IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 25, 2019, accepted July 5, 2019, date of publication July 10, 2019, date of current version August 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927878

Mining Spatiotemporal Diffusion Network: A New Framework of Active Surveillance Planning

HECHANG CHEN[1,2], BO YANG[1,2], JIMING LIU[3], (Fellow, IEEE), XIAO-NONG ZHOU[4], AND PHILIP S. YU[5,6], (Fellow, IEEE)

[1]College of Computer Science and Technology, Jilin University, Changchun 130012, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Changchun 130012, China
[3]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[4]Chinese Center for Diseases Control and Prevention, National Institute of Parasitic Diseases, Shanghai 200025, China
[5]Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA
[6]Institute for Data Science, Tsinghua University, Beijing 100084, China

Corresponding author: Bo Yang (ybo@jlu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61876069 and Grant 61572226, and in part by the Jilin Province Key Scientific and Technological Research and Development Project under Grant 20180201067GX and Grant 20180201044GX.

ABSTRACT Infectious diseases pose a constant and serious threat to human life. One way to prevent infectious disease spread is through active surveillance: monitoring patients to discover disease incidences before they get out of hand. However, active surveillance can be difficult to implement, especially when the monitored area is vast and resources are limited. Incidences of infectious disease that arrive with visitors from abroad are a further challenge. When faced with imported incidences and a large region to monitor, it is critical that public health authorities precisely allocate their sparse resources to high-priority areas to maximize the efficacy of active surveillance. In this paper, the difficulties of active surveillance are considered, and we offer a computational framework to address these challenges by modeling and mining the spatiotemporal patterns of infectious risks from heterogeneous data sources. Malaria is used as an empirical case study (with real-world data) to validate our proposed method and enhance our findings.

INDEX TERMS Active surveillance, diseases control, spatiotemporal patterns, spatiotemporal diffusion networks.

I. INTRODUCTION

Globally, prevention and control of infectious diseases is a top priority in public health around the world. Despite centuries of work toward cures and treatments, 13 million people die from infectious diseases each year, mostly in developing countries [2]. The World Health Organization (WHO) estimates that 3.3 billion people are infected with malaria, for instance, of which 1.2 billion are life threatening cases. In 2013 alone, 198 million people were infected and 584,000 died [3]. As infectious disease is a serious threat to human life and a heavy social and economic burden, policymakers aim to reduce the risk of infection through intervention strategies.

There are two common strategies for prevention and control of infectious disease: passive infectious disease surveillance and active infectious disease surveillance. Passive surveillance records incidents through patient reports to public health agencies. In the past, this method was used to monitor the pandemic influenza outbreak [4], however, researchers found that the confirmed influenza infection rate was much higher than the rate reported for hospitalization [5]. For instance, the confirmed infection rate for Hong Kong's H1N1 outbreak was 22.4 times higher than the rate reported for hospitalization [6]. These studies show that passive surveillance does not discover or enable the timely treatment of infected individuals, which leads to a loss of life and is socioeconomically detrimental, especially in cases of emergent diseases. To address the challenges of passive surveillance strategies such as door-to-door surveys

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

or actively screening patients, i.e., active infectious disease surveillance, have been introduced in some departments, particularly in regions at a high risk of infection. For example, active surveillance methods are widely used to eliminate the spread of malaria and dengue fever in developing countries [7], [8], and to prevent the outbreak of infectious diseases after natural disasters [9], [10]. Active surveillance is more effective than passive surveillance in preventing the spread of infectious disease for two main reasons. First, active surveillance can collect relatively complete data, and therefore help us to understand the extent of ongoing infectious diseases. Second, the data collected through active surveillance are more timely, and thus not only provide real-time information about disease symptoms, distribution, and diffusion trends, but also can effectively prevent secondary infections.

However, active surveillance is difficult to implement in large monitoring areas when resources like antibiotics and health care workers are very limited, particularly in remote and impoverished regions. The situation is even worse in cases of imported disease, because incidents caused by human mobility behaviors are very difficult to discover and prevent. With all of these challenges, it is essential for public health authorities to carefully ration their limited resources by determining when and where should be monitored with high priority to maximize the outcomes of active surveillance. Existing infectious disease prediction methods based on biology [11], [12], statistics [13], [14], and machine learning [15], [16] aim to fit the model by considering biological characteristics and weather conditions. Existing infectious disease diffusion network modeling methods collect location data in an attempt to model the diffusion process with mobile phones [17], [18], wearable sensors [19], [20], and satellites [21], [22]. However, none of these methods can explain why so many infectious cases are imported, how those cases are generated, or what socioeconomic factors dominate the generation of infection. To answer these questions, a model that can explain human mobility patterns and reveal the principles of spatiotemporal patterns of disease spreading is needed.

To address the aforementioned problems, we aim to predict the risks of infectious disease by modeling and mining human mobility patterns from heterogeneous data for use in active surveillance planning. The basic idea was briefly introduced in a conference paper [1] and has been implemented to demonstrate its potential applications [23]. Here, the basic idea is systematically introduced and greatly extended by supplementing new models, new algorithms, theoretical analysis, and sufficient validations. The main contributions of this study are summarized as follows.

1) A framework of active surveillance planning for imported infections (ASPII) that hinges on the modeling and mining spatiotemporal patterns of infectious risks is proposed. To the best of our knowledge, this is the first effort to propose a general framework for active surveillance planning, which fuses heterogeneous data and integrates combined models, as shown in Fig. 1.

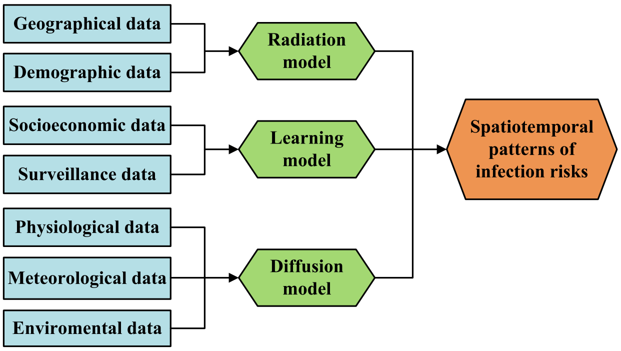


FIGURE 1. An overview of heterogeneous data mining with combined models to predict the spatiotemporal patterns of infection risks. Demographic and geographical data are used to construct a radiation model. Socioeconomic and surveillance data are used to construct a learning model. Physiological, meteorological and environmental data are used to construct an epidemic diffusion model.

2) A mixture optimization method is proposed as a way to accurately discover high-priority regions. This method aims to estimate the parameters of socioeconomic factors and cluster monitoring regions adaptively, so as to accurately predict infectious risks when monitoring regions with different regional characteristics.

3) Real-world data are collected and used to extensively validate the proposed models and methods. Using malaria as a case study, a variety of data from Tengchong county, Yunnan province, China, and ten regions of Myanmar is used to test the efficacy of the proposed method.

The remainder of this paper is organized as follows. Section 2 outlines the task of active surveillance planning and explains our proposed framework by introducing a novel concept of a spatiotemporal diffusion network and corresponding modeling and mining methods. In Section 3, we validate the proposed framework, models and methods using real-world data of malaria infection as a case study. Section 4 discusses the connections between our work and the existing literature. Finally, Section 5 recaps our work by highlighting its main features and contributions, and introducing possible extensions of the novel framework.

II. ACTIVE SURVEILLANCE PLANNING FRAMEWORK

In this section, the computational framework of active surveillance planning for imported infections (ASPII) is proposed by elaborating on the following goals: defining the task of active surveillance planning; modeling the spatiotemporal patterns of infection risks; predicting risks based on the use of data to discover high-priority areas; alleviating the influence of spatial heterogeneity on risk predictions; successfully embedding the epidemiological model of a specific disease into a computational framework.

A. PROBLEM DEFINITION

The main objective of active surveillance planning is to determine where and when to search for cases of infection so as to maximize the effectiveness of available surveillance resources. Formally, we define this planning task as an optimization problem as follows:

Out of all regions of interest (ROI), to select the minimum number of targets that are prioritized to scan, which will cover a predefined percentage, according to the given available resources, of all potential incidences within a period in the future.

To address this task, we propose an infection-risk-based planning framework which consists of two procedures: prediction and planning. In prediction, the infection risk, defined as the probability that an individual will be infected in specific region, will be evaluated for further planning. In planning, regions with high infection risks will be properly selected for planning to maximize the outcome of available resources. Specifically, the framework is to be carried out via the following steps:

Step I: predict the infection risk for each region within a specified time window. For example, a quarter for a short-term planning, or more challenging, a year for a long-term planning;

Step II: estimate the number of infection incidences for each region within the time window as the product of population and infection risk;

Step III: rank all regions in descending order according to their respective incident numbers;

Step IV: select a minimum k so that the overall infection ratio of the top-k regions is above a predefined threshold;

Step VI: output the top-k regions as high-priority targets to be searched for the given time window.

The first procedure of prediction is the foundation of the proposed framework: given the features of regions, e.g., socioeconomic factors, the epidemic diffusion model, and the surveillance data, the objective is to predict the spatiotemporal patterns of infection risks. We will further describe the above problem definition by presenting a model to represent the spatiotemporal patterns of infection risks, and an algorithm to mine such patterns from heterogeneous data.

B. MODELING SPATIOTEMPORAL PATTERNS OF INFECTION RISKS

Human mobility greatly influences the generation and distribution of infection, which should be sufficiently considered during modeling.

We characterize human mobility in terms of a transition matrix, in which entries depict the likelihood of movement from one region to another. This matrix can be directly constructed from data recording the frequency of movement between locations. In practice, however, it is very difficult to get such detailed and sufficient information for privacy reasons, especially information on those migrating across international borders. In light of this, our model will treat the transition matrix as a hidden variable to be estimated, which is collectively regulated by socioeconomic, geographical and transportation factors.

In the course of elimination of infectious disease, the infection cases are increasingly observed as imported. For example, in Tengchong County, Yunnan Province, China, more than 98% of reported malaria infections from 2005 to 2011 were confirmed as imported cases from Myanmar. This was also the case for dengue, cholera and avian influenza. As reported, most dengue cases in Guangzhou (a big city in south China) were imported from southeast Asia (Vietnam, Thailand, Myanmar) through traveling [39]. All cholera cases in Caracas (capital of Venezuela) reported in 2011 were imported from areas of Dominica [40]. In the early stage of avian influenza (H1N1 in 2009 and H7N9 in 2013), reported cases in many cities in China were imported [41], [42]. The propagations of such diseases (where imported cases are dominant during an entire outbreak or early stages of an outbreak) are similar: some areas have a high infection risk for a certain disease, persons from other areas go into these areas for work, travel, or other purposes and get infected, and infection cases are imported when they return to their original places.

More generally, human mobility behaviors driving the importation-featured epidemic process can be properly modeled using spatiotemporal diffusion network that contains multiple types of nodes and links, as shown in Fig. 2.

