Neurocomputing xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Heterogeneous neural metric learning for spatio-temporal modeling of infectious diseases with incomplete data

Qi Tan [a,b], Yang Liu [a,b], Jiming Liu [a,b,*], Benyun Shi [b,f], Shang Xia [b,c,d,e], Xiao-Nong Zhou [b,c,d,e]

[a] Hong Kong Baptist University, Kowloon Tong, Hong Kong, PR China
[b] CDC-NIPD & HKBU-CSD Joint Research Laboratory for Intelligent Disease Surveillance and Control, Shanghai, PR China
[c] National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention, Shanghai, PR China
[d] Key Laboratory of Parasite and Vector Biology, MOH, Shanghai, PR China
[e] WHO Collaborating Center for Malaria, Schistosomiasis and Filariasi, Shanghai, PR China
[f] School of Computer Science and Technology, Nanjing Tech University, Nanjing, Jiangsu, PR China

ARTICLE INFO

Article history:
Received 15 July 2019
Revised 25 October 2019
Accepted 24 December 2019
Available online xxxx

Keywords:
Spatio-temporal modeling
incomplete-data
heterogeneous data sources
infectious disease
kernel method
metric learning
heterogeneous neural metric learning (HNML)

ABSTRACT

Infectious disease data, recording the numbers of infection cases in different locations and time, is one of the most typical categories of spatio-temporal data and plays an important role in the infectious disease control and prevention. However, due to the insufficient resources and manpower, the observations and records of infection cases are inevitably missing in some locations and time, which brings difficulties to the accurate risk assessment and timely disease control. Imputing the missing infectious disease data is challenging as the infectious disease diffusion can be potentially caused and affected by many risk factors. To address the above-mentioned challenges, a novel machine learning method, Heterogeneous Neural Metric Learning (HNML), is developed to restore the integrity of case reporting data using both the incomplete reported cases and the underlying disease-related risk factors from heterogeneous data sources. We empirically validate the effectiveness of our developed method on a representative infectious disease, malaria. We test the developed method under three common real-life data missing patterns with different levels of missing rates. By incorporating the disease-related risk factors as external resources through the proposed HNML method, we demonstrate significant accuracy improvement over the baseline and state-of-the-art inference methods for predicting unobserved malaria cases based on the incomplete reporting data. The results suggest that the disease-related risk factors can provide valuable information about the transmission patterns of infectious diseases and should be taken into account when implementing the surveillance.

1. Introduction

Data with spatial and temporal attributes are very popular in many real-world applications, such as epidemiology [1], crowd flow [2], and air quality [3]. However, such spatio-temporal data are often incomplete due to various reasons, such as high cost of collection [4], sensor failure [5], or unstable transmission [6]. The data incompleteness makes it difficult to monitor and analyze the real-world spatio-temporal dynamics.

Infectious disease data, recording the numbers of infection cases in different locations and time, represents one of the most typical categories of spatio-temporal data and plays an important role in many applications of public health, such as hotspot detection [7], infectious disease risk forecast [8], and transmission network mining [9]. However, due to the insufficient resources and manpower, the observations and records of infection cases are inevitably missing in some locations and time. Taking malaria, one of most serious infectious diseases, as an example. For eliminating malaria, the WHO has called for countries to establish nation-wide epidemiological intelligence strategies to engage in effective surveillance for malaria early detection and prevention [10], which requires lots of experienced public health workers. However, the human resources are very insufficient particularly in remote and poor regions [11], making infection case data missing in some locations and time. The consequence of incompleteness in infectious disease data could be serious as the missing data bring difficulties to the accurate risk assessment and timely disease control.

* Corresponding author at: Hong Kong Baptist University, Kowloon Tong, Hong Kong, PR China.
E-mail address: jiming@comp.hkbu.edu.hk (J. Liu).

https://doi.org/10.1016/j.neucom.2019.12.145
0925-2312/

Moreover, the infectious disease diffusion dynamic is quite complex as it can be potentially caused and affected by many risk factors, making it difficult to impute the missing infectious disease data by using only the partially observed case data. For instance, the transmission risk for malaria depends on the per capita mosquito density, which is closely related to environmental factors, such as temperature and rainfall. Empirical studies have also revealed population mobility patterns (and thus the routes by which malaria is transmitted) to be related to the geographical distances between locations and the socioeconomic factors [12,13].

To efficiently implement the surveillance strategy, officials need to ensure real-time reporting of case data [14]. The persistent problem of incomplete case reporting data and the complexity of disease-related risk factors can be formulated as follows: **Given the incomplete case reporting data, how can the heterogeneous data sources for complex disease-related risk factors be used to effectively restore the integrity of case reporting data, i.e., estimate missing values, such that the number of infection cases in unobserved time periods and locations can be accurately inferred?**

Often, in order to restore the integrity of the case reporting data, a spatio-temporal modeling strategy would be adopted to accommodate the disease infections across both time and geographical locations by reflecting complex dual-dimensional correlations. There are some classical machine learning methods that capture the spatial, temporal, or spatio-temporal correlations of data, such as Kriging [15] (spatial method), Gaussian process (GP) [16] (temporal method), and K-nearest-neighbor (KNN) imputation [17] (spatio-temporal method). However, these may not be sufficient for restoring the integrity of case reporting data in infectious diseases, as they only consider the target variable itself (in our scenario, this would be the number of reported cases), while ignoring disease-related risk factors, which have been shown to be closely related to the transmission and spread of infectious diseases and should be taken into consideration.

Several recent machine learning approaches have been proposed that take advantage of external information to help inference [2]. These methods are well-grounded and have performed well across various inference tasks. However, they also require the completeness in historical observations of the target variable (here again, the target variable means the number of reported cases). Again, the infectious disease surveillance data cannot satisfy this prerequisite, especially for the hard-to-reach areas. Therefore, these approaches are not directly applicable to our task.

To restore the integrity of case reporting data using both the incomplete reported cases and the underlying disease-related risk factors for inferring the number of infectious disease cases in unobserved locations, we develop a novel machine learning method dubbed Heterogeneous Neural Metric Learning (HNML). Unlike existing spatio-temporal methods for missing data estimation, which only model static spatio-temporal correlations for the target variable, our method recovers missing data using both the target variable and the underlying disease-related risk factors, thus making the estimation more reliable. Compared with other approaches that incorporate external information to help with inference, our method does not hold a strong assumption about the completeness of historical data, which makes the proposed method more useful in the practical setting under consideration.

We empirically validate the effectiveness of our developed machine learning solution on a representative infectious disease, malaria. We use the 2005–2009 malaria case reporting data collected from the malaria endemic China-Myanmar border region. To systematically evaluate the performance of our method, we test it under three data missing patterns (spatial missing, temporal missing, and spatio-temporal missing) resulted from three common surveillance strategies with different levels of missing rates (from 10% to 50%). We also compare our method with the existing inference methods (including both the classical and the state-of-the-art methods). The results demonstrate that our method makes inferences on the unobserved malaria cases with higher accuracy, indicating its effectiveness in restoring the integrity of case reporting data with missing case data.

Note that in this paper, we use the word "heterogeneous" to emphasize four unique characteristics of the spatio-temporal modeling of infectious diseases with incomplete data:

1. *Various intrinsic properties*. For example, in malaria transmission modeling, the temperature and rainfall datasets describe the environmental property, while the social-economic dataset characterizes the property of human activities.[1]
2. *Various roles in shaping epidemiology dynamics*. The temperature and rainfall play a role in vector reproduction, which triggers the epidemiological transmission within one location, while the social-economic activity determines the inter-location transmission.
3. *Various availabilities*. The temperature dataset is generally easy to obtain from the satellite remote sensing, which covers a large spatial region in a high resolution. However, the social-economic dataset is recorded manually, which is difficult to collect, especially in the remote area.
4. *Various spatio-temporal resolutions*. The spatial and temporal resolutions of temperature dataset are 1 km (km) and daily, respectively. While the social-economic dataset is collected in the town-level spatial resolution and annually temporal resolution.

Such distinct characteristics in the heterogeneous data sources make the imputation task quite challenging.

Our contributions in this paper could be summarized as follows.

1. First, we develop a machine learning method to conduct data imputation in scenarios featuring a variety of missing data patterns by extracting information from underlying related factors. Unlike existing spatial, temporal, or spatio-temporal methods, our method takes advantage of additional information about underlying related factors and thus outperforms existing methods.
2. Second, we propose to take underlying disease-related risk factors into account when imputing missing values and inferring unobserved cases of infection using heterogeneous disease-related data sources. Specifically, we incorporate environmental factors (temperature and rainfall), geographical factors (latitude and longitude), and 22 socioeconomic factors collected from a multitude of data sources into our disease transmission model.
3. Third, we empirically evaluate our method's performance using ground-truth based on the 2005–2009 malaria case reporting data collected from the malaria endemic China-Myanmar border region, under three common real-life data missing patterns (spatial missing, temporal missing, and spatio-temporal missing) with different levels of missing rates (10%–50%). The results show that our method provides more accurate inferences than existing methods. The results also suggest that by appropriately incorporating underlying disease-related risk factors, extra information on malaria transmission patterns can be obtained and inference accuracy can be enhanced. The results thus provide a scientific foundation for public health authorities to implement real-time surveillance for malaria elimination and

[1] Please refer to Section 5.1 for further details about the environmental, geographic and social-economic data sources for the empirical study in Yunnan province.

our approach could also be a useful framework for other applications with similar spatio-temporal missing scenarios and heterogeneous data sources.

2. Related work

By modeling spatial correlations in the data, spatial methods, such as the Kriging [15] and inverse distance weighting [18] methods, map the propagation of information across geographical space, and use the mapping of spatial correlations to recover missing information. Using an entirely different assumption, temporal methods, such as the Gaussian process (GP) [16] and auto regressive moving average (ARMA) [19] methods, ignore the horizontal propagation of information across geographical space, and recover data via a vertical mapping between reported cases across timeframes. Though the spatial and temporal methods each have their own merits for imputing missing information in specific scenarios, neither is suitable for modeling infectious diseases, where spreading can occur both spatially and temporally at the same time. For instance, if a spatial method is applied to recover missing data while ignoring the continuity of a disease spreading across timeframes, it is likely that the decision would be biased toward areas with more complete temporal disease data. By the same token, using temporal methods while overlooking the possibility of geographic spreading due to infected individuals' mobility across locations, the conclusions would also be highly biased.

