_{in}\| + \theta_3 \|\Gamma_{out}\| - \alpha_{s_i} - \rho x_i) s_i + B \end{aligned} \tag{27}$$

where $\|\cdot\|$ is the size of set and B is the constant in regardless of s_i. Considering the property of equation (27), the problem can be solved by Theorem B.3.

LEMMA B.1. *The value of s_i is directly proportional to $\|\Gamma_{in}\|$ and inversely proportional to $\|\Gamma_{out}\|$, when fixing the value of s_j.*

PROOF. Because the possible value set of s_j is a finite set of discrete number, where $j = 1, .., L(j \neq i)$, when increasing s_i, the number of $j \in \{1..L\}$ that satisfies the condition $s_i - s_j - c_{ji}$ increases. So s_i is directly proportional to $\|\Gamma_{in}\|$. Similarly, s_i is inversely proportional to $\|\Gamma_{out}\|$. □

LEMMA B.2. *The domain D of s_i can be divided into N finite sub-domains, and for each sub-domain D_n, $\forall s_i \in D_n$ map to the same values of $\|\Gamma_{in}\|$ and $\|\Gamma_{out}\|$.*

PROOF. As s_i is proportional to $\|\Gamma_{in}\|$ and the value of $\|\Gamma_{in}\|$ is finite, the domain D of s_i can be divided into finite sub-domains $D^{in} = \{D_1^{in}, ..., D_p^{in}\}$. For each sub-domain $D_p^{in} \in D^{in}$, $\forall s_i \in D_p^{in}$ has the same $\|\Gamma_{in}\|$ value. Similarly, $\forall s_i \in D_p^{out}$ has the same $\|\Gamma_{out}\|$ value. Because all the sub-domains in D^{in} increases and sub-domains in D^{out} decreases, there are finite combination of the domain D^{in} and D^{out}, in which s_i maps to the same $\|\Gamma_{in}\|$ and $\|\Gamma_{out}\|$ values. □

THEOREM B.3. *Minimization of the equation* (27) *can be solved by minimizing the values of its finite convex subproblems.*

PROOF. As s_i can be divided into finite domains with corresponding $\|\Gamma_{in}\|$ and $\|\Gamma_{out}\|$, the problem of equation (27) can be divided into finite subproblems according to the domain division. Each problem with fixed $\|\Gamma_{in}\|$ and $\|\Gamma_{out}\|$ is a quadratic programming problem which is convex, thus the minimum value of each subproblems is the solution of equation (27). □

C PROOF OF THEOREM 5.5

PROOF. The primal residual can be easily deduced from the primal feasibility according to the objective function directly. The deduction of the dual residual is elaborated in the following. The dual feasibility of the objective function is listed in the order $\mathcal{W}$, $\mathcal{Q}$, $\mathcal{V}$, $\mathcal{R}$, $\mathcal{X}$, $\mathcal{M}_i$, $\mathcal{U}$, $\mathcal{S}$:

$$\begin{aligned} &0 \in \partial\mathcal{L}(\mathcal{W}) + \alpha_q^* + \alpha_r^*, 0 \in \alpha_q^* + \alpha_v^*, 0 \in \alpha_v^*, 0 \in \alpha_r^* \\ &0 \in \sum_{i=1}^{K} \beta_i^* + \alpha_u^* + \alpha_s^*, 0 \in \sum_{i=1}^{K} \beta_i^*, 0 \in \alpha_u^*, 0 \in \alpha_s^* \end{aligned} \tag{28}$$

where the variables with superscript $*$ denote the optimal solutions. Notice that the auxiliary variables always satisfy the dual feasibility requirement, therefore they have no dual residual. For variable $\mathcal{X}$, we know that:

$$\begin{aligned} 0 \in &\sum_{i=1}^{K} \beta_i^* + \alpha_u^* + \alpha_s^* = \sum_{i=1}^{K} (\beta_i^k + \rho(\mathcal{X}^{k+1} - \mathcal{M}_i^k)) \\ &+ \alpha_u^k + \rho(\mathcal{X}^{k+1} - \mathcal{U}^k) + \alpha_s^k + \rho(\mathcal{X}^{k+1} - \mathcal{S}^k) \\ \in &\sum_{i=1}^{K} (\beta_i^{k+1} + \rho(\mathcal{M}_i^{k+1} - \mathcal{M}_i^k)) + \alpha_u^{k+1} + \rho(\mathcal{U}^{k+1} - \mathcal{U}^k) \\ &+ \alpha_s^{k+1} + \rho(\mathcal{S}^{k+1} - \mathcal{S}^k) \end{aligned} \tag{29}$$