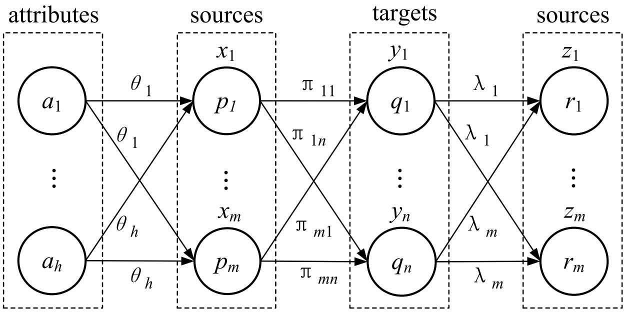


FIGURE 2. A spatiotemporal diffusion network. *p*, π, *q* and λ represent the probabilities of whether to go out, where to go, whether to be infected, and when to go back, respectively, corresponding to the four phases of human mobility.

In the network, node a_i denotes one of h socioeconomic attributes driving people to migrate for work or travel. Node x_i denotes one of m source locations, and node weight p_i depicts the likelihood that residents of the source location migrate out. According to surveys, in most imported cases the main purpose of migration is for short-term work [27], [28]. Correspondingly, we assume that p_i will be driven by a non-linear combination of h socioeconomic attributes in terms of a set of weights $\theta = (\theta_1, \ldots, \theta_h)$, where $\theta_j (1 \leq j \leq h)$ characterizes the impact of attribute a_j on p_i. Link weight π_{ij} depicts the probability of a resident of x_i going to y_j, one of n target locations. Node weight q_i depicts the risk of getting infected in y_i. Link weight λ_i depicts the temporal probability distribution of a resident going back to their source location z_i from target locations after working. More specifically, λ_i is a T-dimension vector $(\lambda_{i1}, \ldots, \lambda_{it})$, where $\lambda_{it} (1 \leq t \leq T)$ denotes the probability of going back to the source location z_i in the time interval t. The number of time intervals is set according to the time scale of surveillance planning and the granularity of available surveillance data. For example, one can set $T = 4$ for seasonal planning or set $T = 12$ for monthly planning. Node z_i denotes the same source location as x_i, while its weight r_i depicts the ratio of infected cases in the location. Note that r varies with both space and time due to monthly or annually variations in input data. That is why we say the vector r (embedded throughout the spatiotemporal diffusion network) represents both spatial and temporal patterns of infection risks.

Notably, along with the expected spatiotemporal pattern of infection risks provided by vector r, the dominant socioeconomic factors in terms of vector θ and the hidden transition matrix in terms of vector p and matrix π can also be estimated simultaneously from this network.

C. MINING SPATIOTEMPORAL PATTERNS OF INFECTION RISKS

To determine the parameters of the spatiotemporal diffusion network from available data, we will need to mine the spatiotemporal patterns of infection risks.

The task is not trivial because most of the parameters cannot be directly figured out due to a lack of data. Before presenting detailed formulas, we will briefly introduce our basic strategies of parameter estimation. Vector p can be calculated with a regression model using the input of socioeconomic data and weights θ. We suggest to use logistic regression to represent the causality between attributes a and p, because in our preliminary studies we found that many socioeconomic attributes are in an approximate logistic-linear scale to the actual infection risks. Once residents make the decision to migrate for work, they should already know where they will go. In real data, a very small portion of infected cases reported where they were working, so it is not reliable to statistically infer vector π from such a little amount of data. Instead, a more reasonable π value can be estimated with an economical job-finding model [29], [30], such as a radiation model that uses inputs of demographical, geographical, and transportation data of both source location x and target location y. Without surveillance data of targets, the infection ratios in various locations cannot be directly counted. Using malaria as an example, we can turn to biological and epidemiological models, such as a vector capacity model [12], to estimate vector q by inputting environmental and meteorological data of target locations, including variables such as temperature, rainfall and humidity. In practice, we do not have data that explicitly records the distribution of λ_i, i.e., how many people will return from target locations during a certain period for each source location. Alternatively, the distribution can be estimated based on surveillance data by assuming that the ratio of residents going back to their source location z_i from target locations during an time interval will be approximately proportional to the infection ratio of the same interval at z_i. In this regard, the time interval should be set long enough to cover the incubation time of the specific infectious disease. For example, in our case study the interval is set to three months ($T = 4$) to safely cover the longest incubation interval of Plasmodium vivax, the causative agent of malaria. Thus, vector r can be represented as a multiplication of p, π, q and λ. Finally, we infer weights θ from surveillance data using a statistical inference method such as maximum likelihood estimation (MLE).

According to the logistic regression model, for source location x_i we have:

$$p_i = g(\theta X_i) = \frac{e^{\theta X_i}}{1 + e^{\theta X_i}} \tag{1}$$

where $X_i = (x_{i1}, \ldots, x_{ih})^{\mathrm{T}}$ and each component x_{ij} denotes the value of attribute a_j of location x_i.

The above model indicates that all source locations share a common θ to regulate their respective p_i. That is, each attribute will have an equal impact on different locations. However, this might not be reasonable due to the variation of socioeconomic levels in distinct locations. To characterize this variation, we relax this constraint by introducing a cluster based logistic regression model. In this model, source locations are clustered so that locations within the same cluster share a common θ to regulate their respective p_i. Otherwise, θ will be different. This flexible model enables us to provide a more detailed causal analysis, i.e., determining which socioeconomic factors will dominate the infection of a specific location, and then plan a more targeted intervention for the location accordingly.

Assume that m source locations are assigned to Ω clusters. Let $Z = (z_{ij})_{\Omega \times m}$ be an indicator for the cluster, where $z_{ij} = 1$ if location x_j belongs to cluster i. Otherwise it equals zero. In terms of Z, Eq. 1 can be rewritten as:

$$p_i = g(Z_i^T \theta X_i) = \frac{e^{Z_i^T \theta X_i}}{1 + e^{Z_i^T \theta X_i}} \tag{2}$$

where both Z_i and X_i are long vectors and θ is a matrix. Z_i corresponds to the i-th column of Z. Note that, in this model, θ is extended from a vector to a Ω by h matrix. For all locations within cluster i, the same weights $\theta_i = (\theta_{i1,} \cdots, \theta_{ih})$ are used to regulate their respective out-going probabilities. Eq. 1 is actually a special case of Eq. 2 when all locations are assigned to the same cluster.

We improved the population radiation model [29] to estimate π, as follows:

$$\pi_{ij} = \frac{pop_i \times pop_j}{(pop_i + s_{ij}) \cdot (pop_i + pop_j + s_{ij})} \tag{3}$$

where pop_i and pop_j are the populations of source location x_i and target location y_j, respectively. Let l_{ij} be the distance between x_i and y_j, and s_{ij} is the total population in the circle which has a radius of l_{ij} centered at x_i by excluding the target population. Fig. 3(a) is an example in which we assume that the man is considering where to work next, and that the location indicated by a red marker is one of the candidates. l_{ij} is the distance between the source and target location, and s_{ij} is the sum of the population within the orange area except the target location. After evaluating the feasibility of all the

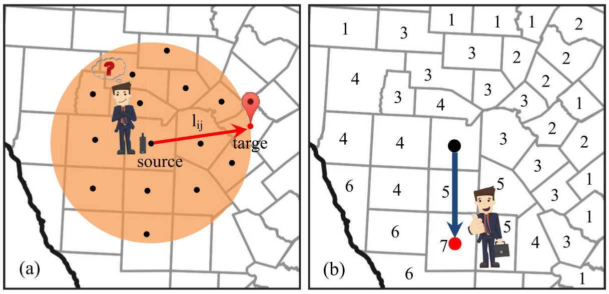


FIGURE 3. An example of the population radiation model. (a) Depiction of the evaluation process of where to go (target) from the source location. The source location is in the center and a candidate target is indicated by a red marker. (b) Final decision of where to go. The number contained in each region is the probability (%) of its selection.

candidate locations according to the radiation model, the final decision of where to go will be made by this man, as shown in Fig. 3(b).

In this framework, the infection ratios q in different target locations can be estimated through biological and epidemiological models that consider the specific characteristics of infectious diseases. See Section II.E for a case study of malaria. The distribution λ can be estimated using surveillance data.

Based on the above analysis, the infection risk of source location z_i at time interval t in a certain year u can be calculated by:

$$r_{it}^{u} = p_{it}^{u}(\sum_{j=1}^{n} \pi_{ij}^{u} q_{jt}^{u})\lambda_{it}^{u} \tag{4}$$

The total surveillance data of Y years can be represented as a cube tensor denoted as $C = [c_{it}^{u}]_{Y\times m\times T}$, where $C_{it}^{u} = r_{it}^{u} \cdot pop_i$ denotes the number of incidences reported at location z_i during the time interval t of year u, pop_i is the population of region i. In terms of parameters θ and Z, the likelihood of surveillance data C is:

$$\mathcal{L}(C;\theta,Z)=\prod_{u=1}^{Y}\prod_{i=1}^{m}\prod_{t=1}^{T}\binom{pop_i^{y}}{C_{it}^{u}}(r_{it}^{u})^{C_{it}^{u}}(1-r_{it}^{u})^{pop_i^{y}-C_{it}^{u}} \tag{5}$$

According to MLE, one can estimate θ and Z from C by solving the following constraint optimization problem:

$$\begin{aligned} &\max \log L(C;\theta,Z) \\ &s.t. \quad \forall j \sum_{i=1}^{\Omega} z_{ij} = 1, \ \forall i,j \ \ z_{ij} \in \{0,1\} \end{aligned} \tag{6}$$

Moreover, we add the group sparse feature of θ to the above objective function:

$$\mathcal{J} = -\log L(C;\theta,Z) + \varphi \sum_{k=1}^{h} ||\theta_{\cdot k}||_2 \tag{7}$$

where $\theta_{\cdot k} = (\theta_{1k},\ldots,\theta_{\Omega k})$ contains the weights of the k-th attribute for all the Ω location clusters. φ balances the two terms, i.e., data fitting and group sparsity. Remember, θ_i consists of the weights of h socioeconomic factors for location cluster i. With the penalty of group sparsity, which forces the weights of socioeconomic factors close to zero, one can obtain a sparse representation of each θ_i. In this way, the most important socioeconomic factors driving residents to migrate for work from different locations can be readily recognized.

D. OPTIMIZATION ALGORITHM

There are two parameters, θ and Z, need to be estimated to minimize the objective function Eq. 7. As θ is continuous while Z is discrete, we propose a mixture optimization method, in which θ and Z are estimated alternatively.

The parameter θ can be estimated by performing the following gradient descents iteratively,

$$\theta = \theta - \alpha \cdot \frac{\partial \mathcal{J}}{\partial \theta} \tag{8}$$

where α is a predefined learning rate.