Spatio-temporal methods take both spatial information and temporal information into consideration when imputing missing values. To capture more complex spatio-temporal correlations, other approaches have been developed recently. Yi et al. [20] proposed a spatio-temporal multiview-based learning method, ST-MVL, to infer missing values by considering the spatial and temporal correlations from both global and local views. Senanayake et al. [8] introduced a variation on the GP regression technique that can characterize both spatial and temporal variations of influenza cases. These methods are well grounded and have been shown to perform well in various applications of missing data imputation.

In the general missing data imputation, besides the most commonly used methods in this category are mean imputation and K-nearest-neighbor (KNN) imputation [21,22], Jinsung proposed a generative adversarial based method, called generative adversarial imputation nets (GAIN). Beside a generator to generate the missing value, GAIN incorporate a discriminator which attempts to classify the actually observed components and imputed components. In this way, the generator is focused to learn to generate according to the true data distribution.

However, each of these methods models spatio-temporal correlations for only the target variable, while ignoring the underlying factors, which have been shown to play an important role in making any inferences about the target variable [11]. Consider Plasmodium vivax (P. vivax), a typical category of malaria, as an example. Infected persons transmit P. vivax to susceptible persons via female anopheles mosquito bites [23]. Different locations may have different per capita mosquito densities because of heterogeneous environmental and demographic factors (such as temperature, rainfall, and population density), and thus may yield different P. vivax incidences (i.e., the numbers of infected cases). Empirical studies have also demonstrated the effects of population movement on the spread of mosquito-borne infectious diseases [24,25]. Population movement to and from particular locations varies because of heterogeneous socioeconomic and geographical factors, which may result in heterogeneous spatial correlations of disease incidences. Therefore, the underlying risk factors should be taken into consideration when inferring missing values of P. vivax incidence at different spatio-temporal coordinates. In our recent work [26], we incorporated underlying risk factors for missing data imputation, and achieved state-of-the-art performance in terms of inference accuracy. It should be mentioned that in [26], we made two assumptions: 1) the effects from underlying risk factors could be linearly combined; and 2) the disease-related risk factors in different locations should be homogeneous. The first assumption simplifies the complex correlation between different risk factors, while the second assumption might be overly optimistic about some of the real-world situations, where the attributes obtained from different locations could be highly heterogeneous.

3. Proposed method

To restore the integrity of case reporting data using both the incomplete reported cases and the underlying disease-related risk factors to infer the missing numbers of infection cases, we develop a machine learning method called Heterogeneous Neural Metric Learning (HNML). Fig. 1 illustrates the idea behind our method. As shown in Fig. 1(a), given incomplete historical data, HNML integrates different disease-related risk factors from heterogeneous data sources, such as environmental, geographical, and socioeconomic factors, to restore the integrity of the case map and infer the number of infection cases in unobserved locations. In the maps shown in Fig. 1(a), t indicates the time point, the gray indicates missing data, and the red indicates available data. The intensity of the red indicates the number of infection cases in a specific location/region, with higher intensity corresponding with larger numbers. Fig. 1(b) shows the detailed structure of HNML. Specifically, HNML utilizes disease-related risk factors to learn an embedding for each location; the learned embeddings are used to characterize the correlations between different locations. HNML then integrates the incomplete historical data and the correlations between different locations to restore the integrity of case reporting data and make inferences on unobserved data via regression.

3.1. Problem statement

Before introducing our method, we formally define the problem of spatio-temporal missing data estimation. Let $\mathbf{Y} \in \Re^{N\times T}$ denote the case reporting data collected in N locations during T time steps, $\mathbf{Y}_t \in \Re^N$ denote the reported cases in time t, and $\mathbf{Y}_t^i$ is the i^{th} element of $\mathbf{Y}_t$. Let $\mathcal{W} \in \Re^{D_W\times N\times T}$ and $\mathcal{X} \in \Re^{D_X\times N\times T}$ denote the attributes of temporally and spatially related risk factors, respectively. Furthermore, $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_t, \ldots, \mathbf{X}_T\}$, where $\mathbf{X}_t \in \Re^{D_X\times N}$ represents the attributes at time t, and $\mathbf{X}_t^i \in \Re^{D_X}$ is the i^{th} row. The same representation applies for $\mathcal{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_t, \ldots, \mathbf{W}_T\}$. There are temporal correlations between $\mathbf{Y}_t^i$ and $\mathbf{Y}_{t-1}^i$ and spatial correlations between $\mathbf{Y}_t^i$ and $\mathbf{Y}_t^{\sim i}$, where $\sim i$ indicates all other locations except location i. The spatio-temporal correlations are hidden and should be estimated from the target variable $\mathbf{Y}$ and attributes $\mathcal{W}$ and $\mathcal{X}$. Note that in our scenarios, some data in $\mathbf{Y}$ could be missing.

The spatio-temporal missing data estimation problem is defined as follows. Let $\mathbf{Y}_t^m$ and $\mathbf{Y}_t^o$ denote the missing and observed values of the target variable, respectively, at time t. The length of the vectors in different time steps may be various, depending on the missing patterns. The goal of missing data estimation is to accurately estimate the values of the target variable in unobserved locations $\mathbf{Y}_t^m$, based on the available observations $\mathbf{Y}_{[1:t]}$ and the attributes $\mathbf{W}_{[1:t]}, \mathbf{X}_{[1:t]}$. The mathematical notations used in this paper are summarized and explained in Table 1.

3.2. Integrating disease-related risk factors via HNML

In this subsection, we introduce our method for incorporating various data sources on location-specific disease-related risk fac-

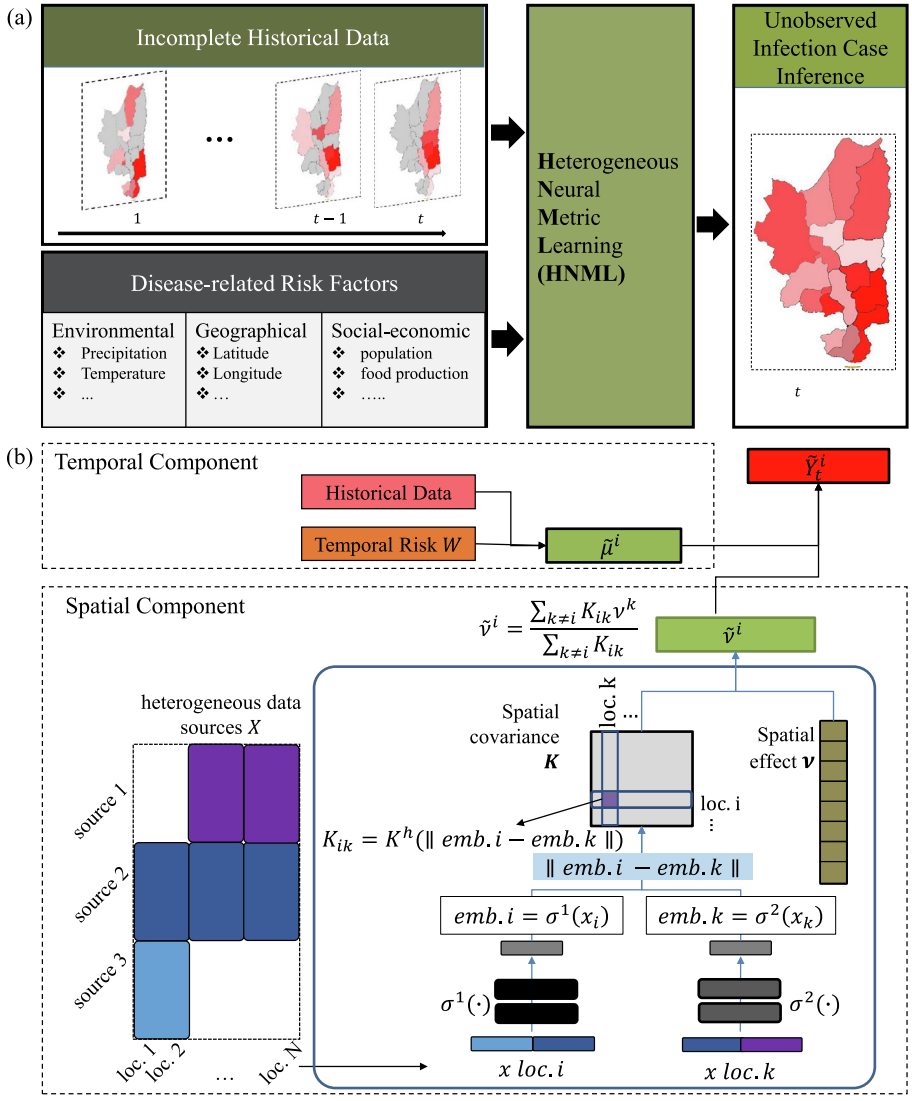


Fig. 1. Illustration of Heterogeneous Neural Metric Learning (HNML) for inferring the current number of infection cases in unobserved locations from incomplete historical data and heterogeneous data sources. (a) Given the incomplete historical data, HNML integrates different disease-related risk factors from heterogeneous data sources, such as environmental, geographical, and socioeconomic factors, to infer unobserved infection cases in different locations. (b) The detailed structure of the HNML. HNML utilizes disease-related risk factors to learn an embedding for each location; the learned embeddings are used to characterize the correlations between different locations. As an example for illustration here, HNML learns two mappings from two types of feature space, i.e., source 1 + 2 and source 2 + 3, to the common space. HNML then integrates the incomplete historical data and the correlations between different locations to restore the integrity of case reporting data and make inferences for unobserved data via regression.

tors. Mathematically, the target variable is represented as the linear combination of a temporal effect, a spatial effect, and a noise term:

$$\mathbf{Y}_t = \boldsymbol{\mu}_t + \boldsymbol{\nu}_t + \boldsymbol{\epsilon}_t, \ \ \boldsymbol{\epsilon}_t \sim N(0, \sigma^{*2}\mathbf{I}), \tag{1}$$

where $\boldsymbol{\mu}_t \in \Re^N$ is a vector representing a temporal dynamic process and $\boldsymbol{\nu}_t \in \Re^N$ is a vector representing the spatial component on different locations.