Therefore, the dual residual of $\mathcal{X}$ is $\rho\|\mathcal{M}_i^k - \mathcal{M}_i^{k+1}) + \mathcal{U}^k - \mathcal{U}^{k+1} + \mathcal{S}^k - \mathcal{S}^{k+1}\|_F$. Similarly, the same method can be applied to prove the dual residuals with respect to other parameters. □

REFERENCES

[1] Henk-Jan Brinkman and Cullen S Hendrix. *Food insecurity and violent conflict: Causes, consequences, and addressing the challenges*. World Food Programme, 2011.

[2] Joe Weinberg, Ryan Bakker, et al. Let them eat cake: Food prices, domestic policy and social unrest. *Conflict Management and Peace Science*, 32(3):309–326, 2015.

[3] Oeindrila Dube and Juan Vargas. Commodity price shocks and civil conflict: Evidence from colombia*. *The Review of Economic Studies*, 2013.

[4] World Bank Group. *World Development Indicators 2012*. World Bank Publications, 2012.

[5] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[6] Liang Zhao, Jiangzhuo Chen, Feng Chen, Wei Wang, Chang-Tien Lu, and Naren Ramakrishnan. Simnest: Social media nested epidemic simulation via online semi-supervised deep learning. In *Proceedings of the 2015 IEEE International Conference on Data Mining (ICDM)*, ICDM '15, pages 639–648, Washington, DC, USA, 2015. IEEE Computer Society.

[7] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[8] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010.

[9] Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1503–1512, New York, NY, USA, 2015. ACM.

[10] Matthew S. Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115 – 125, 2014.

[11] Susan E Hardy, Heather Allore, and Stephanie A Studenski. Missing data: a special challenge in aging research. *Journal of the American Geriatrics Society*, 57(4):722–729, 2009.

[12] Sujuan Gao. A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data. *Statistics in Medicine*, 23(2):211–219, 2004.

[13] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, 11:2287–2322, August 2010.

[14] Jin Huang, Feiping Nie, Heng Huang, Yu Lei, and Chris Ding. Social trust prediction using rank-k matrix recovery. IJCAI '13, pages 2647–2653. AAAI Press, 2013.

[15] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theor.*, 56(5):2053–2080, May 2010.

[16] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.

[17] M. Signoretto, R. Van de Plas, B. De Moor, and J. A. K. Suykens. Tensor versus matrix completion: A comparison with application to spectral data. *IEEE Signal Processing Letters*, 18(7):403–406, July 2011.

[18] Hua Wang, Feiping Nie, and Heng Huang. Low-rank tensor completion with spatio-temporal consistency. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 2846–2852. AAAI Press, 2014.

[19] Frank Harrell. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[20] Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015. PMID: 26759522.

[21] Heewon Park and Fumitake Sakaori. Lag weighted lasso for time series model. *Computational Statistics*, 28(2):493–504, 2013.

[22] Robert Tibshirani and Xiaotong Suo. An ordered lasso and sparse time-lagged regression. *Technometrics*, 58(4):415–423, 2016.

[23] L. Eldn. *Matrix Methods in Data Mining and Pattern Recognition*. Society for Industrial and Applied Mathematics, 2007.

[24] Owen A Lamont and Richard H Thaler. Anomalies: The law of one price in financial markets. *The Journal of Economic Perspectives*, 17(4):191–202, 2003.

[25] Ryan Compton, Craig Lee, Jiejun Xu, Luis Artieda-Moncada, Tsai-Ching Lu, Lalindra De Silva, and Michael Macy. Using publicly visible social media to build detailed forecasts of civil unrest. *Security Informatics*, 3(1):1–10, 2014.

[26] Wang Chong, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1903–1910. IEEE, 2009.

[27] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[28] MITRE. https://www.mitre.org/, September 2016.

[29] Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, and Saraf et. al. 'beating the news' with embers: Forecasting civil unrest using open source indicators. KDD '14, pages 1799–1808, New York, NY, USA, 2014. ACM.

[30] Joseph O Ogutu, Torben Schulz-Streeck, and Hans-Peter Piepho. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In *BMC proceedings*, volume 6, page 1. BioMed Central, 2012.