The partial derivative is calculated as follows:

$$\frac{\partial \mathcal{J}}{\partial \theta_{i\cdot}} = -\sum_{u=1}^{Y}\sum_{t=1}^{T} f(Z_{it}^{u}X_{it}^{u}) + \theta_{i\cdot} \cdot \left(\sum_{k=1}^{\Omega}\theta_{ik}^{2}\right)^{-1/2} \tag{9}$$

where

$$\begin{cases} f(S)=\dfrac{c_{it}^{u}\cdot S\cdot e^{Z_i\theta X_i}}{1+e^{Z_i\theta X_i}}-\dfrac{(pop_i^{(u)}-c_{it}^{u})\cdot S\cdot e^{Z_i\theta X_i}}{(1+e^{Z_i\theta X_i}-\omega)(1+e^{Z_i\theta X_i})}\cdot\omega \\ \omega=\lambda_{it}^{(u)}\sum_{j=1}^{n}\pi_{ij}^{(u)}q_{jt}^{(u)} \end{cases}$$

We adopt the simulated annealing method to estimate Z. We first calculate the empirical risks, or fitting errors, for each source location based on the current clustering indicator, denoted as Z_0, and rank all locations in descending order according to their empirical risks. We then randomly reassign the top L locations to new clusters, and obtain an updated clustering indicator, denoted as Z_1. Then the objectives $\mathcal{J}_0$ and $\mathcal{J}_1$ are calculated based on Z_0 and Z_1, respectively. We accept Z_1 if $\mathcal{J}_1 < \mathcal{J}_0$, and if not, we accept Z_1 with a probability $exp((\mathcal{J}_0 - \mathcal{J}_1)/I)$, where I is the number of remaining iterations. The complete optimization algorithm is given in Algorithm 1. $rand(a, b)$ returns a random number between a and b. $E(i)$, $pred(i)$ and $real(i)$ are the empirical risk, predicted case number, and real number of cases in a location i, respectively.

A cross-validation method is adopted to determine a reasonable Ω, i.e., the number of clusters of source locations. In this method, we first bipartition whole learning data into a training set and testing set. The training set is used to estimate θ and Z for a given $\Omega(1 \leq \Omega \leq m)$. The testing set is used to evaluate the performance of such estimations by its infection risk prediction accuracy, based on Eq. 4. From all m candidates, we select the candidate with the best performance as the real number of clusters. With the selected Ω and corresponding θ and Z, infection risks can then be predicted and active surveillance for a given time interval can be planned accordingly.

Algorithm 1 The Alternative Optimization of θ and Z

```
Input: $X, \pi, q, \lambda$
Output: $\theta, Z$
parameters: $\Omega, L, I_0, \alpha, \epsilon_0, \epsilon_1$
Initialize $\theta$ and $Z$ as $\theta_0$ and $Z_0$;
LOOP
[Optimize $\theta$ given $Z$]
        LOOP
            $\theta^{(t)} \leftarrow \theta^{(t-1)} - \alpha \cdot \frac{\partial \mathcal{J}}{\partial \theta^{(t-1)}}$;
        UNTIL $|\theta^{(t)} - \theta^{(t-1)}| < \epsilon_1$
[Optimize $Z$ given $\theta$]
    $I \leftarrow I_0$
    WHILE $I > 0$ DO
        $E(i) \leftarrow \frac{|pred(i)-real(i)|}{real(i)}$, $1 \leq i \leq m$;
        Sort all $m$ locations in a descending order
        according to $E$;
        $Z_1 \leftarrow Z_0$;
        FOR $i = 1$ TO $L$ DO
            $r \leftarrow rand(1, \Omega)$;
            $Z_1(i, :) \leftarrow 0; Z_1(i, r) \leftarrow 1$;
        END FOR
        Compute $\mathcal{J}_1$ according to Eq. 7 based on $Z_1$;
        IF $J_1 < J_0$ OR $rand(0, 1) < exp((\mathcal{J}_0$
        $-\mathcal{J}_1)/T)$
            $\epsilon \leftarrow |\mathcal{J}_1 - \mathcal{J}_0|; Z_0 \leftarrow Z_1; \mathcal{J}_0 \leftarrow \mathcal{J}_1$;
        END IF
        $I \leftarrow I - 1$;
    END WHILE
UNTIL $\epsilon < \epsilon_0$
```

E. EMBEDDING AN EPIDEMIOLOGICAL MODEL

Malaria is one of the most serious infectious diseases and typically occurs as imported cases in developing countries. Here we use malaria as an example to illustrate how to compute the infection ratios of target locations in terms of vector q using an epidemiological model and how to embed it into the framework proposed above.

The infection risk of malaria is determined by the ability of mosquitoes to transmit Plasmodium, generally referred to as vectorial capacity, which can be formally expressed by the following VCAP model [12]:

$$V = -(\mu \alpha^2)\rho^{\tau}/ln(\rho) \tag{10}$$

where V depicts the vectorial capacity in a certain area, μ is equilibrium mosquito density per human, α is the expected number of bites on humans per mosquito per day, ρ is the probability that a mosquito survives through one whole day, and τ is the extrinsic incubation period of malaria parasites or the time taken for extrinsic cycle completion.

The parameters of the VCAP can be determined by temperature and rainfall [31], two major environmental and meteorological factors triggering malaria epidemics in warm, semiarid and high altitude areas. Specifically, we have: $\mu = 10 * prct$, $\alpha = 0.7/gtr$, $\rho = 0.5^{1/gtr}$, $\tau = 111/(tep_{min} - 1/gtr - 16)$, $gtr = 365.5/(tep_{min} - 7.9) + 0.5$, where $prct$ and tem_{min} denote the rainfall and the lowest temperature of an area during a time interval. Moreover, the infection risk of a target location can be estimated as follows [11]:

$$q = (\beta V - \sigma)/(\beta V + \sigma \frac{\alpha}{\eta}) \tag{11}$$

where β is the probability that an uninfected human is infected after being bitten by an infectious mosquito, σ denotes the recovery rate of humans and η denotes the per-capita daily death rate of a mosquito that is equal to $ln(\rho)$.

Note that all of the parameters in the above models, except β and σ, can be determined based on rainfall and temperature. With the help of epidemiologists, we set $\beta = 0.5$ and $\sigma = 0.001$ in our empirical study, following their studies on malaria infection in Myanmar.

F. SELECTION OF TOP-K REGIONS

On the basis of the proposed active surveillance framework, the potential infection risks of the monitoring regions can be calculated by embedding a specific epidemiological model, e.g., VCAP model. Then top-k regions with high priority can be selected as follow-up monitoring objects, considering both the potential infection risks and available surveillance resources. Assuming that the amount of available resources τ can only cover part of these candidate regions. The resource to be consumed by region i is δ_i. The selection of top-k regions takes the following steps:

Step I: rank all candidate regions in descending order to obtain an orderly infectious risk vector $R = \{R_1, \ldots, R_m\}$;

Step II: calculate the total resources need to be paid if top-k regions were selected from R for monitoring, i.e., $\sum_{i=1}^{k} \delta_i$;

Step III: repeat Step II starting from $k = 1$ until $\sum_{i=1}^{k} \delta_i \geq \tau$, then, the number of top-k areas is determined.

Generally speaking, candidate regions with higher potential infection risk need to invest more resources, i.e., $R_i \propto \delta_i$. Because the number of people to be monitored is large and the frequency to be monitored is high in such regions.

III. A CASE STUDY OF ACTIVE SURVEILLANCE

In this section, the effectiveness of the proposed framework will be thoroughly validated using the Tengchong malaria epidemic as a real-world case study.

Yunnan province experienced the most serious malaria outbreak in China because of its climate, which is ideal for mosquito proliferation. From 1999 to 2005, Yunnan was ranked first nationally for the number of malaria cases [26]. Yunnan also shares a long international border with Myanmar, which has experienced one of the most severe malaria epidemics in Asia. Difficulties in malaria control and resurgent epidemics have been closely associated with the frequent migration of people across the border, which has no natural barriers [32]. One of the dominant causes of malaria infection in Yunnan is an increase in migration (often illegal) across the border from highly infected areas. Each year, the Yunnan-Myanmar border in Tengchong County is crossed over 10,000 times and from 2005 to 2011 more than 98%

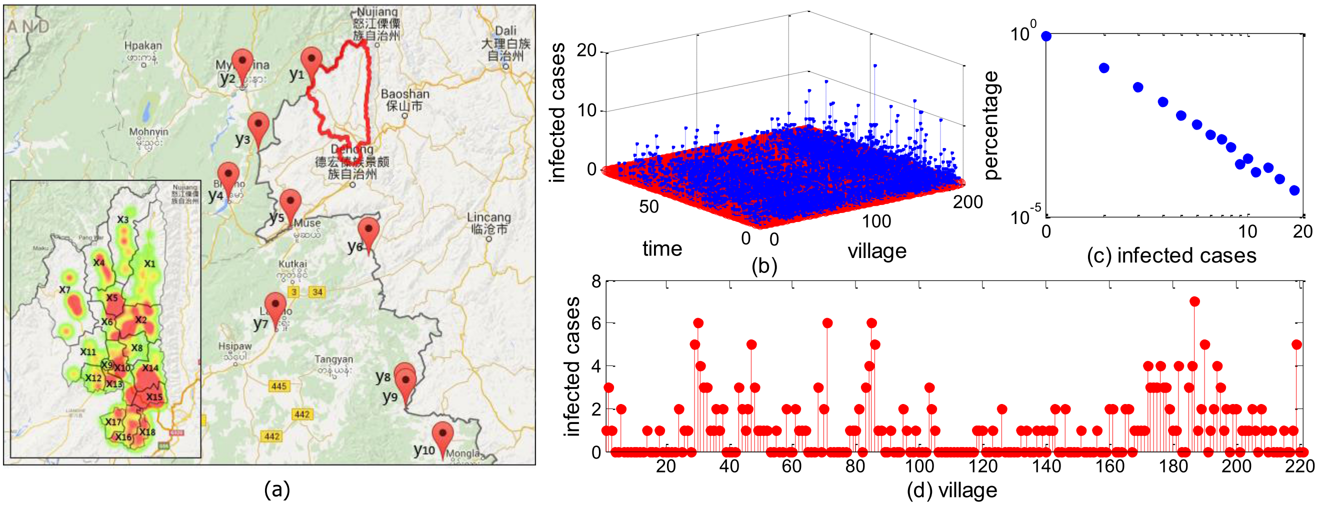


FIGURE 4. The spatiotemporal distribution of surveillance data. (a) Map of the Yunnan-Myanmar border. The red line is the border of Tengchong county, and the inset shows the distribution of infected cases in the 18 towns of Tengchong county, represented by a heat map. The locations indicated by red markers are the 10 candidate destinations going from Tengchong to Myanmar. (b) The land-scape of total monthly cases in 221 villages over seven years. (c) The power-law distribution of observed monthly infections in 221 villages over seven years. *x* is the number of infected cases and *y* is the percentage of villages with a certain number of cases. (d) The accumulated cases of respective villages in 2011.

of reported malaria infections were confirmed as imported cases from Myanmar. Tourism in Tengchong has also boomed recently. In 2012 alone, Yunnan hosted more than five million tourists. This increase in mobility significantly increases the risk of nationwide, or even worldwide, malaria spread.

Aside from significant social and economic changes in Yunnan, corresponding malaria control strategies have also changed. In addition to traditional passive surveillance and vector controls, active surveillance and intervention have also been introduced, particularly in regions with a high risk of infection. The key players implementing active surveillance are the local CDC and surveillance agencies, who visit villages and inquire door-to-door whether there is/has been a fever case. These local health agencies also perform surveys each fortnight to discover secondary cases before the start of the next cycle according to the incubation interval of P. vivax. For instance, the incubation interval of P. vivax is 12 to 18 days, while P. falciparum is 9 to 14 days [33]. Active surveillance is extremely expensive and time-consuming. It requires a massive group of experienced public health workers. However, resources are very limited, particularly in remote and poor regions. For instance, Tengchong has 18 towns (consisting of 221 villages), 167,964 households, and 658,207 residents that are distributed in a vast area of 5,845 square kilometers (as of 2011). Yet, in the local Tengchong CDC only a few workers or investigators are available to perform active surveys. Therefore, it is extremely important for public health authorities to know how to distribute their very sparse resources to high-priority regions to maximize active surveillance effectiveness.