3.2.1. Epidemiological dynamics of disease transmission

Various disease transmission models have been studies for different diseases. To assess the malaria transmission potential among the locations, we consider the inter-location transmission and disease-related risk factors, such as environmental and demographic factors, in the temporal dynamic. Following the malaria transmission model developed in [27], the temporal correlations are related to the location-specific attributes as follows:

$$\boldsymbol{\mu}_t = f(\mathcal{B}, \mathbf{Y}) = \mathbf{Y}_{t-1}\mathbf{Z} + \mathbf{B}\mathbf{W}_{t-1}, \tag{2}$$

where $\mathcal{B} = \{\mathbf{B}, \mathbf{Z}\}$, $\mathbf{B}$ is the coefficients of influence from risk factors, and $\mathbf{Z} \in \Re^{N\times N}$ is the inter-location transmission matrix.

3.2.2. Spatial correlations

The spatial correlation between different locations is modeled in the form of Nadaraya-Watson estimator [28]:

$$\boldsymbol{\nu}_t^i = \frac{\sum_{j=1,j\neq i}^{N} K_{ji} * \boldsymbol{\nu}_t^j}{\sum_{k=1,k\neq i}^{N} K_{ki}}, \tag{3}$$

where $\boldsymbol{\nu}_t^i$ is the i_{th} element of $\boldsymbol{\nu}_t$, representing the spatial component of location i at time step t, $K_{ji} = K_h(d(\mathbf{x}_t^j, \mathbf{x}_t^i))$ is the $(j,i)_{th}$ element of spatial covariance matrix $\mathbf{K}$, and $K_h(d(\mathbf{x}_t^j, \mathbf{x}_t^i))$ is a kernel function for measuring the closeness/similarity between two locations in terms of their attributes. For simplicity, we omit the time index of the attributes (for example, $\mathbf{x}^j := \mathbf{x}_t^j$). The spatial effect term in a specific

Table 1
Mathematical notations used in this paper and their corresponding explanations.

Symbols	Meanings
$\mathbf{Y}$	Incidences collected in N locations during T time steps
N	Number of locations
T	Number of time steps
$\mathbf{Y}_t$	Incidences collected in N locations at time point t
$\mathbf{Y}_t^i$	the i-th element of $\mathbf{Y}_t$
$\mathcal{W}$, $\mathcal{X}$	Covariate data for temporal correlations and spatial correlations
D_W, D_X	Number of attributes in covariate data for temporal correlations and spatial correlations
$\mathbf{W}_t$, $\mathbf{X}_t$	Attribute of covariate data for temporal correlations and spatial correlations at time step t
$\mathbf{x}_t^i$	Covariate data of location i for spatial correlations at time step t
$\mathbf{Y}_{[1:t]}$	Incidences collected in N locations from time step 1 to time step t
$\mathbf{Y}_t^m$, $\mathbf{Y}_t^o$	The missing and observed values of the target variable at time t.
$\mathbf{W}_{[1:t]}$, $\mathbf{X}_{[1:t]}$	Attribute of covariate data for temporal correlations and spatial correlations from time step 1 to time step t
$\boldsymbol{\mu}$ $\boldsymbol{\nu}$	Temporal and spatial components in the incidences data
$\boldsymbol{\mu}_t$, $\boldsymbol{\nu}_t$	Temporal and spatial components in the incidences data at time step t
$\boldsymbol{\nu}_t^i$	spatial components in the incidences of location i at time step t
$\boldsymbol{\epsilon}_t$	Noise component in the incidences data
$\mathbf{Z}$	$N \times N$ inter-location transition matrix to be estimated.
$\mathbf{B}$	$N \times D_W$ matrix representing the coefficients of influence from risk factors to the number of infection
K_h	Kernel function for spatial correlations
$\mathbf{K}$	Spatial covariance matrix
K_{ji}	The $(j,i)_{th}$ element of spatial covariance matrix
$d(\mathbf{x}^j, \mathbf{x}^i)$	Parametric distance function that measures the distance between two feature vectors.
$I(j)$	Indicator function which indicates which type of feature space that location j belongs to.
G	Linear mapping for metric learning
$\mathbf{A}$	Positive semi-definite rescale matrix.
$\sigma^{I(j)}(\cdot)$	Nonlinear mapping for location j
$\boldsymbol{\Theta}$	The parameters in the neural networks
$emb.j$	Embedding representation of location j in metric learning
$eemb.j$	Extra embedding representation of location j in metric learning

location is the weighted average of those in corresponding locations. For instance, if two locations have similar attributes, their spatial variations would exhibit the same trend. By assuming that similar locational attributes indicate a high probability of experiencing similar spatial effects, we can use the Gaussian kernel with adjusted distance input:

$$K_h(d(\mathbf{x}^j, \mathbf{x}^i)) = \frac{1}{\sigma'\sqrt{2\pi}} \exp\left(-\frac{d(\mathbf{x}^j, \mathbf{x}^i)}{\sigma'^2}\right). \tag{4}$$

In previous studies, only geographical distance was utilized to determine the closeness across locations. With external information for the location profile, we use the metric learning technique to learn an appropriate distance function with higher-dimension attributes. In linear metric learning, a linear mapping, encoded as a matrix G, is learned such that the learned distance is $\|Gx_j - Gx_i\|^2$. In fact, a general Mahalanobis distance would be learned: $d_A(x_j, x_i) = (x_j - x_i)^T A(x_j - x_i)$, where A is a positive semi-definite matrix.

In this work, we consider a more general nonlinear distance function:

$$d(\mathbf{x}^j, \mathbf{x}^i) = \|\sigma^{I(j)}(\mathbf{x}^j) - \sigma^{I(i)}(\mathbf{x}^i))\|^2, \tag{5}$$

where $I(j)$ and $I(i)$ are indicators function indicating which type of feature space that location belongs to (e.g., $I(j) = i$ indicates the j_{th} location belong to the i_{th} feature), and $\sigma^{I(j)}(\cdot)$ and $\sigma^{I(i)}(\cdot)$ are nonlinear mappings for location j and i, respectively. We learn a function that maps input patterns into a target space such that the L_2 norm in the target space approximates the "semantic" distance in the input space. Recent studies have demonstrated the effectiveness of deep embedding [2,29]. We thus consider $\sigma^{I(j)}(\cdot)$ as a nonlinear mapping from the feature space of location j to the common space. In most of the existing methods, the feature for each sample is complete. However, because of the difficulty of data collection or system error in real-world applications, some data sources are not available in certain locations. Thus we need to learn a unique mapping for each feature space onto the common space. In our learning framework, the nonlinear distance is formulated as:

$$\begin{aligned} d(\mathbf{x}^j, \mathbf{x}^i) &= \|emb.j - emb.i\|^2, \\ s.t.\ \ emb.j &= \sigma^{I(j)}(\mathbf{x}^j), emb.i = \sigma^{I(i)}(\mathbf{x}^i), \end{aligned} \tag{6}$$

where $emb.j$ and $emb.i$ are the embedding representation of location j and i respectively, and $\sigma^{I(j)}(\cdot)$ and $\sigma^{I(i)}(\cdot)$ are modeled using neural networks. The framework of our method is shown in Fig. 1(b).

Remarks. Although the temporal effect and the spatial effect are integrated without a hyper-parameters adjusting the relative weights of these two effects, the relative weights of spatial and temporal effects could be considered as being automatically learned in our algorithm. Assume we introduce a hyper-parameter α for the relative weights of spatial and temporal effects into the spatio-temporal model:

$$\mathbf{Y}_t = \alpha\boldsymbol{\mu}_t + (1-\alpha)\boldsymbol{\nu}_t + \boldsymbol{\epsilon}_t,\ \boldsymbol{\epsilon}_t \sim N(0, \sigma^{*2}\mathbf{I}), \tag{7}$$

where $1 \geqslant \alpha \geqslant 0$. Substituting Eqs. (2) and (3) into Eq. (7), we have: $\mathbf{Y}_t = \mathbf{Y}_{t-1}\mathbf{Z}\alpha + \alpha\mathbf{B}\mathbf{W}_{t-1} + (1-\alpha)\boldsymbol{\nu}_t + \boldsymbol{\epsilon}_t$. We could redefine the parameters as $\mathbf{Z}_\alpha := \mathbf{Z}\alpha, \mathbf{B}_\alpha := \alpha\mathbf{B}$, and $\boldsymbol{\nu}_{\alpha t} := (1-\alpha)\boldsymbol{\nu}_t$, which are equivalent to the previous parameters $\mathbf{Z}, \mathbf{B}$, and $\boldsymbol{\nu}_t$, respectively. Thus the relative weights of spatial and temporal effects could be automatically optimized in our algorithm to fit the training data.

3.2.3. Spatial correlations with unknown factors

In real-world infectious disease spread, there may be unknown factors that affect the dynamics of the target variable but are not covered by the heterogeneous data sources. Neglecting these factors may limit the power of the data-driven disease model. Taking the unknown risk factors into consideration, we learn an extra representation for each location:

$$\begin{aligned} d(\mathbf{x}^j, \mathbf{x}^i) &= \|emb.j - emb.i\|^2, \\ s.t.\ \ emb.j &= [\sigma^{I(j)}(\mathbf{x}^j); eemb.j], emb.i = [\sigma^{I(i)}(\mathbf{x}^i); eemb.i]. \end{aligned} \tag{8}$$

The $eemb.i$ is an extra representation vector to be estimated for location i for spatial covariance. The proposed method for considering the unknown factors (referred to as HNML-UF) is illustrated in Fig. 2.