As the vast majority (over 98%) of reported cases are imported from Myanmar, it seems that resources could be allocated by simply ranking villages by their cross-border migration numbers. However, in practice it is very difficult to get detailed information about how many people in a specific village, town, or county have passed through the border monthly or yearly, as there are over 20 official immigration channels and many more secret and illegal channels provided by snakeheads along the border. Therefore, in this case study we examine how the proposed framework can be used to estimate the spatiotemporal distribution of malaria risks and then reasonably allocate resources for active surveillance.

This task is particularly challenging because malaria transmission can be affected by multiple factors such as biology, environment, and meteorology, which directly impinge on interactions between hosts, vectors, and parasites at varying degrees and scales. The challenge is further exacerbated by human mobility, which is driven by various socioeconomic factors including income, food and meat production, agricultural population, and household size. Mobility plays a particularly important role in Yunnan's malaria distribution in that most of its cases are imported from neighboring countries rather than secondary infections from internal epidemics.

In the following, we first introduce the data collection from multiple sources, which are used for modeling and inference, and then conduct validations and analysis.

A. DATA COLLECTION AND PREPROCESS

Collecting sufficient and accurate data from multiple heterogeneous sources is quite challenging. The data sources we explored to mine the spatiotemporal diffusion network are summarized as follows.

For Tengchong County, surveillance data on monthly malaria cases at the village level for seven years (2005-2011) were obtained from the annual reports of the National Institute of Parasitic Disease, Chinese CDC. The data contains a total of 7,835 incidences distributed among 221 villages, as illustrated in Fig. 4(b). Annual demographic data at

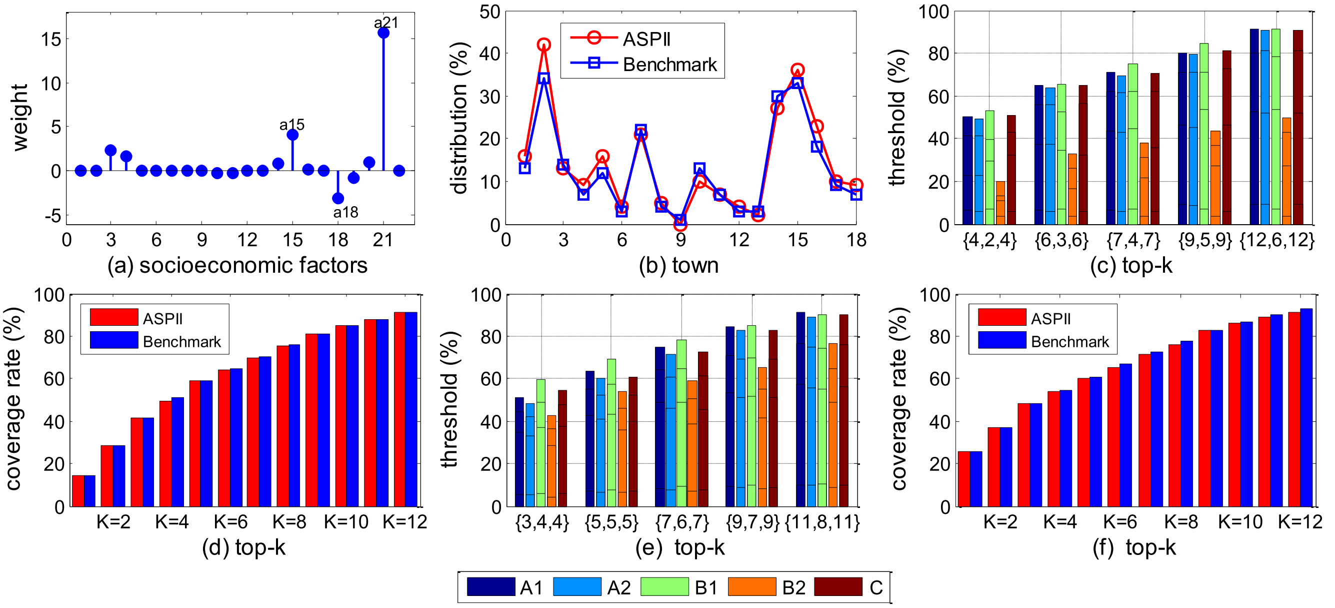


FIGURE 5. The effectiveness of active surveillance in 2010 and 2011. (a) Weights of socioeconomic attributes for the largest cluster in 2011. (b) Distributions of real and predicted cases of 18 towns for 2011. (c) Plans for active surveillance under different coverage thresholds in 2011. (d) Coverage rate from $K = 1$ to $K = 12$ in 2011. (e) Plans for active surveillance under different coverage thresholds in 2010. (f) Coverage rate from $K = 1$ to $K = 12$ in 2010.

the village level was obtained from the Chinese Natural Resources Database. Socioeconomic data at the town level were obtained from the annual reports issued by the Tengchong government, which contains a total of 22 socioeconomic attributes, denoted by $a_1, \ldots, a_{22}$, respectively.

Note that the surveillance data is quite sparse at the village level, particularly for 2010 and 2011, as illustrated by Fig. 4(c) and (d). From 2005 to 2011, more than 80% of villages have no monthly observed infections and the average infection incidence for each village per year is about five. As a result, a large bias of parameter estimation will be introduced if the model (Eq. 7) directly fits surveillance data at the village level. In addition, annual socioeconomic data at the village level is not available. For these reasons, in our empirical study we use 18 towns (consisting of 221 villages) as source locations, denoted by $x_1, \ldots, x_{18}$ (Fig. 4(a)).

Comparatively speaking, it is difficult to directly obtain Myanmar data because the government provides little official data. From supplemental information of surveillance data, we determined a total of 72 locations in Myanmar where individuals who had imported cases of the disease to Tengchong had been previously visited. Most of these locations are concentrated in 10 cities or towns in Myanmar near the Yunnan-Myanmar border, marked by red tags in Fig. 4(a), which are used as the target locations denoted by $y_1, \ldots, y_{10}$, respectively. The temperature, humidity and rainfall data of these target areas are obtained by integrating information from three sources: the IRI/LDEO Climate Data Library, Tropical Rainfall Measuring Mission (TRMM) and MODerate-resolution Imaging Spectroradiometer (MODIS). The last two datasets are provided by NASA, and useful data can be extracted from them with the remote sense image processing software ENVI (ENvironment for Visualizing Images). As missing values are contained in the data on temperature and humidity, a trigonometric polynomial curve fitting method is utilized for data completion. In this method, the value of temperature or humidity with respect to time t is formulated as $f(t) = a_0 + a_1 \cdot cos(t \cdot \omega) + a_2 \cdot sin(t \cdot \omega)$, where a_0, a_1, a_2 and ω can be estimated by historical data. The demographic and socioeconomic data of those targets are extracted and thereafter integrated from multiple online archives, such as Myanmar diaries, Tiptopglobe, Collins maps, and Wikipedia. Geographical and transportation data about the source and target locations were obtained from Google Earth.

B. VALIDATION OF SURVEILLANCE EFFECTIVENESS

We use surveillance data from year 2005 to year $(u - 1)$ for training and year u for testing, where $u \in \{2010, 2011\}$. We will first analyze how socioeconomic factors affect imported incidences in terms of the estimated θ, and then test the accuracy of infection risk prediction in terms of vector r and the effectiveness of active surveillance planning under different coverage thresholds.

The results are presented in Fig. 5 and Fig. 6. According to the estimated clustering indicator Z, 18 towns in Tengchong are clustered into six groups, as shown in Fig. 6, in which different colors represent different clusters. Accordingly, six weight vectors, i.e., $\theta_1, \ldots, \theta_6$, are obtained. As an example, Fig. 5(a) plots out θ_1, the weights of 22 socioeconomic attributes for the largest cluster consisting of eight towns, in which there are six positive weights, two negative weights

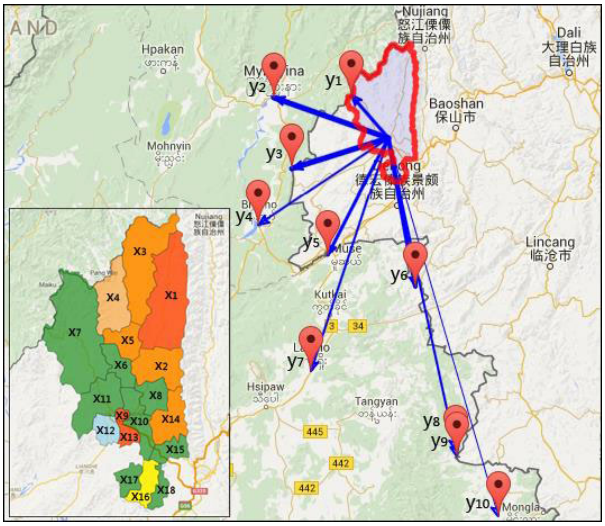


FIGURE 6. The inferred information of hidden cross-border migration distribution and the estimated clustering indicator *Z*. The width of the blue lines is proportional to the probability of migration from the source location to the destination. In the inset, color indicates group assignment, and locations with the same color are in the same cluster.

and 14 zero weights. Three dominant factors can be found from θ_1. Two of these are ''rural per capita net income (a_{21})'' and ''number of live pigs on hand at year's end (a_{15})'' with maximum positive weights, and the last is ''total agricultural machinery power (a_{18})'' with a maximum negative weight. According to Eq. 2, attributes with positive or negative weight will suppress or promote the probability of migrating for work, respectively. This implies that one can cut down the infection ratios of the eight towns by increasing their income level and the number of live pigs on hand at the end of the year, while decreasing their total agricultural machinery power. Note that most of the weights of socioeconomic attributes are close to zero, indicating that the group lasso penalty is effective to find out important attributes. Moreover, all these dominate factors recognized by are closely related to people's lives, which implies that the group lasso penalty is capable of finding dominate factors in line with the reality. Furthermore, according to the optimal θ estimated by the proposed mixture optimization method, the accumulated number of incidences for each town in 2011 can be predicted, as shown in Fig. 5(b). The blue line represents the actual values, which are normalized by the sum of total cases, and the red line represents the predictions by ASPII. It is clear that the prediction fits the truth quite well for most towns.

The predictions enable decisions about which towns should be considered high-priority in active surveillance, by ranking all of the towns according to the predicted incident numbers and selecting the highest risk towns based on a predefined threshold. Fig. 5(c) shows the plans of active surveillance for 2011 under five coverage thresholds respectively, 50%, 60%, 70%, 80%, 90% from left to right. Plan A is given by the proposed method integrating a cluster based logistic model (Eq. 2). A_1 and A_2 denote the coverage rates of predicted cases and real cases, respectively, of selected top-k towns with plan A. Specifically, A1 and A2 can be calculated by $\sum_{i=1}^{k} \hat{C}_{it}^{u}$ and $\sum_{i=1}^{k} C_{it}^{u}$, respectively, where i belongs to top-k regions given by A, and $\hat{C}_{it}^{u}$ and C_{it}^{u} are the predicted and real infection cases of region i at time t in year u. Plan B is given by the proposed method integrating a basic logistic model (Eq. 1) without clustering indicator Z. Similarly, B_1 and B_2 denote the coverage rates of predicted cases and real cases with plan B, respectively. As a benchmark to compare, plan C is given based on the real cases of 2011. The triplets on the x-axis denote the number of towns selected using the three plans. For example, in the case of threshold 50%, the triplet ''{4,2,4}'' in the leftmost denotes that plans A, B and C select top-4, top-2, and top-4 towns for active searching, respectively. Note that each bar of coverage rate consists of four portions, denoting the contributions of four quarters of a year from bottom to top, respectively. One can see that compared with plan B, the coverage rates of real cases (including the portions of four quarters) achieved by plan A are very close to the benchmark in all five cases. This indicates that the cluster based logistic model is more accurate in finding high-priority regions that are geographically heterogeneous.

We further compare the top-k towns selected by ASPII against the benchmark in terms of the coverage rates of real cases, as shown in Fig. 5(d). The coverage rates achieved by ASPII are still very close to the benchmark from top-1 to top-12. Similarly, Fig. 5(e) provides the plans of A, B and C for 2010 under five coverage thresholds, and Fig. 5(f) provides the coverage rates of top-k towns selected by ASPII for 2010 from top-1 to top-12.

The hidden distribution of cross-border migration estimated by p and π can serve as a reference for local government to prevent illegal stowaways. As an example, Fig. 6 shows part of the distribution in 2010, where arrow width indicates the magnitude of outgoing probabilities from Tengchong to targets y_i in Myanmar. It indicates that targets closer to the source location have larger outgoing probabilities. More than 80% of patients are middle-aged farmers who have crossed the border for short-term employment, e.g., logging and mining during slow farming seasons, according to statistics. This could be the main reason why residents from the source locations prefer to work in places closer to rather than farther from their homes.

C. VALIDATION OF CLUSTER-BASED REGRESSION MODEL

The inset in Fig. 6 shows the clustering of 18 townships obtained by ASPII when applied to the data for 2011, where the locations painted with the same color are allocated to the same cluster. In this case, most locations in the same cluster are geographically adjacent. The cluster based logistic regression model (Eq. 2) plays an important role in boosting the prediction performance. This can be validated by experiments, as shown in Fig. 7, in which the performance of ASPII is compared to ASPII-S in terms of seasonal infection risk prediction. ASPII-S is the same as ASPII except that the cluster indicator Z is predefined but not optimized. The predictions of ASPII fit the ground truth better than ASPII-S

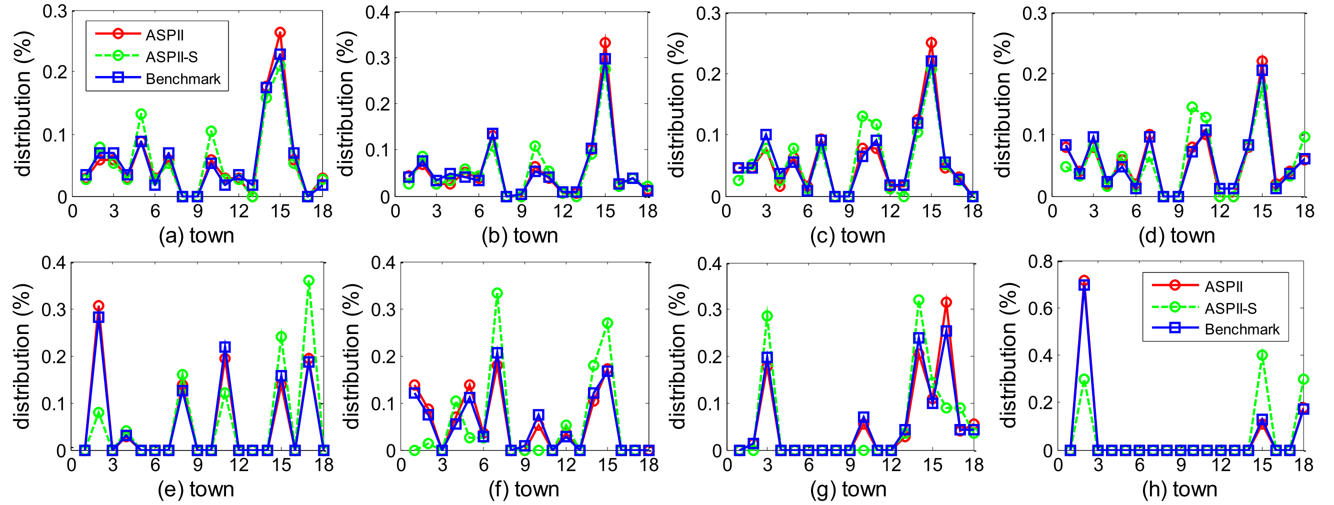


FIGURE 7. Distributions of seasonal infection cases in 18 towns. (a)-(d) are predictive results of seasonal infection risks for 18 towns in 2010. (e)-(h) are predictive results of seasonal infection risks for 18 towns in 2011. Rectangle-marked solid lines show the actual infection numbers normalized by total infections. Circle-marked solid lines show the predictions of ASPII, in which both θ and Z are optimized. Circle-marked dashed lines show the predictions of ASPII-S, as a comparison method, in which θ is optimized, while Z is predefined according to the clustering suggested by [1].

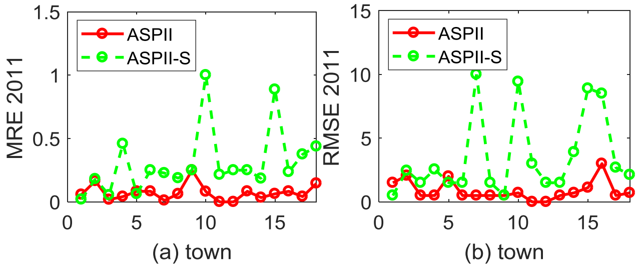


FIGURE 8. Comparison of ASPII and ASPII-S in terms of prediction error for year 2011. The x-axis represents the index of 18 townships and the y-axis represents the prediction error in terms of MRE and RSME.

for all seasons in years 2010 and 2011. Fig. 8 shows the prediction errors of the two methods for 2011 in terms of MRE (mean relative error) and RMSE (root mean squared error), which are defined as $MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}$ and $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$, where y_i and $\hat{y}_i$ are the actual and predicted number of infections respectively. Here we set n to 4, corresponding to the four seasons in a year. The predictions of ASPII are much better than ASPII-S in terms of both MRE and RMSE.

D. COMPARISON WITH MODEL-FREE STRATEGIES

To further check the effectiveness of ASPII, which is a model-based active surveillance, in this section we compare it to three model-free active monitoring strategies. These model-free strategies are based on either historical statistical data, or geographical features, or both.

1) HISTORICAL CASES-BASED ACTIVE SURVEILLANCE (H-AS)

According to H-AS, for a given region, the more infection cases that have occurred in the past, the higher the infection risk in the future, and thus the more likely it should be monitored. Accordingly, the infection risk for region i at time $t+1$ is calculated as:

$$r_{HAS}(i, t+1) = \frac{\sum_{u=1}^{Y}\sum_{t=1}^{T} C_{it}^{u}}{\sum_{i=1}^{m}\sum_{u=1}^{Y}\sum_{t=1}^{T} C_{it}^{u}}$$

As defined previously, C_{it}^{u} counts the real infected cases of region i at time t in year u, and m denotes the number of regions.

2) GEOGRAPHIC FEATURES-BASED ACTIVE SURVEILLANCE (G-AS)

According to G-AS, closer that a given region is to high-risk areas, the higher the future infection risk, and thus the greater the need for monitoring. Accordingly, the infection risk of region i at time $t+1$ is calculated as:

$$r_{GAS}(i, t+1) = \frac{\sum_{u=1}^{Y}\sum_{t=1}^{T}\sum_{j=1}^{n} l_{ij} q_{it}^{u}}{\sum_{i=1}^{m}\sum_{u=1}^{Y}\sum_{t=1}^{T}\sum_{j=1}^{n} l_{ij} q_{it}^{u}}$$

l_{ij} is the distance between location i in Tengchong and location j in Myanmar. $l_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}$, where x and y denote latitude and longitude. The infection ratio q is estimated by the VCAP model (Eq. 11). n denotes the number of target regions.

3) THE COMBINATION OF H-AS AND G-AS (HG-AS)

In this strategy, historical cases and geographical features are comprehensively considered when designing active surveillance. Accordingly, the infection risk of region i at time $t+1$ is calculated as:

$$r_{HGAS}(i, t+1) = \frac{r_{HAS}(i, t+1) \cdot r_{GAS}(i, t+1)}{\sum_{i=1}^{m} r_{HAS}(i, t+1) \cdot r_{GAS}(i, t+1)}$$

In the three strategies, infection risks are first computed and ranked and top-k regions are then selected for

TABLE 1. Precision and recall under different thresholds in 2010 (∗ denotes our method).

		Precision under different thresholds							Recall under different thresholds						
Season	Method	30%	40%	50%	60%	70%	80%	Average	30%	40%	50%	60%	70%	80%	Average
I	ASPII*	1	1	1	1	1	1	**1(1)**	1	1	0.75	1	1	1	**0.96(1)**
	H-AS	1	0.67	0.75	0.60	0.71	0.78	0.75(2)	1	1	0.75	0.60	0.83	0.88	0.84(2)
	G-AS	0	0	0.22	0.36	0.38	0.43	0.23(4)	0	0	0.50	0.80	0.83	0.75	0.48(4)
	HG-AS	0.50	0.67	0.75	0.67	0.71	0.78	0.68(3)	0.50	1	0.75	0.80	0.83	0.88	0.79(3)
II	ASPII*	1	1	1	1	0.83	0.88	**0.95(1)**	0.50	1	1	1	0.83	0.78	**0.85(1)**
	H-AS	0.50	0.33	0.50	0.60	0.71	0.89	0.59(2)	0.50	0.50	0.67	0.75	0.83	0.89	0.69(2)
	G-AS	0	0.14	0.11	0.27	0.31	0.50	0.22(4)	0	0.50	0.33	0.75	0.67	0.78	0.50(4)
	HG-AS	0.50	0.33	0.50	0.60	0.71	0.78	0.57(3)	0.50	0.50	0.67	0.75	0.83	0.78	0.67(3)
III	ASPII*	1	0.67	1	0.80	1	1	**0.91(1)**	1	0.67	1	0.80	0.86	0.89	**0.87(1)**
	H-AS	1	0.67	0.50	0.40	0.71	0.89	0.69(2)	1	0.67	0.50	0.40	0.71	0.89	0.69(2)
	G-AS	0	0.14	0.22	0.36	0.46	0.50	0.28(4)	0	0.33	0.50	0.80	0.86	0.78	0.54(4)
	HG-AS	1	0.67	0.50	0.40	0.71	0.89	0.69(2)	1	0.67	0.50	0.40	0.71	0.89	0.69(2)
IV	ASPII*	0.50	0.67	0.75	0.83	1	0.89	**0.77(1)**	0.50	0.67	0.75	0.83	1	1	**0.79(1)**
	H-AS	0.50	0.33	0.25	0.40	0.57	0.67	0.45(3)	0.50	0.33	0.25	0.33	0.57	0.75	0.46(4)
	G-AS	0	0.14	0.22	0.45	0.46	0.50	0.30(4)	0	0.33	0.50	0.83	0.86	0.88	0.57(2)
	HG-AS	0.50	0.33	0.25	0.40	0.71	0.67	0.48(2)	0.50	0.33	0.25	0.33	0.71	0.75	0.48(3)

TABLE 2. Precision and recall under different thresholds in 2011 (∗ denotes our method).