3.3. Inferences on unobserved data

Lasso regression is used to learn the temporal autoregressive model, which incorporates variable selection and sparse structure using L_1 regularization. The loss function of the kernel model is given as: $L_s = \sum_t \|\boldsymbol{\nu}_t^i - \hat{\boldsymbol{\nu}}_t^i\|_2$, where $\hat{\boldsymbol{\nu}}_t^i$ is the estimated value obtained via regression model. The derivative with respect to $emb.i$ is given as follows:

$$\frac{\partial L_s}{\partial emb.i} = \sum_t (\hat{\boldsymbol{\nu}}_t^i - \boldsymbol{\nu}_t^i) \sum_j (\hat{\boldsymbol{\nu}}_t^j - \boldsymbol{\nu}_t^j) K_{ij} (emb.i - emb.j). \tag{9}$$

The neural mapping is learned via error back-propagation. To integrate the spatial and temporal learning processes, an alternating algorithm is developed. The loss function is formulated as follows:

$$L(\mathcal{B}, \boldsymbol{\Theta}) = \sum_i \sum_t \|\mathbf{Y}_{i,t} - (f(\mathcal{B}, \mathbf{Y}))_i - \frac{\sum_{j=1, j\neq i}^N K_{ji} * \boldsymbol{\nu}_t^j}{\sum_{k=1, k\neq i}^N K_{ki}}\|_2^2, \tag{10}$$

where the subscript i represents the i-th element of a vector and Θ is the parameter in the neural network model that used to calculate spatial covariance $\mathbf{K}$. The alternating algorithm aims to decrease the loss by optimizing matrices $\mathcal{B}$ and $\mathbf{\Theta}$ alternately. The details are shown in Algorithm 1. As mentioned above, the EM algorithm is used to reduce the bias caused by missing values. The missing values are first imputed using statistical methods, e.g., the Mean method. Then, the following two steps are performed alternately: i) learning the spatio-temporal model on the complete dataset, and ii) imputing the missing values based on the learned spatio-temporal model. Finally, missing values in the target variable, i.e., $\mathbf{Y}_t^m$, are estimated by combining the spatial and temporal effects: $\mathbf{Y}_t^m = \boldsymbol{\mu}_t^m + \boldsymbol{v}_t^m$.

Algorithm 1. Spatio-temporal imputation via Heterogeneous Neural Metric Learning (HNML)

Input: Incomplete spatio-temporal incidences data $\mathbf{Y}$; Attributes $\mathcal{X}, \mathcal{W}$.
Output: Learned parameter $\{\mathcal{B}, \mathbf{\Theta}\}$
1: Initial $\boldsymbol{v} \in \Re^{N \times T}$,;
2: **while** not converged **do**
3: **while** not converged **do**
4: /* Update temporal component
5: Estimate $\mathcal{B}$ from $\mathbf{Y} - \boldsymbol{v}$;
6: Update $\boldsymbol{\mu}$ from new $\mathcal{B}$;
7: /* Update spatial component
8: Estimate $\mathbf{\Theta}$ from $\boldsymbol{v} = Y - \boldsymbol{\mu}$;
9: Update $\boldsymbol{v}$: $\boldsymbol{v}_t^i = \frac{\sum_{j=1, j \neq i}^{N} K_{ji} * \boldsymbol{v}_t^j}{\sum_{k=1, k \neq i}^{N} K_{ki}}$
10: **end while**
11: /* Impute the missing values/*
12: $\mathbf{Y}_t^m = \boldsymbol{\mu}_t^m + \boldsymbol{v}_t^m$
13: **end while**

4. Synthetic experiments

We evaluate the performance of our method on a systematically designed synthetic dataset.

4.1. Synthetic dataset generation

The synthetic dataset is systematically generated as follows:

1. Temporal Component: First, a first-order autoregressive model was used to simulate the temporal correlations $\mu_t = Y_{t-1}\beta,\ \beta \in R^{N \times N}$. We included a seasonal evolution matrix, $\beta = \beta_o + H * s_t$, where s_t is a seasonal indicator that can be treated as the temporal correlation attributes.
2. Spatial Component: Without loss of generality, we assumed there are two types of risk factors, each of which consists of D attributes. To simulate dynamic spatial closeness effect, for each data source, we divided the locations into g groups, $\mathcal{C} = \{C_1, \cdots, C_g\}$. For locations in each group, the corresponding attributes were sampled from the group-specified distribution: $(X^i)_d \sim N(m_g, \sigma_g), i \in C_g$. In this way, we could quantify the ground-truth similar locations of each location. We sampled each element of the first row of the input matrix of neural model $W^k \in \Re^{D \times hidden\_size}, k \in \{1, 2\}$ using the uniform distribution $U(0, 1)$ and sampled each element of the other rows using a normal distribution $N(0, 0.01)$, i.e., only the first attribute of each factor plays an role in shaping spatial patterns. The location-specific attributes were fixed in each period but varying between different periods, which implies that the strength of correlations in different periods were slightly different. In our simulation, when entering a new period, we resampled each attribute value $(X_p^i)_d \sim N((X_{p-1}^i)_d, \sigma_v)$.
3. Unknown factors: We randomly selected 50% of the locations, and the second risk factor of the selected locations would be missing in the training and test phases.
4. We ran the dynamic model to generate the time series data.

4.2. Design of evaluations

We conduct a series of experimental evaluations to systematically assess the performance of our method. Specifically, we:

1. **Explore three common real-life data missing patterns**. We vary the synthetic infectious disease data to generate training data with different missing patterns. Here, each variation is created by deliberately withholding a portion of the data from the original complete dataset. This creates datasets that resemble the missing data patterns found in real world contexts and that can be used to model a restoration of data integrity.
   Infection cases from a few randomly selected locations are removed to simulate a real-life spatial missing (S-M) pattern, while random infection cases from various time points are removed to mimic a real-life temporal missing (T-M) pattern. Infection cases are also withheld randomly along both spatial and temporal dimensions to imitate a real-life spatio-temporal missing (ST-M) pattern. The formal setting of three real-life missing patterns are given as follows:
   - Spatial Missing (S-M): We randomly select $MissingRate * N$ locations, where $MissingRate$ denotes the missing rate, and N is the total number of locations. The target variable in the selected locations will be considered as missing.
   - Temporal Missing (T-M): In each year (i.e., 12 months), a time window containing $(MissingRate) * 12$ successive time-frames is randomly selected for each location. The target variable in the selected time window will be considered as missing.
   - Spatio-Temporal Missing (ST-M): The target variable in the location i at time t will be considered as missing if $r_t^i < MissingRate$, where $r_t^i \sim U(0, 1)$.
2. **Test our method at five different levels of missing rates**. To evaluate the robustness of our method, we set five levels of missing rates for the observed data, from a small portion to a large part: $MissingRate = 10\%, 20\%, 30\%, 40\%, 50\%$. Higher missing rate indicates less observed data, which brings more challenge to the inference task as the information obtained from the highly incomplete data could seriously distorts the true distribution of the complete data.
3. **Compare our method with four representative inference methods**. We select Kriging [15], Gaussian process (GP) [16], Mean imputation [30], K-nearest-neighbor (KNN) imputation [17] for comparison. Kriging is a classical spatial method that maps the propagation of information across geographical space and uses the mapping of spatial correlations to recover missing information. Temporal Gaussian process is a classical temporal method that recovers data via a vertical mapping between reported cases across time frames. K-nearest-neighbor imputation and Mean imputation are the classical imputation methods for numerical spatio-temporal data. In the spatio-temporal content, KNN selects the close locations in one time frame by measuring the similarity of data in the previous time frames.

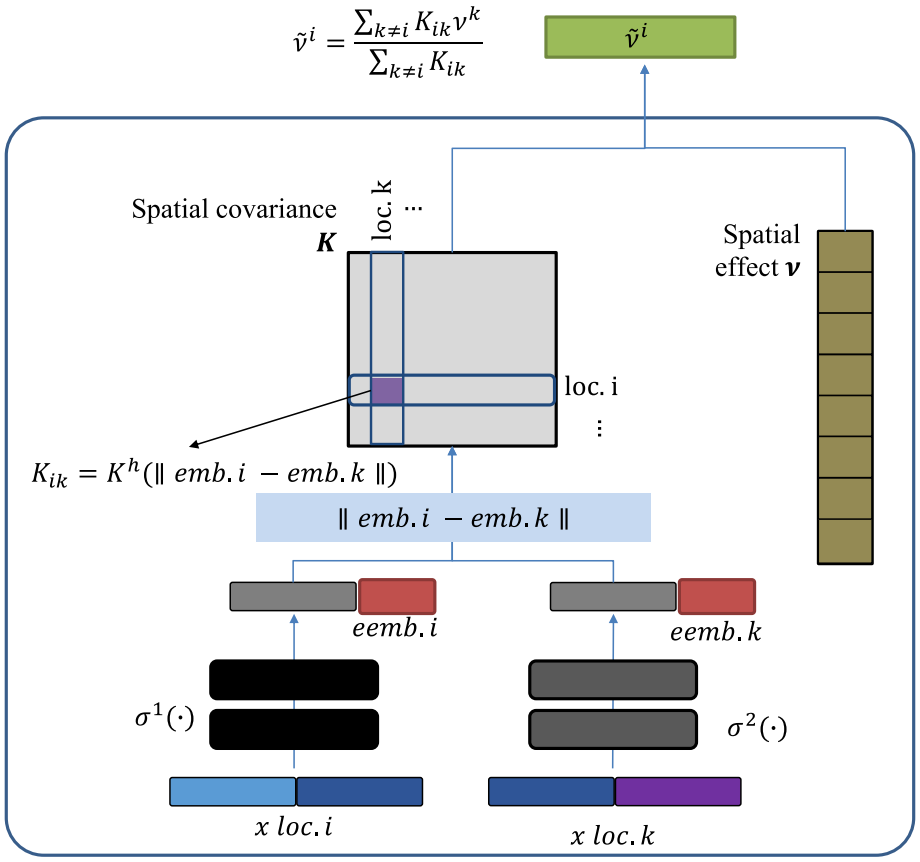


Fig. 2. Illustration of Heterogeneous Neural Metric Learning with Unknown Factors (HNML-UF) for considering unknown factors in spatial correlation modeling. To further consider the unknown factors that shaping the spatial correlations, an extra representation vector *eemb*, as indicated as the red rectangle, is to be estimated for each location.