		Precision under different thresholds							Recall under different thresholds						
Season	Method	30%	40%	50%	60%	70%	80%	Average	30%	40%	50%	60%	70%	80%	Average
I	ASPII*	1	1	1	1	0.75	0.75	**0.92(1)**	0.50	1	1	1	0.75	0.75	**0.83(1)**
	H-AS	0	0	0	0.20	0.29	0.22	0.12(3)	0	0	0	0.33	0.50	0.50	0.22(3)
	G-AS	0.20	0.14	0.11	0.27	0.31	0.29	0.22(2)	0.50	0.50	0.50	1	1	1	0.75(2)
	HG-AS	0	0	0	0.17	0.29	0.22	0.11(4)	0	0	0	0.33	0.50	0.50	0.22(3)
II	ASPII*	1	1	0.75	0.75	1	1	**0.92(1)**	1	1	0.75	0.75	1	1	**0.92(1)**
	H-AS	0.50	0.33	0.50	0.40	0.57	0.67	0.50(3)	0.50	0.33	0.50	0.50	0.80	1	0.61(3)
	G-AS	0	0.29	0.22	0.27	0.31	0.36	0.24(4)	0	0.67	0.50	0.75	0.80	0.83	0.59(4)
	HG-AS	0.50	0.33	0.50	0.40	0.71	0.67	0.52(2)	0.50	0.33	0.50	0.50	1	1	0.64(2)
III	ASPII*	1	1	1	1	1	1	**1(1)**	0.50	1	0.67	1	1	0.80	**0.83(1)**
	H-AS	0.50	0.33	0.25	0.20	0.43	0.56	0.38(2)	0.50	0.50	0.33	0.33	0.75	1	0.57(2)
	G-AS	0	0	0.11	0.09	0.15	0.21	0.10(4)	0	0	0.33	0.33	0.50	0.60	0.29(4)
	HG-AS	0.50	0.33	0.25	0.20	0.29	0.56	0.35(3)	0.50	0.50	0.33	0.33	0.50	1	0.53(3)
IV	ASPII*	1	1	1	1	1	1	**1(1)**	1	1	1	1	0.50	1	0.92(2)
	H-AS	0	0	0	0.17	0.14	0.11	0.07(4)	0	0	0	1	0.50	0.50	0.33(4)
	G-AS	0.20	0.14	0.11	0.09	0.15	0.14	0.14(2)	1	1	1	1	1	1	**1(1)**
	HG-AS	0	0	0.25	0.17	0.14	0.11	0.11(3)	0	0	1	1	0.50	0.50	0.50(3)

high-priority monitoring according to the budget. In order to quantitatively compare these model-free strategies to the model-based ASPII, precision and recall are adopted. These two metrics are often used to validate rank-based algorithms, and are defined as follows:

$$Precision\,(\tau) = \frac{|real\,(\tau) \cap pred\,(\tau)|}{|pred\,(\tau)|}$$

$$Recall\,(\tau) = \frac{|real\,(\tau) \cap pred\,(\tau)|}{|real\,(\tau)|}$$

where $real\,(\tau)$ and $pred\,(\tau)$ denote the sets of selected regions to monitor according to the ground truth and the predictions under a predefined threshold τ, respectively.

The comparison results are given in Tables 1 and 2. Overall, the model based ASPII (marked by ∗) significantly outperforms the model-free strategies for all cases except the last season of 2011, where the recall of ASPII is less than that of G-AS at a threshold of 70%. Note that in this case ASPII is much more precise than G-AS. Compared with 2010, the performance of the three model-free strategies sharply drops when they are applied to the data for 2011. The main reasons are that the surveillance data of 2011 become much sparser than 2010, and the patterns of infection risk in 2011 changed a lot from 2010. It indicates that these model-free strategies cannot adapt to changes in circumstance that can occur, e.g., the rapid drop in cases of infection. Therefore, the model-free strategies are not stable, nor robust.

E. COMPARISON WITH MODEL-BASED STRATEGIES

In this experiment, the effectiveness of ASPII is validated by comparing it with existing model-based predictive models such as linear regression [13] and neural network [16]. The linear regression (LN) considers the relationship between socio-economic factors and the probability of going out as linear, and the infection risk can be calculated as $r_{LN}\ (i, t+1) = \beta_i X_{i,t+1}$, where $X_{i,t+1}$ is a vector of socio-economic factors of region i at time $t+1$, and β_i is the weight vector of socio-economic factors of region i. In the neural network (NN), the number of neurons contained in input, hidden, and output layers are 22, 12, and 1, respectively. The activation function adopted is Sigmoid.

TABLE 3. Precision and recall under different thresholds in 2010 (∗ denotes our method).

		Precision under different thresholds							Recall under different thresholds						
Season	Method	30%	40%	50%	60%	70%	80%	Average	30%	40%	50%	60%	70%	80%	Average
I	ASPII*	1	1	1	1	1	1	**1(1)**	1	1	0.75	1	1	1	**0.96(1)**
	LN	0	0	0.40	0.50	0.50	0.60	0.33(3)	0	0	0.50	0.60	0.67	0.75	0.42(3)
	NN	0.50	0.50	0.40	0.50	0.67	0.67	0.54(2)	0.50	1	0.50	0.60	1	1	0.77(2)
II	ASPII*	1	1	1	1	0.83	0.88	**0.95(1)**	0.50	1	1	1	0.83	0.78	**0.85(1)**
	LN	0.33	0.33	0.60	0.50	0.50	0.60	0.48(2)	0.50	0.50	1	0.75	0.67	0.67	0.68(2)
	NN	0	0.25	0.60	0.43	0.50	0.64	0.40(3)	0	0.50	1	0.75	0.67	0.78	0.62(3)
III	ASPII*	1	0.67	1	0.80	1	1	**0.91(1)**	1	0.67	1	0.80	0.86	0.89	**0.87(1)**
	LN	0	0.33	0.80	0.83	0.75	0.80	0.59(2)	0	0.33	1	1	0.86	0.89	0.68(3)
	NN	0	0.50	0.80	0.71	0.75	0.73	0.58(3)	0	0.67	1	1	0.86	0.89	0.74(2)
IV	ASPII*	0.50	0.67	0.75	0.83	1	0.89	**0.77(1)**	0.50	0.67	0.75	0.83	1	1	**0.79(1)**
	LN	0.33	0.67	0.80	1	0.88	0.70	0.73(2)	0.50	0.67	0.67	0.75	1	0.88	0.75(3)
	NN	0.33	0.50	0.60	1	0.88	0.70	0.67(3)	0.50	0.67	0.75	0.75	1	0.88	0.76(2)

TABLE 4. Precision and recall under different thresholds in 2011 (∗ denotes our method).

		Precision under different thresholds							Recall under different thresholds						
Season	Method	30%	40%	50%	60%	70%	80%	Average	30%	40%	50%	60%	70%	80%	Average
I	ASPII*	1	1	1	1	0.75	0.75	**0.92(1)**	0.50	1	1	1	0.75	0.75	**0.83(1)**
	LN	0	0.33	0.25	0.40	0.33	0.29	0.27(3)	0	0.50	0.50	0.67	0.50	0.50	0.44(3)
	NN	0.50	0.33	0.25	0.17	0.38	0.40	0.34(2)	0.50	0.50	0.50	0.33	0.75	1	0.60(2)
II	ASPII*	1	1	0.75	0.75	1	1	**0.92(1)**	1	1	0.75	0.75	1	1	**0.92(1)**
	LN	0	0.33	0.50	0.40	0.33	0.43	0.33(2)	0	0.33	0.50	0.50	0.40	0.50	0.37(2)
	NN	0	0	0.25	0.33	0.29	0.44	0.22(3)	0	0	0.25	0.50	0.40	0.67	0.30(3)
III	ASPII*	1	1	1	1	1	1	**1(1)**	0.50	1	0.67	1	1	0.80	**0.83(1)**
	LN	0	0	0.25	0.20	0.33	0.43	0.20(2)	0	0	0.33	0.33	0.50	0.60	0.29(2)
	NN	0	0	0.20	0.14	0.12	0.30	0.13(3)	0	0	0.33	0.33	0.25	0.60	0.25(3)
IV	ASPII*	1	1	1	1	1	1	**1(1)**	1	1	1	1	0.50	1	**0.92(2)**
	LN	0	0	0	0	0	0	0(3)	0	0	0	0	0	0	0(3)
	NN	0	0	0	0	0	0.10	0.02(2)	0	0	0	0	0	0.50	0.08(2)

The results are shown in Tables 3 and 4. The ASPII significantly outperforms the LN and NN in terms of both precision and recall rate for all seasons in 2010 and 2011. In 2011, the accuracy of LN and NN was dramatically lower than it was in 2010. In particular, for the fourth season of 2011 LN and NN completely failed to predict the regions with high infection risks. The possible reasons are as follows. 1) The rapidly decreased infected cases in 2011 resulted in the problem of underfitting when training the model, especially for neural network. 2) The LN and NN are model-based algorithms, which only focus on how to fit the model, ignoring the process of human mobility that dominates the diffusion of infection risks.

F. THE STRATEGY OF MODEL SELECTION

Recall the cluster based logistic regression model (Eq. 2), in which Ω denotes the number of clusters. The magnitude of Ω actually reflects the heterogeneity of source locations in terms of both their spatial and socioeconomic features, and influences the prediction of infection risk. The goal of model selection in our active surveillance framework is to find a reasonable value for Ω to maximize prediction accuracy. Model selection is often difficult, as mentioned in Section 2.4, which is why we suggest using cross-validation (CV), a simple but powerful tool, to achieve a reasonable value.

In the case study, surveillance data from 2005 to 2009 are used for model selection. Specifically, a four-fold CV is adopted, in which the surveillance data from 2005 to year $(u-1)$ are used to train the model and the data of year u is used to test the model, where $u \in \{2006, 2007, 2008, 2009\}$. For each fold of the CV, ASPII runs 50 times to obtain an average prediction error. As Fig. 9 shows, the x-axis denotes the number of clusters and y-axis denotes the prediction error of infection cases in terms of RMSE. An overall pattern can be observed: the prediction error first declines rapidly and then tends to become stable when Ω is greater than three.

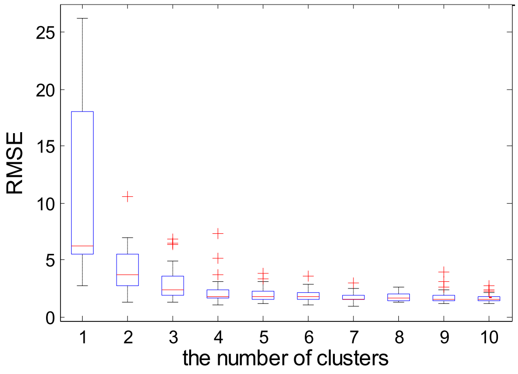


FIGURE 9. The process of model selection to determine a reasonable number of clusters with cross-validation. The bottom and the top of the boxes correspond to the 25*th* and 75*th* percentiles, and the horizontal segment indicates the median. The ends of the whiskers correspond to the 5*th* and 95*th* percentiles. The red crosses are outliers located outside the 90% confidence interval, i.e., events falling below the 5*th* percentile or above the 95*th* percentile.

To avoid the frequent outliers (marked by the crosses above the boxes) that result from the small Ω value (e.g., $\Omega = 4$ or 5) and overfitting caused by complex models (e.g., $\Omega = 9$ or 10), the three values of six, seven and eight are more reasonable. In this case study, we select $\Omega = 6$, and the corresponding clusters obtained from the ASPII algorithm are shown in Fig. 6.

G. SUMMARY OF VALIDATIONS

By adopting the group lasso penalty, sparse representation of the weights of 22 socioeconomic factors was obtained (Fig. 5(a)). Three key features of rural per capita net income (a_{21}), number of live pigs on hand at year's end (a_{15}), and total agricultural machinery power (a_{18}) are readily identified. Based on the estimated weights, a good prediction of infection risk can be made using ASPII (Fig. 5(b)). By comparing the results obtained from plan A, which applies a cluster based logistic regression model, to plan B, which applies a plain logistic regression model, under different thresholds (Fig. 5(c) and (e)) we find that grouping source locations into a reasonable number of clusters is a better way to alleviate the influence of geographical heterogeneity on the prediction of infection risks. This point is further supported by the experiments shown in Fig. 7 and 8. The predictions achieved with ASPII, in which the cluster indicator Z is adaptively estimated from data, are much better than those of ASPII-S, in which Z is a predefined constant. In addition, three model-free active surveillance strategies were tested and compared with the model-based ASPII. We found that if the temporal patterns of infection cases change greatly from previous patterns, the performance of model-free strategies greatly drops, but ASPII still works robustly (Tables 1 and 2). These results indicate that the model-based method captures the essential mechanism of malaria diffusion that is driven by human socioeconomic activity and mobility. It is the mechanism, rather than the temporal pattern of data, that allows us to design an accurate active surveillance framework.

IV. RELATED WORKS AND DISCUSSION

Accurately forecasting the trend of malaria diffusion is necessary for early detection and intervention. Related models and methods for forecasting disease diffusion were developed based on epidemiological models and time series analysis techniques. Smith et al. proposed the VACP model and the EIR (entomological inoculation rate) model to predict malaria spread by analyzing the biological characteristics of mosquito life cycles and sporozoite rates [11]. Noting that weather conditions have a great effect on the spread of vector-borne diseases, Ceccato *et al.* incorporated two additional factors, temperature and rainfall, into the original VCAP model [12]. Laneri *et al.* improved the traditional SEIR model by considering rainfall as a new factor and proposed the vector-SEIRS model [34]. Conventional regression models were applied to malaria spread prediction and analysis. Safi *et al.* explored malaria propagation in Afghanistan from 2001 to 2005 using a linear regression model, in which environmental factors such as vegetation, rainfall, and temperature were taken as inputs [13]. Yang *et al.* fused a linear regression model and Bayesian inference to infer the number of infectious cases at a metapopulation level [14]. Poisson regression [35] and nonlinear regression [36] were also applied for infectious spread prediction.

Human mobility data has also been used to examine the mechanism of infectious disease spread. Colizza *et al.* [43] analyzed infectiousness at the metapopulation level, using data on airline travel flow between 3,100 urban areas in 220 different countries. Andrew *et al.* recorded the mobility of 770,369 Zanzibar inhabitants using anonymous cell phone data and built a transition network of inhabitants [17]. With this network, they aimed to observe the mobility patterns of Plasmodium carriers to infer infection risks of malaria. Similarly, Amy *et al.* [18] studied the spatial distribution of malaria in Kenya by tracking individuals' trajectories. Their work also relied on mobile phone location data. Wearable sensor devices were also used to monitor human mobility behaviors [19], [20]. For example, in hospitals patients' mobility behaviors were monitored with wearable sensor devices, and their positions were recorded every 20 seconds [20]. In the recent program of malaria elimination launched in Namibia, the movement patterns of more than a million people were analyzed by integrating mobile phone and satellite data, and a movement network was constructed to predict new cases [22].

Epidemiologists have adopted machine learning methods to predict infection risks. Huang *et al.* combined the hierarchical Bayesian model and Markov Chain Monte Carlo (MCMC) method to predict the infection risk of malaria in China [15]. Sudheer *et al.* proposed a FFA-SVM algorithm by fusing Firefly algorithm (FFA) and support vector machines (SVM). In their method, FFA was used to select key factors such as rainfall, temperature, and humidity that had the greatest effect on infectious spread [37]. Kiang *et al.* [16] and Gao *et al.* [38] adopted artificial neural networks (ANN) to predict malaria diffusion in Thailand and Yunnan. Simulation models were used to investigate the effectiveness of intervention strategies to curb the diffusion of infectious disease. Longini *et al.* [44] adopted a stochastic influenza model to simulate the outbreak of influenza in a population of 500,000 people in rural Southeast Asia. Eubank *et al.* [45] generated dynamic bipartite graphs by simulating large-scale mobility activities of urban traffic to capture the movements of individuals between specific locations. They found that combining targeted vaccination and early detection is better than resorting to mass vaccination in a population. Fang *et al.* proposed an effective feature selection algorithm named TF-LTR to adaptively select representative query terms for trend surveillance, e.g., trend of epidemics. It is capable of providing effective support for authorities to respond to fast-changing events so as to make appropriate decisions [46].

Spatiotemporal data mining methods have also been adopted recently. By adopting multiple BigData analytics technologies, Zadeh *et al.* analyzed the spatiotemporal patterns of human behavior during the flu season using the real-world data collected from Twitter and Cerner HealthFacts. They found that flu-related traffic on social media is closely related with actual flu outbreaks, and, moreover, clinical flu encounters lag behind online posts [47]. To analyze spatiotemporal patterns of different surveillance regions, Liu *et al.* first modeled multidimensional spatiotemporal data into homogeneous partitions by proposing a piecewise rank-one tensor decomposition algorithm. Then, extract the latent patterns in each partition for comparison and visual summarization [48]. Wu *et al.* designed a visualization system to represent changes of spatiotemporal patterns, by which the evolution of areas of concern to us can be visually analyzed over time. It allows analysts to track the spatiotemporal changes within these areas and understand why such changes occur [49]. Aiming at detecting outliers in surveillance system, DJENOURI et al. proposed a method to construct a distribution probability database using historical spatiotemporal data first. Then, storing inliers contained in the coming data into the database by outlier detection mechanisms, while the existing outliers are excluded from the database [50].

The various studies cited above indicate that climate, environment and human mobility each play an important role in shaping the spread of infectious disease. More importantly, we believe that these works provide a solid foundation that supports our work. They show that infectious spread is simultaneously affected by multiple factors and these factors should be integrated into one comprehensive model. In addition to climate, environment and human mobility, our work attempted to discover other factors to construct a more comprehensive mechanism of spread. With a complete mechanism, we can better understand the complexity of infectious spread, which is driven by the synthetic influence of multiple factors, particularly socioeconomic activity. In practice, this mechanism will enable more powerful active surveillance strategies that will effectively utilize very limited monitoring resources.

V. CONCLUSION

This work proposed a new framework of active surveillance planning, called ASPII. This framework enables examination of the real driving forces of infectious spread from a more comprehensive perspective, covering multiple factors. Specifically, the influence of human mobility on epidemic prevalence was characterized with a radiation model using geographical and demographic data. The influence of socioeconomic activities and their heterogeneity in terms of geographical differences were characterized using a spatial-specific regression model based on socioeconomic data. The influences of physiology and climate change were characterized in an epidemiological model using meteorological and environmental data. Finally, all of these models and heterogeneous data were integrated into a unified mechanism with the proposed spatiotemporal diffusion network, which is a new representation of infectious complexity. This framework enables precise prediction of the spatiotemporal patterns of infectious risk across large regions, and then finds high-priority targets to monitor by accurately considering the available resource constraints.

In our case study, the cross-border diffusion mechanism of malaria between China and Myanmar was systematically examined. The proposed framework allowed us to address the goals that were put forth in the introduction, which were not previously well addressed. Questions related to why so many cases were imported, how they were generated, and what socioeconomic factors dominated the generation process were all answered with our framework. Further, corresponding active surveillance plans were designed for Tengchong county, and their effectiveness has been rigorously validated using real-world data.

Although our empirical study focused on malaria, it should be noted that the challenges and solutions proposed and demonstrated here are general and can readily be extended to other vector-borne diseases such as dengue, cholera, Zika, and avian influenza, in which imported cases are dominant during an outbreak or a stage of an outbreak. Our framework can flexibly address these diseases by replacing the VCAP, the diffusion model designed for malaria, with other epidemiological models to calculate the corresponding infection risk.

REFERENCES

[1] B. Yang, H. Guo, Y. Yang, B. Shi, X. Zhou, and J. Liu, "Modeling and mining spatiotemporal patterns of infection risk from heterogeneous data for active surveillance planning," in *Proc. 28th AAAI Conf. Artif. Intell.*, Canada, Jul. 2014, pp. 493–499.

[2] M. L. Cohen, "Changing patterns of infectious disease," *Nature*, vol. 406, no. 6797, pp. 762–767, 2000.

[3] World Health Organization. (2014). *World Malaria Report 2014*. [Online]. Available: www.who.int/malaria/

[4] D. M. Fleming, P. Chakraverty, and P. Litton, "Combined clinical and virological surveillance of influenza in winters of 1992 and 1993-4," *BMJ*, vol. 311, no. 7000, pp. 290–291, 1995.

[5] W. W. Thompson, L. Comanor, and D. K. Shay, "Epidemiology of seasonal influenza: Use of surveillance data and statistical models to estimate the burden of disease," *J. Infectious Diseases*, vol. 194, pp. S82–S91, Nov. 2006.

[6] J. T. Wu, E. S. K. Ma, C. K. Lee, D. K. W. Chu, P.-L. Ho, A. L. Shen, A. Ho, I. F. N. Hung, S. Riley, and L. M. Ho, "The infection attack rate and severity of 2009 pandemic H1N1 influenza in Hong Kong," *Clin. Infectious Diseases*, vol. 51, no. 10, pp. 1184–1191, 2010.

[7] H. Chen, B. Yang, and J. Liu, "Partially observable reinforcement learning for sustainable active surveillance," in *Proc. Int. Conf. Knowl. Sci., Eng. Manage.* Changchun, China, Aug. 2018, pp. 425–437.

[8] H. J. W. Sturrock, M. S. Hsiang, J. M. Cohen, D. L. Smith, B. Greenhouse, T. Bousema, and R. D. Gosling, "Targeting asymptomatic malaria infections: Active surveillance in control and elimination," *PLoS Med.*, vol. 10, no. 6, 2013, Art. no. 1001467.

[9] Y. Wang, P. Hao, B. Lu, H. Yu, W. Huang, H. Hou, and H. Dai, "Causes of infection after earthquake, China, 2008," *Emerg. Infectious Diseases*, vol. 16, no. 6, pp. 974–975, 2010.

[10] I. K. Kouadio, S. Aljunid, T. Kamigaki, K. Hammad, and H. Oshitani, "Infectious diseases following natural disasters: Prevention and control measures," *Expert Rev. Anti-Infective Therapy*, vol. 10, no. 1, pp. 95–104, 2012.