The performance of each method is evaluated in terms of the mean average error (MAE): $MAE = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}|y_t^i - \hat{y}_t^i|$, where y_t^i is the ground-truth and $\hat{y}_t^i$ is the estimated number of infected cases in location i at time t.

4.3. Results

The performance of all the methods is evaluated in terms of the mean average error (MAE). Table 2 showed the results of all the methods under different missing patterns and missing rates. Our method outperformed the existing methods in all the scenarios. Moreover, HNML-UF consistently performed better than HNML, indicating the capacity of HNML-UF in modeling unknown factors.

5. An empirical study in Yunnan province

We empirically validate the effectiveness of our developed machine learning solution using the 2005–2009 malaria case reporting data collected from Yunnan, a malaria-endemic province located in the China-Myanmar border region (as shown in Fig. 3), because malaria is one of most serious infectious disease and the Yunnan province is a typical region in the phase toward malaria elimination.

5.1. Scenario description

The population's cross-border activities may result in imported cases of malaria in the border area, which can trigger local infections [9,31,11]. It is therefore crucial to effectively implement the active surveillance strategy in the area to assess malaria risks for different locations. Given that China has entered the malaria elimination phase, it is essential that imported infections are taken into account, as imported cases currently dominate the total number of infection cases.

Fig. 3(a) and (b) illustrate the area in Yunnan province on the China-Myanmar border. It can be observed that along the over 2,000 km long borderline, many villages are situated in mountainous areas with ecological environment that is perfect for the reproduction of malaria vectors. Not to mention the high proportion of mobile population between Yunnan province and Myanmar, doing businesses and acting as carriers who could easily bring infectious diseases from one side of the border to the other. Fig. 3(c) shows the map of Tengchong, a county located in Yunnan that has long been considered as one of the most at-risk regions prone to malaria epidemics and one of the top priorities in the malaria elimination initiative [32,33].

Tengchong comprises 18 townships with around 660,000 residents distributed in a mountainous area of approximately 5–6 thousand square kilometers. The small surveillance team from the Tengchong Center for Disease Control (CDC), however, is only able to allocate a few staff members to conduct case surveys. The lack of resources dedicated to comprehensive surveillance inevitably leads to inefficacies and inaccuracies in case reporting. In practice, the patterns of missing case reporting data can be multifaceted [26]. For instance, in sentinel surveillance, qualified agents, with suitable laboratory facilities and experienced staffs, identify and report all incidences of disease within target locations. Incidence data can be continuously collected in carefully selected locations while the infection case data could be missing for other

Table 2
Performance evaluation (in terms of MAE) of our method and existing inference methods with different missing patterns and varying missing data rates on the synthetic dataset. The best result for each scenario is underlined and highlighted in bold.

Settings		Methods					
Missing Scenario	Missing Rate	MEAN	KNN	Kriging	GP	HNML	HNML-UF
Spatial Missing (**S-M**)	10%	421.424	164.830	353.068	464.94	128.478	**102.121**
	20%	379.603	217.324	303.793	363.703	109.766	**74.637**
	30%	382.410	414.469	323.416	378.233	122.061	**71.194**
	40%	430.247	777.030	387.188	338.833	159.880	**79.162**
	50%	437.567	1072.890	460.260	474.460	215.957	**141.800**
Temporal Missing (**T-M**)	10%	325.129	125.592	111.543	225.647	71.051	**42.537**
	20%	440.142	189.881	269.803	308.184	111.664	**68.174**
	30%	400.975	713.908	338.288	336.046	158.075	**92.478**
	40%	325.621	1231.705	293.901	318.671	166.553	**96.395**
	50%	1649.822	2442.283	351.284	394.571	253.294	**166.743**
Spatio-Temporal Missing (**ST-M**)	10%	427.586	193.991	92.341	273.68	111.048	**90.887**
	20%	410.324	174.425	170.611	277.04	120.203	**97.039**
	30%	295.855	216.095	206.119	251.484	111.509	**75.346**
	40%	407.164	509.054	248.883	295.662	155.104	**106.152**
	50%	312.014	772.923	243.901	237.75	170.459	**99.316**

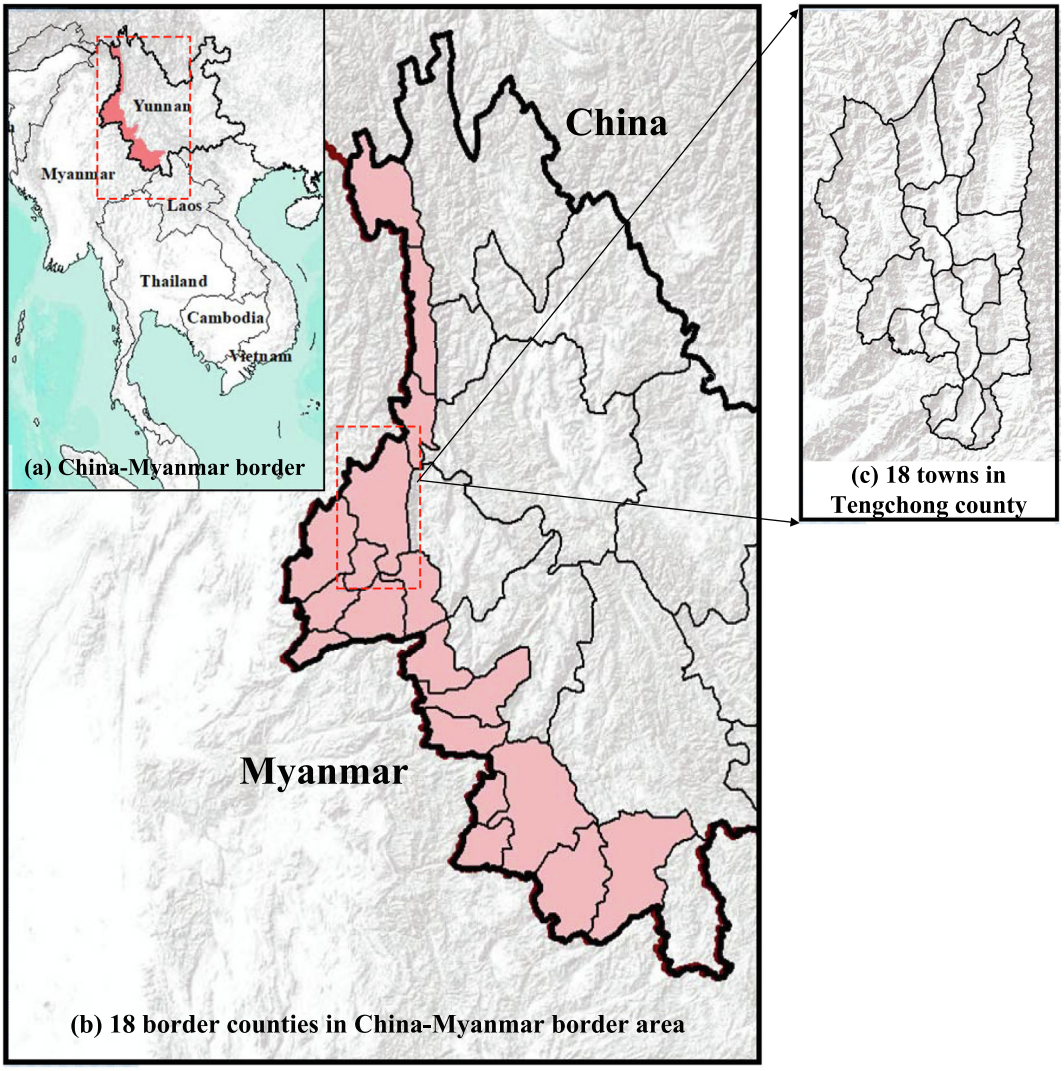


Fig. 3. Mapping of 18 counties in Yunnan province on the China-Myanmar border. (a) and (b) highlight the area in Yunnan province on the China-Myanmar border, covering 18 counties. (c) enlarges a specific county (Tengchong County, including 18 towns) from (b).

locations/regions (e.g., townships in a county) for entire periods of time (referred to as *spatial missing*). On the other hand, surveillance activities may be carried out during certain potential high-risk periods, while the infection case data could be missing for other time points/intervals (e.g., weeks or months) in all of the locations/regions (referred to as *temporal missing*). In the hybrid random surveillance strategy, the infection case data could also be missing along both spatial and temporal dimensions (referred to as *spatio-temporal missing*). Regardless of the reasons and patterns, missing data are always undesirable, as even a small proportion of missing data can affect the validity of research results [34–37].

In addition to the aforementioned challenges, various disease-related risk factors affect disease transmission. For the underlying disease-related risk factors, we collected environmental attributes (temperature and rainfall) and geographical attributes (latitude and longitude) for each of the 62 towns. To incorporate socioeconomic activity, we collected 22 attributes (summarized in Table 3). However, these socioeconomic attributes are only available for the 18 towns in Tengchong.

Table 3
Summary of the underlying risk factors considered in this study, including 2 environmental attributes, 2 geographic attributes, and 22 socioeconomic attributes.