[11] D. L. Smith and F. E. McKenzie, "Statics and dynamics of malaria infection in Anopheles mosquitoes," *Malaria J.*, vol. 3, no. 1, p. 13, 2004.

[12] P. Ceccato, C. Vancutsem, R. Klaver, J. Rowland, and S. J. Connor, "A vectorial capacity product to monitor changing malaria transmission potential in epidemic regions of Africa," *J. Tropical Med.*, vol. 2012, 2012, Art. no. 595948.

[13] N. Safi, F. Adimi, R. P. Soebiyanto, and R. K. Kiang, "Toward malaria risk prediction in Afghanistan using remote sensing," *Malaria J.*, vol. 9, no. 1, p. 125, 2010.

[14] X. Yang, J. Liu, W. K. Cheung, and X.-N. Zhou, "Inferring metapopulation based disease transmission networks," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, 2014, pp. 385–399.

[15] F. Huang, S. Zhou, S. Zhang, H. Zhang, and W. Li, "Meteorological factors–based spatio-temporal mapping and predicting malaria in Central China," *Amer. J. Tropical Med. Hygiene*, vol. 85, no. 3, pp. 560–567, 2011.

[16] R. Kiang, F. Adimi, V. Soika, J. Nigro, P. Singhasivanon, J. Sirichaisinthop, and S. Looareesuwan, "Meteorological, environmental remote sensing and neural network analysis of the epidemiology of malaria transmission in Thailand," *Geospatial Health*, vol. 1, no. 1, pp. 71–84, 2006.

[17] J. T. Andrew, Y. Qiu, D. L. Smith, O. Sabot, S. A. Abdullah, and M. Bruno, "The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents," *Malaria J.*, vol. 8, p. 287, Dec. 2009.

[18] W. Amy, E. Nathan, J. Andrew, and D. L. Smith, "Quantifying the impact of human mobility on malaria," *Amer. Assoc. Adv. Sci.*, vol. 338, pp. 266–270, Oct. 2012.

[19] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, and J. Pinton, "Dynamics of person-to-person interactions from distributed RFID sensor networks," *PLoS ONE*, vol. 5, Jul. 2010, Art. no. e11596.

[20] L. Isella, M. Romano, A. Barrat, C. Cattuto, V. Colizza, W. Van den Broeck, F. Gesualdo, E. Pandolfi, L. Ravà, C. Rizzo, and A. E. Tozzi, "Close encounters in a pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors," *PLoS ONE*, vol. 6, no. 2, Feb. 2011, Art. no. e17144.

[21] H. Chen, B. Yang, H. Pei, and J. Liu, "Next generation technology for epidemic prevention and control: Data-driven contact tracking," *IEEE Access*, vol. 7, pp. 2633–2642, 2018.

[22] A. J. Tatem, Z. Huang, C. Narib, U. Kumar, D. Kandula, D. K. Pindolia, D. L. Smith, J. M. Cohen, B. Graupe, P. Uusiku, and C. Lourenço, "Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning," *Malaria J.*, vol. 13, no. 1, p. 52, 2014.

[23] X. Wang, B. Yang, J. Huang, H. Chen, X. Gu, Y. Bai, and Z. Du, "IASM: A system for the intelligent active surveillance of malaria," *Comput. Math. Methods Med.*, vol. 2016, 2016, Art. no. 2080937.

[24] World Health Organization. (2013). *World Malaria Report 2013*. [Online]. Available: www.who.int/malaria/

[25] L. Tang, "Progress in malaria control in China," *Chin. Med. J.*, vol. 113, no. 1, pp. 89–92, 2000.

[26] F. Hui, B. Xu, Z. W. Chen, X. Cheng, L. Liang, H.-B. Huang, L. Fang, H. Yang, H. Zhou, and H. Yang, "Spatio-temporal distribution of malaria in Yunnan Province, China," *Amer. J. Tropical Med. Hygiene*, vol. 81, no. 3, pp. 503–509, 2009.

[27] Y. Liu, M. S. Hsiang, H. Zhou, W. Wang, Y. Cao, R. D. Gosling, J. Cao, and Q. Gao, "Malaria in overseas labourers returning to China: An analysis of imported malaria in Jiangsu Province, 2001–2011," *Malaria J.*, vol. 13, no. 1, p. 29, 2014.

[28] R. Yan, S. Zhou, and Z. Xia, "Spatial-temporal characteristics of malaria transmission in China," *J. Pathogen Biol.*, vol. 9, no. 3, pp. 198–219, 2014.

[29] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, no. 7392, pp. 96–100, 2012.

[30] A. P. Masucci, J. Serras, A. Johansson, and M. Batty, "Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 88, no. 2, 2013, Art. no. 022812.

[31] K. P. Paaijmans, A. F. Read, and M. B. Thomas, "Understanding the link between malaria risk and climate," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 33, pp. 13844–13849, 2009.

[32] K. Na-Bangchang and K. Congpuong, "Current malaria status and distribution of drug resistance in East and Southeast Asia with special focus to Thailand," *Tohoku J. Exp. Med.*, vol. 211, no. 2, pp. 99–113, 2007.

[33] Queensland Health. (2012). *Malaria: Queensland Health Guidelines for Public Health Units*. [Online]. Available: www.health.qld.gov.au/cdcg/index/malaria.asp

[34] K. Laneri, A. Bhadra, E. L. Ionides, M. Bouma, R. C. Dhiman, R. S. Yadav, and M. Pascual, "Forcing versus feedback: Epidemic malaria and monsoon rains in northwest India," *PLoS Comput. Biol.*, vol. 6, no. 9, 2010, Art. no. e1000898.

[35] Y. Zhang, K. W. Cheung, and J. Liu, "A unified framework for epidemic prediction based on poisson regression," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 2878–2892, Nov. 2015.

[36] C. Chatterjee and R. R. Sarkar, "Multi-step polynomial regression method to model and forecast malaria incidence," *PLoS ONE*, vol. 4, no. 3, p. e4726, 2009.

[37] C. Sudheer, S. K. Sohani, D. Kumar, A. Malik, B. R. Chahar, A. K. Nema, B. K. Panigrahi, and R. C. Dhiman, "A support vector machine-firefly algorithm based forecasting model to determine malaria transmission," *Neurocomputing*, vol. 129, pp. 279–288, Apr. 2014.

[38] C. Y. Gao, H. Y. Xiong, and D. Yi, "Study on meteorological factors-based neural network model of malaria," *Zhonghua Liuxingbingxue Zazhi*, vol. 24, no. 9, pp. 831–834, 2003.

[39] S. Sang, B. Chen, H. Wu, Z. Yang, B. Di, L. Wang, Q. Liu, X. Tao, and X. Liu, "Dengue is still an imported disease in China: A case study in Guangzhou," *Infection, Genet. Evol.*, vol. 32, pp. 178–190, Jun. 2015.

[40] Venezuela (Bolivarian Republic), South America. *Ma Santé Dans Les Amériques*. [Online]. Available: http://www.paho.org/salud-en-las-americas-2017/?p=2391&lang=fr

[41] J. Lu, J. Wu, X. Zeng, D. Guan, L. Zou, L. Yi, H. Ni, M. Kang, X. Zhang, H. Zhong, X. He, C. Monagin, J. Lin, and C. Ke, "Continuing reassortment leads to the genetic diversity of influenza virus H7N9 in Guangdong, China," *J. Virol.*, vol. 88, no. 15, pp. 8297–8306, 2014.

[42] J. T. F. Lau, S. Griffiths, K. C. Choi, and H. Y. Tsui, "Widespread public misconception in the early phase of the H1N1 influenza epidemic," *J. Infection*, vol. 59, no. 2, pp. 122–127, 2009.

[43] V. Colizza, A. Barrat, M. Barthelemy, A. J. Valleron, and A. Vespignani, "Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions," *PLoS Med.*, vol. 4, no. 1, p. e13, 2007.

[44] I. M. Longini, A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D. A. T. Cummings, and M. E. Halloran, "Containing pandemic influenza at the source," *Science*, vol. 309, no. 5737, pp. 1083–1087, 2005.

[45] S. Eubank, H. Guclu, V. A. Kumar, and M. V. Marathe, "Modelling disease outbreaks in realistic urban social networks," *Nature*, vol. 429, no. 6988, p. 180, 2004.

[46] Z.-H. Fang and C. C. Chen, "A novel trend surveillance system using the information from Web search engines," *Decis. Support Syst.*, vol. 88, pp. 85–97, Aug. 2016.

[47] A. H. Zadeh, H. M. Zolbanin, R. Sharda, and D. Delen, "Social media for nowcasting flu activity: Spatio-temporal big data analysis," *Inf. Syst. Frontiers*, pp. 1–18, Feb. 2019.

[48] D. Liu, P. Xu, and L. Ren, "TPFlow: Progressive partition and multidimensional pattern extraction for large-scale spatio-temporal data analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 1–11, Jan. 2019.

[49] Y. Wu, X. Xie, J. Wang, D. Deng, H. Liang, H. Zhang, W. Chen, and S. Cheng, "Forvizor: Visualizing spatio-temporal team formations in soccer," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 65–75, Jan. 2019.

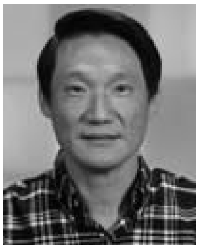
[50] Y. Djenouri, A. Belhadi, J. C.-W. Lin, and A. Cano, "Adapted κ-nearest neighbors for detecting anomalies on spatio-temporal traffic flow," *IEEE Access*, vol. 7, pp. 10015–10027, 2019.

HECHANG CHEN received the M.S. and Ph.D. degrees from the College of Computer Science and Technology, Jilin University, in 2014 and 2018, respectively. He was enrolled in the University of Illinois at Chicago and Hong Kong Baptist University as a visiting Ph.D. student, from 2015 to 2016 and from 2017 to 2018, respectively. His current research interests include heterogeneous data mining and complex network modeling with applications to computational epidemiology.

BO YANG is currently a Professor with the College of Computer Science and Technology, Jilin University. He is also the Director of the Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, China. His current research interests include data mining, complex network analysis, self-organized and self-adaptive multi-agent systems, with applications to knowledge engineering and intelligent health informatics.

JIMING LIU is currently the Chair Professor in computer science and the Dean of the Faculty of Science, Hong Kong Baptist University. He is also the Director of the Centre for Health Informatics and the Co-Founder of the Joint Research Laboratory for Intelligent Disease Surveillance and Control. His current research interests include health informatics, big data analytics, complex systems/networks modeling, and data-intensive epidemiology.

XIAO-NONG ZHOU received the M.S. degree from the Institute of Schistosomiasis Control and Prevention, Jiangsu, China, in 1988, and the Ph.D. degree from the Schistosomiasis Laboratory, University of Copenhagen, Denmark, in 1994. He is currently the Director of the National Institute of Parasitic Diseases (NIPD), China. His research interests include epidemiology, parasitic diseases, medical malacology, and geospatial health.



PHILIP S. YU is currently a Distinguished Professor of computer science with the University of Illinois at Chicago and also holds the Wexler Chair in information technology. Before joining UIC, he was with IBM, where he was the Manager of the Software Tools and Techniques Group, Watson Research Center. He has published more than 1000 papers in refereed journals and conferences. His research interests include data mining, data stream, database, and privacy.