Environmental	Temperature, Rainfall
Geographic	Latitude, Longitude
Socioeconomic	Total households,Total population,Agricultural population, Natural population growth rate,Rural households,Rural population,Number of rural employees, Number of village groups,Year-end cultivated area,Total food production,The production of crops sown in spring,The production of crops sown in tenth lunar month,Per capita production of food, Total production of meat,Year-end pig-sacrifice livestock on hand,Year-end fat pig slaughter,Year-end large domestic animals,Agricultural machinery power, Financial general budget income,Financial general budget expenditure,Per capita income of farmers,Current value of agricultural production

5.2. Design of evaluations

To implement the active surveillance, the local CDC should visit towns house by house to enquire whether there is/was an infection case, which is extremely time and manpower consuming. However, as the human resources are very limited particularly in the remote and poor regions, only partial infection cases could be timely reported. The spatial and temporal coverage of infection surveillance depends on the surveillance strategies, such as sentinel surveillance, periodic surveillance or random surveillance, adopted by the local CDC.

To understand the capacity of our method in inferring the number of malaria cases in unobserved locations, a study has been conducted to evaluate disparities between our method's estimates of infected cases and the observed cases based on historical data. For the purpose of evaluation, we establish a ground-truth that features a specified number of infected cases within defined spatial and/or temporal boundaries as a reference. Specifically, we collected the monthly case reporting data from 62 towns in Yunnan from 2005 to 2009. Of the 62 towns, 18 are from Tengchong County while the remaining 44 border or are in close proximity to Tengchong. The case reporting data from 2009 are used as the ground-truth for evaluating the performance of our method's inferences.

Consistent with synthetic experiments, we systematically evaluated the performance of our method with.

1. Three common real-life data missing patterns: spatial missing (S-M), temporal missing (T-M), and spatio-temporal missing (ST-M). Fig. 4(a) illustrates an example of the S-M pattern in 18 towns in Tengchong County, Fig. 4(b) illustrates an example of the T-M pattern, and Fig. 4(c) illustrates an example of the ST-M pattern.
2. Five different levels of missing rates: *MissingRate* = $10\%, 20\%, 30\%, 40\%, 50\%$.
3. Four classical inference methods: Kriging [15], Gaussian process (GP) [16], Mean imputation [30], and K-nearest-neighbor (KNN) imputation [17].

In this real-world spatio-temporal disease dataset, we further compare our methods with four state-of-the-art spatio-temporal imputation and prediction methods as follows.

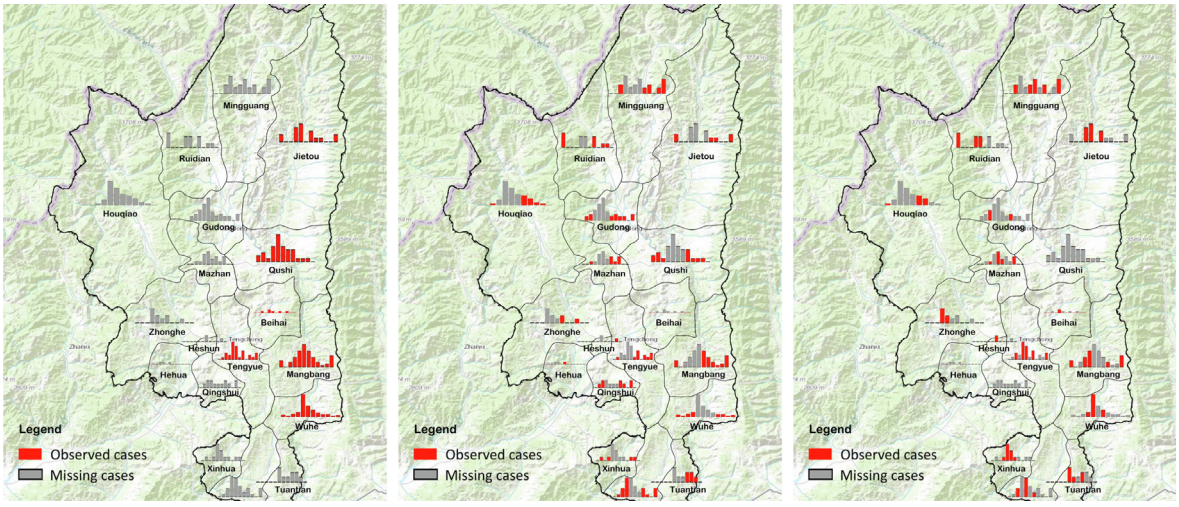
1. ST-ResNet [38]. ST-ResNet, originally developed to forecast the crowd flow within city, is able to collectively incorporate many factors for spatio-temporal prediction. ST-ResNet require the complete history data for input. In the experiments, we complete missing values in the history data using the estimated values by ST-ResNet.
2. Generative Adversarial Imputation Nets (GAIN) [39]. GAIN is the state-of-art method for missing value imputation. Beside a generator to generate the missing value, GAIN incorporate a discriminator which attempts to classify the actually observed components and imputed components. In this way, the generator is focused to learn to generate according to the true data distribution.
3. Spatio-temporal Imputation via Kernel-based Learning (STI-KL) [26]. STI-KL incorporates underlying risk factors for missing data imputation, while it assumes that the effects from underlying risk factors could be linearly combined, and the disease-related risk factors in different locations should be homogeneous.
4. Spatio-temporal multiview-based learning method (ST-MVL) [20]. ST-MVL infers missing values by considering the spatial and temporal correlations among the target value from both global and local views.

We set the hidden size as 32 and number of ST-block as 2 for ST-RestNet; the hidden size as 32 and number of layers as 2 for GAIN; the hidden size as 10 and number of layers as 1 for HNML, as the neural model in HNML only account for spatial correlation.

5.3. Results

5.3.1. Evaluation on 18 towns in Tengchong with homogeneous attributes

We first evaluate the performance of our method by comparing it with other baseline and state-of-the-art inference methods on 18 towns in Tengchong County. Both HNML and HNML-UF are our methods. HNML-UF models the unknown factors while HNML does not. The attributes of the underlying risk factors collected from these 18 towns are homogeneous, i.e., the environmental, geographic, and socioeconomic factors are complete The results are shown in Table 4. Our method was found to outperform the others in most scenarios, indicating that considering underlying risk factors is useful in inferring infectious disease dynamics with missing data. Moreover, we observe that HNML and HNML-UF perform similarly in this evaluation. The information available for underlying risk factors in these 18 towns are complete, and thus the unknown factor module in our method does not necessarily further



(a) Spatial Missing (S-M) (b) Temporal Missing (T-M) (c) Spatio-Temporal Missing (ST-M)

Fig. 4. Illustration of examples of three missing patterns, i.e., Spatial Missing, Temporal Missing and Spatio-Temporal Missing, in the real-world malaria surveillance at Tengchong, a malaria endemic county in the Yunnan province of China, which borders Myanmar [26]. The data marked in red are observed while those marked in gray are missing. (a) Spatial Missing (S-M), (b) Temporal Missing (T-M), and (c) Spatio-Temporal. Missing (ST-M).

Table 4
Performance evaluation (in terms of MAE) of our method and existing inference methods with different missing patterns and varying missing data rates on 18 towns in Tengchong with homogeneous features from underlying risk factors. The best result for each scenario is underlined and highlighted in bold.

Settings		Methods									
Missing Mode	Missing Rate	MEAN	KNN	Kriging	GP	ST-MVL	GAIN	ST-ResNet	STI -KL	HNML	HNML -UF
Spatial Missing (**S-M**)	10%	2.36	2.26	3.06	2.74	2.14	3.93	1.89	1.76	**1.51**	1.52
	20%	2.43	2.29	2.55	2.39	2.29	2.16	2.24	1.59	2.01	**1.51**
	30%	2.81	2.72	3.12	3.36	1.87	2.48	2.48	1.66	1.94	**1.40**
	40%	2.7	2.61	2.71	2.49	**1.92**	2.79	2.06	2.18	2.04	2.75
	50%	2.93	3.89	2.94	2.72	1.79	3.05	2.82	1.72	2.56	**1.62**
Temporal Missing (**T-M**)	10%	2.20	1.71	2.07	1.85	1.92	2.72	1.82	1.60	**1.50**	1.70
	20%	2.10	1.89	2.00	1.99	1.92	2.54	1.85	1.74	**1.54**	1.92
	30%	2.44	2.91	2.38	2.31	2.39	2.93	2.19	2.29	**2.08**	2.21
	40%	2.66	4.68	2.62	2.38	**1.97**	3.20	2.47	2.61	2.33	2.51
	50%	5.21	10.44	2.51	2.42	**1.89**	3.13	2.43	2.99	2.58	2.83
Spatio-Temporal Missing (**ST-M**)	10%	2.53	1.94	2.33	2.56	1.85	3.38	1.79	2.02	1.93	**1.75**
	20%	2.94	2.46	2.84	2.82	2.00	3.45	1.89	2.46	2.03	**1.82**
	30%	2.94	2.93	2.83	2.92	2.15	3.54	**1.82**	2.23	2.00	1.84
	40%	2.62	3.04	2.57	2.52	2.03	2.65	2.00	2.11	2.05	**1.89**
	50%	2.82	5.22	2.78	2.76	2.07	2.00	**1.90**	2.60	2.12	1.99

improve the inference accuracy. In our algorithm, we first estimate the parameters for the temporal effect. Note that when the data are temporally missing, it is more challenging to estimate the temporal component in one location, making the inference of spatial effect unstable as the spatial effect and temporal effect are linearly integrated. In this case, with less parameters, HNML produces more robust and accurate estimation. When the temporal dynamic could be estimated relatively accurately, HNML-UF could better estimate the spatial effect.

5.3.2. Evaluation on 62 towns in and outside Tengchong with heterogeneous attributes

We then evaluate the performance of our methods and other imputation methods on the 62 towns (all in Yunnan province). For the 18 towns in Tengchong, the information on environmental, geographic, and socioeconomic factors is complete, whereas the 44 towns outside Tengchong have complete information for environmental and geographic factors but information for the 22 socioeconomic factors is entirely missing. Table 5 reports the results of the different methods. Note that the method STI-KL [26] is not applicable to the case of heterogeneous attributes, as it requires the homogeneity of disease-related risk factors in different locations. Similar to the previous evaluation, our method outperforms the others in terms of the inference accuracy in most scenarios. However, unlike the previous evaluation, in which HNML and HNML-UF perform similarly, in this evaluation HNML-UF consistently outperforms HNML in each scenario. In this evaluation, information for socioeconomic factors is missing for most of the towns; this makes

Table 5
Performance evaluation (in terms of MAE) of our method and existing inference methods with different missing patterns and varying missing data rates on 62 towns (18 in Tengchong and 44 outside Tengchong) in Yunnan province, with heterogeneous features from underlying risk factors. The best result for each scenario is underlined and highlighted in bold.

Settings		Methods								
Missing Mode	Missing Rate	MEAN	KNN	Kriging	GP	ST-MVL	GAIN	ST-ResNet	HNML	HNML -UF
Spatial Missing (**S-M**)	10%	0.724	0.647	0.740	1.599	0.976	1.90	0.709	0.623	**0.556**
	20%	0.716	1.050	0.701	1.476	0.844	1.28	0.909	0.734	**0.594**
	30%	0.713	1.452	0.713	1.437	0.819	1.23	0.740	0.727	**0.604**
	40%	0.737	1.654	0.747	1.316	0.747	1.10	0.726	0.652	**0.610**
	50%	0.728	1.496	0.764	1.186	0.799	1.09	0.758	0.734	**0.712**
Temporal Missing (**T-M**)	10%	0.702	0.941	0.683	1.608	0.818	0.96	0.722	0.645	**0.609**
	20%	0.803	1.695	0.786	1.287	0.858	1.31	0.740	0.744	**0.720**
	30%	0.770	1.821	0.780	1.036	0.792	1.09	**0.682**	0.716	0.707
	40%	0.791	2.837	0.796	1.032	0.869	1.23	0.736	0.752	**0.720**
	50%	2.018	3.691	0.790	0.840	0.874	1.26	**0.672**	0.707	0.705
Spatio-Temporal Missing (**ST-M**)	10%	0.741	0.579	0.745	1.608	0.846	1.32	1.112	0.719	**0.563**
	20%	0.736	0.705	0.714	1.595	0.853	1.63	0.755	0.652	**0.546**
	30%	0.749	0.798	0.742	1.260	0.850	1.48	0.677	0.699	**0.572**
	40%	0.742	1.004	0.747	1.118	0.817	1.30	0.660	0.670	**0.631**
	50%	0.740	1.512	0.739	0.967	0.860	1.11	0.665	0.676	**0.655**

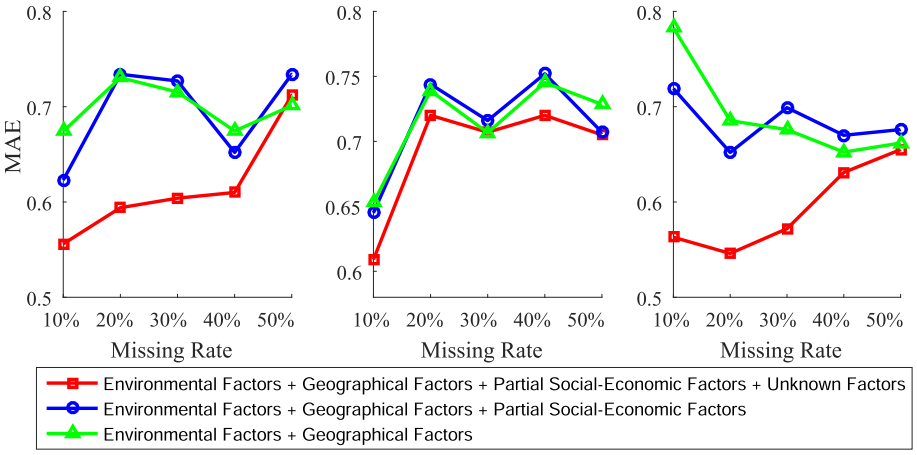


Fig. 5. The effect of integrating varying underlying risk factors in missing value imputation. If the unknown factor module in our method is turned off, incorporating only the environmental and geographical factors in the model has a comparable performance to incorporating environmental, geographical, and socioeconomic factors; if the unknown factor module is turned on, incorporating the extra socioeconomic factors obviously enhances performance.

the attributes of each town heterogeneous. Unlike HNML, which directly infers the disease dynamics based on the data with missing attributes, HNML-UF utilizes the unknown factor module to model the missing socioeconomic factors for the 44 towns based on the available information, and then uses the completed factors to infer the disease dynamics. As the missing socioeconomic factors are reasonably estimated via the unknown factor module and utilized for imputation, it is not surprising that HNML-UF achieves higher accuracy then HNML.

The above experiments demonstrate that incorporating the underlying risk factors as external resources via a machine learning method is helpful for inferring infectious disease dynamics when data are missing. We now examine the effect of integrating varying amounts of underlying risk factors for missing value imputation. Specifically, we consider three levels for the amount of factors to be incorporated:

1. We use only the environmental and geographical factors for all of the 62 towns, and use HNML (i.e., the unknown factor module is turned off) for missing value imputation (the results are marked as green lines in Fig. 5);
2. We use the environmental factors, geographical factors, and socioeconomic factors for the 18 towns in Tengchong, and the environmental and geographical factors for the remaining 44 towns, and use HNML (i.e., the unknown factor module is turned off) for missing value imputation (the results are marked as blue lines in Fig. 5);
3. We use the environmental factors, geographical factors, and socioeconomic factors for the 18 towns in Tengchong, and the environmental and geographical factors for the remaining 44 towns, and use HNML-UF (i.e., the unknown factor module is turned on) for missing value imputation (the results are marked as red lines in Fig. 5).

We observe that if the unknown factor module in our method is turned off, adding extra socioeconomic factors does not necessarily improve performance. If the unknown factor module in our method is turned on, then incorporating the extra socioeconomic factors obviously enhances performance.

We further use the Friedman's test and Holm's test to analyze the performance of different approaches in this challenging 62-town inference task. Following [40], we first count the mean ranks for dif-

Table 6
Holm's step-down procedure for analyzing the performance of HNML-UF against other methods.

i	Method	z	p	$\alpha/(k-i)$
1	GAIN	8.185	2.220E−16	0.013
2	KNN	8.096	6.661E−16	0.014
3	GP	7.566	3.841E−14	0.017
4	MEAN	5.710	1.131E−08	0.020
5	Kriging	3.677	2.360E−04	0.025
6	ST-MVL	3.500	4.649E−04	0.033
7	ST-ResNet	1.998	0.0457	0.050
8	HNML	1.556	0.1198	0.100

ferent methods: MEAN (4.5), KNN (7.3), Kriging (4.4), GP (7.6), ST-MVL (5.9), GAIN (7.7), ST-ResNet (3.3), HNML (3.0), HNML-UF (1.2). Using Matlab statistic toolbox, the Friedman test statistic is 93.5168 and the p-value is 8.9673e-17. Thus we reject the null hypothesis ($8.9673e-17 < 0.1$), which means the performances of these methods are significantly different. Then we use Holm's test for post hoc tests with confidence level $\alpha = 0.1$. We compare HNML-UF with other methods. Table 6 illustrates the Holm's step-down procedure. We can reject the first 7 null hypotheses, but we cannot reject the 8_{th} hypothesis, which means HNML-UF significantly outperforms the other methods except HNML.

6. Conclusion and future work

In this study, we developed a machine learning method for making inferences based on incomplete data to support decision-making in infectious disease control and prevention. To solve this challenging yet important problem, we presented a machine learning method that incorporates location-specific attributes of underlying disease-related risk factors collected from heterogeneous data sources. To evaluate the performance of our method, we conducted experiments based on a real-world dataset, i.e., the Yunnan malaria dataset. We compared the results of our method with several representative inference methods under three typical scenarios in which some data points are missing. Our method displayed statistically significant improvements over the existing alternatives, supporting our claim that underlying disease-related risks, such as environmental, geographic, and socio-economic factors, can provide useful information about infectious disease transmission patterns and should be considered when inferring the current/future number of infectious disease cases in unobserved locations.

Although we focused mainly on evaluating our proposed method on malaria, the proposed method is able to model various spatio-temporal diffusion dynamics, which are also affected by multiple complex factors. For instances, empirical studies has revealed that the transmission of dengue, a vector-borne disease, is influenced by the entomologic, demographic, and environmental risk factors [41]; the urban air quality is under the effect of the traffic condition and meteorology [42]; and the urban crow flow is affected by the weather, points of interest, and external events [2]. By systematically incorporating the data sources of related factors and adapting to domain-specific dynamic models, our method could flexibly model the spatio-temporal diffusion patterns of various spatio-temporal dynamics with incomplete data. We will further employ our method for modeling other infectious diseases transmission and other applications with similar spatio-temporal missing scenarios and heterogeneous data sources.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

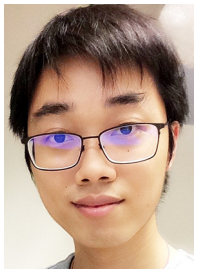
Acknowledgements

The authors would like to acknowledge the funding support from Hong Kong Research Grants Council (RGC/HKBU12201318, RGC/HKBU12201619, RGC/HKBU12202220) for the research work being presented in this article.

References

[1] Y. Matsubara, Y. Sakurai, W.G. van Panhuis, C. Faloutsos, Funnel: automatic mining of spatially coevolving epidemics, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 105–114.
[2] J. Zhang, Y. Zheng, D. Qi, Deep spatio-temporal residual networks for citywide crowd flows prediction, in: Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017, pp. 1655–1661.
[3] X. Yi, J. Zhang, Z. Wang, T. Li, Y. Zheng, Deep distributed fusion network for air quality prediction, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018, pp. 965–973.
[4] S.A. McDonald, B. Devleesschauwer, N. Speybroeck, N. Hens, N. Praet, P.R. Torgerson, A.H. Havelaar, F. Wu, M. Tremblay, E.W. Amene, D. Dpfer, Data-driven methods for imputing national-level incidence in global burden of disease studies, Bulletin of the World Health Organization 93 (4) (2015) 228–236.
[5] L. Qu, L. Li, Y. Zhang, J. Hu, Ppca-based missing data imputation for traffic flow volume: a systematical approach, IEEE Transactions on Intelligent Transportation Systems 10 (3) (2009) 512–522.
[6] Y. Li, L.E. Parker, Nearest neighbor imputation using spatial–temporal correlations in wireless sensor networks, Information Fusion 15 (2014) 64–79.
[7] S.K. Greene, E.R. Peterson, D. Kapell, A.D. Fine, M. Kulldorff, Daily reportable disease spatiotemporal cluster detection, New York city, New York, USA, 2014–2015, Emerging Infectious Diseases 22 (10) (2016) 1808.
[8] R. Senanayake, O. Simon Timothy, F. Ramos, Predicting spatio-temporal propagation of seasonal influenza using variational gaussian process regression, in: AAAI, 2016, pp. 3901–3907.
[9] B. Shi, J. Liu, X.-N. Zhou, G.-J. Yang, Inferring plasmodium vivax transmission networks from tempo-spatial surveillance data, PLoS Neglected Tropical Diseases 8 (2) (2014) e2682.
[10] World Health Organization, Disease surveillance for malaria elimination: operational manual, 2012.
[11] B. Yang, H. Guo, Y. Yang, B. Shi, X. Zhou, J. Liu, Modeling and mining spatiotemporal patterns of infection risk from heterogeneous data for active surveillance planning, in: Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.
[12] L. Fumanelli, M. Ajelli, P. Manfredi, A. Vespignani, S. Merler, Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread, PLoS Computational Biology 8 (9) (2012) e1002673.
[13] F. Simini, M.C. González, A. Maritan, A.-L. Barabási, A universal model for mobility and migration patterns, Nature 484 (7392) (2012) 96–100.
[14] J. Cao, H.J.W. Sturrock, C. Cotter, S. Zhou, H. Zhou, Y. Liu, L. Tang, R.D. Gosling, R. G.A. Feachem, Q. Gao, Communicating and monitoring surveillance and response activities for malaria elimination: China's 1-3-7 strategy, PLOS Medicine 11 (5) (2014) 1–6.
[15] T. Wu, Y. Li, Spatial interpolation of temperature in the united states using residual kriging, Applied Geography 44 (2013) 112–120.
[16] J. Wang, A. Hertzmann, D.M. Blei, Gaussian process dynamical models, in: Advances in Neural Information Processing Systems, 2005, pp. 1441–1448.
[17] L. Beretta, A. Santaniello, Nearest neighbor imputation algorithms: a critical evaluation, BMC Medical Informatics and Decision Making 16 (3) (2016) 74.
[18] S. Banerjee, B.P. Carlin, A.E. Gelfand, Hierarchical Modeling and Analysis for Spatial Data, CRC Press, Boca Raton, 2014.
[19] R.J.M. Buendia, G.A. Solano, A disease outbreak detection system using autoregressive moving average in time series analysis, in: 2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA), 2015, pp. 1–5.
[20] X. Yi, Y. Zheng, J. Zhang, T. Li, St-mvl: Filling missing values in geo-sensory time series data, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016, pp. 2704–2710.
[21] W. Vach, M. Blettner, Missing data in epidemiologic studies, Encyclopedia of Biostatistics 5.
[22] R.J. Little, D.B. Rubin, Statistical Analysis with Missing Data, John Wiley & Sons, Hoboken, New Jersey, 2014.
[23] S. Mandal, R.R. Sarkar, S. Sinha, Mathematical models of malaria-a review, Malaria Journal 10 (202).
[24] D.L. Smith, T.A. Perkins, et al., Recasting the theory of mosquito-borne pathogen transmission dynamics and control, Transactions of the Royal Society of Tropical Medicine and Hygiene 108 (4) (2014) 185–197.
[25] A. Wesolowski, N. Eagle, A.J. Tatem, D.L. Smith, A.M. Noor, R.W. Snow, C.O. Buckee, Quantifying the impact of human mobility on malaria, Science 338 (6104) (2012) 267–270.
[26] Q. Tan, J. Liu, B. Shi, Y. Liu, X.-N. Zhou, Public health surveillance with incomplete data – spatio-temporal imputation for inferring infectious disease

dynamics, in: Proceedings of the 6th IEEE International Conference on Healthcare Informatics, IEEE, 2018.
[27] Y. Zhang, W. Cheung, J. Liu, A unified framework for epidemic prediction based on poisson regression, IEEE Transactions on Knowledge and Data Engineering 27 (11) (2015) 2878–2892.
[28] E.A. Nadaraya, On estimating regression, Theory of Probability & Its Applications 9 (1) (1964) 141–142.
[29] J. Wiens, E.S. Shenoy, Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology, Clinical Infectious Diseases 66 (1) (2018) 149–153.
[30] A.N. Baraldi, C.K. Enders, An introduction to modern missing data analyses, Journal of School Psychology 48 (1) (2010) 5–37.
[31] B. Shi, Q. Tan, X.-N. Zhou, J. Liu, Mining geographic variations of Plasmodium vivax for active surveillance: a case study in China, Malaria Journal 14 (1) (2015) 216.
[32] X. Wang, L. Yang, T. Jiang, B. Zhang, S. Wang, X. Wu, T. Wang, Y. Li, M. Liu, Q. Peng, W. Zhang, Effects of a malaria elimination program: a retrospective study of 623 cases from 2008 to 2013 in a chinese county hospital near the china myanmar border, Emerging Microbes & Infections 5 (1) (2016) e6.
[33] M. Dhimal, R.B. O'Hara, R. Karki, G.D. Thakur, U. Kuch, B. Ahrens, Spatio-temporal distribution of malaria and its association with climatic factors and vector-control interventions in two high-risk districts of nepal, Malaria Journal 13 (1) (2014) 457.
[34] S. Greenland, W.D. Finkle, A critical look at methods for handling missing covariates in epidemiologic regression analyses, American Journal of Epidemiology 142 (12) (1995) 1255–1264.
[35] J.A.C. Sterne, I.R. White, J.B. Carlin, M. Spratt, P. Royston, M.G. Kenward, A.M. Wood, J.R. Carpenter, Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, BMJ 338.
[36] N.J. Perkins, S.R. Cole, O. Harel, E.J. Tchetgen, B. Sun, E.M. Mitchell, E.F. Schisterman, Principled approaches to missing data in epidemiologic studies, American Journal of Epidemiology 187 (3) (2018) 568–575.
[37] W.R. Myers, Handling missing data in clinical trials: an overview, Drug Information Journal 34 (2) (2000) 525–533.
[38] J. Zhang, Y. Zheng, D. Qi, R. Li, X. Yi, T. Li, Predicting citywide crowd flows using deep spatio-temporal residual networks, Artificial Intelligence 259 (2018) 147–166.
[39] J. Yoon, J. Jordon, M. Schaar, Gain: Missing data imputation using generative adversarial nets, in: International Conference on Machine Learning, 2018, pp. 5675–5684.
[40] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (Jan) (2006) 1–30.
[41] G. Zhu, J. Liu, Q. Tan, B. Shi, Inferring the spatio-temporal patterns of dengue transmission from surveillance data in Guangzhou, China, PLoS Neglected Tropical Diseases 10 (4) (2016) e0004633.
[42] H.-P. Hsieh, S.-D. Lin, Y. Zheng, Inferring air quality for station location recommendation based on urban big data, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 437–446.



Qi Tan received the B.S. from the School of Computer Science and Engineering, South China University of Technology, in 2014. He is pursuing the Ph.D. degree in the Department of Computer Science, Hong Kong Baptist University. His research interests include spatio-temporal data mining and social network analysis with applications for health informatics.

Yang Liu received the B.S. and M.S. degrees in Automation from National University of Defense Technology in 2004 and 2007, respectively. He received the Ph.D. degree in Computing from The Hong Kong Polytechnic University in 2011. Between 2011 and 2012, he was a Postdoctoral Research Associate in the Department of Statistics at Yale University. Dr. Liu is currently an Assistant Professor in the Department of Computer Science at Hong Kong Baptist University. His research interests include artificial intelligence, machine learning, as well as their applications in high-dimensional data mining, complex network analysis, and infectious disease modeling.

Jiming Liu received the MEng and PhD degrees from McGill University. He is currently the chair professor of computer science at Hong Kong Baptist University. His research interests include data analytics, machine learning, complex network analytics, data-driven complex systems modeling, and health informatics. He has served as the editor-in-chief of the Web Intelligence Journal (IOS), and an associate editor of the Big Data and Information Analytics (AIMS), the IEEE Transactions on Knowledge and Data Engineering, the IEEE Transactions on Cybernetics, Neuroscience and Biomedical Engineering (Bentham), and Computational Intelligence (Wiley), among others. He is a fellow of the IEEE.

Benyun received the BSc. degree in Mathematics from Hohai University, Nanjing, China, in 2003, and the M.Phil and Ph.D. degrees in Computer Science from Hong Kong Baptist University in 2008 and 2012. He is currently the professor of School of Computer Science and Technology, Nanjing Tech University. His research interests include Multi-agent Autonomy-Oriented Computing (AOC), Real-world Complex Systems Modeling, Complex Networks, particularly for Energy Distribution Systems and Infectious Disease Epidemiology.

Shang Xia received the PhD degree in computer science from Hong Kong Baptist University. He is currently an associate professor in the National Institute of Parasitic Diseases, Chinese CDC. His research interests include computational epidemiology and health informatics.

Xiao-Nong Zhou obtained his PhD in Biology at University of Copenhagen, Denmark in 1994, following his MSc in Medical Parasitology from Jiangsu Institute of Parasitic Diseases in China. He is a professor and the director of the National Institute of Parasitic Diseases at the Chinese Center for Disease Control and Prevention, based in Shanghai, China. He is a leading expert in the research and control of schistosomiasis and other infectious diseases, with over 30 years' experience in the field. Professor Zhou established a career in infectious disease research across the fields of ecology, population biology, epidemiology, and malacology.