Modeling and Mining Spatiotemporal Patterns of Infection Risk from Heterogeneous Data for Active Surveillance Planning

Bo Yang[1,2], **Hua Guo**[1], **Yi Yang**[1], **Benyun Shi**[3], **Xiaonong Zhou**[4], and **Jiming Liu**[3*]
[1]School of Computer Science and Technology, Jilin University, Changchun, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
[3]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[4]National Institute of Parasitic Diseases, Chinese CDC, Shanghai, China
ybo@jlu.edu.cn; jiming@comp.hkbu.edu.hk

Abstract

Active surveillance is a desirable way to prevent the spread of infectious diseases in that it aims to timely discover individual incidences through an active searching for patients. However, in practice active surveillance is difficult to implement especially when monitoring space is large but available resources are limited. Therefore, it is extremely important for public health authorities to know how to distribute their very sparse resources to high-priority regions so as to maximize the outcomes of active surveillance. In this paper, we raise the problem of active surveillance planning and provide an effective method to address it via modeling and mining spatiotemporal patterns of infection risks from heterogeneous data sources. Taking malaria as an example, we perform an empirical study on real-world data to validate our method and provide our new findings.

Introduction

As compared to passive infectious disease surveillance, i.e., data collection by public health agencies from the patients who come to them, active surveillance is much more desirable and effective to prevent the spread of infectious diseases, as it aims to timely discover individual infection incidences through active searching for patients or house-to-house surveys. However, in practice, active surveillance is difficult to implement especially when monitoring space is large but the resources available for active searching are limited. In those situations, it would be essential for public health authorities to carefully plan their limited resources by determining when and which regions should be searched with a high priority so as to maximize the outcomes of active surveillance. Such a task of resource planning is challenging in that infectious disease diffusion can be potentially caused and affected by many impact factors. The goals of this work is to address this planning task by means of achieving the following objectives:

(1) To develop a computational model of active surveillance planning and then propose a planning method via modeling and mining spatiotemporal patterns of infection risks from heterogeneous data sources, including meteorological, environmental, geographical, transportation, demographic, and socioeconomic, as well as surveillance data (the spatiotemporal distribution of infection incidences annually reported by authority).

(2) Taking malaria as a case study, to conduct an empirical research in Tengchong County, Yunnan Province, China, with the collaboration of epidemiologists and the local CDC and surveillance agencies, to validate the models and algorithms by implementing a prototype and using real-world data, and to uncover the intrinsic causality between socioeconomic factors and imported malaria outbreaks.

We choose malaria as a case study in view of the fact that it is one of the most serious and deadly infectious diseases in the developing countries, and moreover, malaria transmission is complex and challenging to model due to many impact factors. According to the world malaria report issued by World Health Organization (WHO), half of the world's population was at risk of malaria and an estimated 225 million cases led to nearly 0.8 million deaths in 2009 (WHO 2010). In China, the implementation of malaria control measures for more than 30 years has significantly reduced the overall burden in the last century (Tang 2000). However, early in the 21th century, malaria reemerged, representing once again a severe public health threat especially in the remote and poor regions with very limited intervention and medical resources. In 2006 and 2007 alone, a total of more than 0.11 million confirmed and more than 0.13 million suspected cases were reported in China (Zhou et al. 2008). Consequently, an action plan of malaria elimination was launched by the Ministry of Health of China in 2009.

Because of the suitable climate for mosquito habitats, Yunnan has the most serious malaria outbreak in China. From 1999 to 2005, Yunnan was ranked the first for its number of malaria cases in the country (Hui et al. 2009). Moreover, Yunnan shares a long international border with Myanmar, which is one of the most severe epidemic regions in Asia. The impedance of malaria control and resurgent epidemics have been closely associated with the frequent migration of people across the border without natural barriers (Na-Bangchang and Congpuong 2007). The increase of migrants (many of whom are organized by illegal agents) across border regions with a high infectious risk under poor management has been one of the dominant causes of malaria infection in Yunnan nowadays. Taking Tengchong County as

*Corresponding author

an example, the times of crossing Yunnan-Myanmar border from this county alone is over 10,000 per year and from 2005 to 2011 more than 98.57% of total 7,835 reported infections were confirmed as imported cases from Myanmar. Noticeably, the tourism industry of Tengchong is quickly booming recently. In 2012 alone, it attracted more than 5 million tourists. This will present a huge risk causing a nationwide or even worldwide spreading of malaria.

Along with the significant change of social and economic status in Yunnan, the corresponding malaria control strategies have also changed. Besides traditional passive surveillance and vector controls, active surveillance and intervention have also been introduced, particularly in regions at a high risk of infection. The key players implementing active surveillance are the local CDC and surveillance agencies, who should visit villages house by house to enquire whether there is/was a fever case. Furthermore, they should perform surveys fortnightly to safely catch more secondary cases before the commencement of next cycle according to the incubation interval of vivax. For instance, the incubation interval of P.vivax is 12 to 18 days, while for P.falciparum is 9 to 14 days (Queensland-Health 2012). In this respect, active surveillance is extremely cost-expensive and time-consuming and requires massive experienced public health workers. However, the human 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 wide area of 5,845 square kilometers in 2011. Yet, in Tengchong CDC, only a few workers/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 so as to maximize the outcomes of active surveillance.

Since the vast majority (over 98%) of reported cases are imported from Myanmar, it seems that we can simply rank villages according to the number of cross-border migration to plan resources? However, in practice, it is very difficult to get detailed information about how many people in a specific village, a town, or a county, have passed through the border monthly or yearly in that there are over 20 official immigration channels and much more secret and illegal ones provided by snakeheads along the border. Hence, the real-world task, as raised by us having worked directly with the local CDC, can be stated as follows:

Can we find a more feasible and effective method by means of cutting-edge artificial intelligence technologies to estimate the spatiotemporal distribution of malaria risks and then on the basis of it to reasonably allocate human resources for active surveillance?

This task is challenging because malaria transmission can be affected by multiple factors such as biology, environment, and meteorology that directly impinge on the interactions among hosts, vectors, and parasites at varying degrees and scales. Moreover, human mobility driven by socioeconomic factors including income, food and meat production, agricultural population, number of households and so on, will play a particularly important role in Yunnan's malaria distribution in that most of the cases found there are imported from neighboring countries rather than the secondary infections through internal epidemic spreading. According to the surveys (Zinszer et al. 2012; Liu et al. 2012), contemporary spatiotemporal techniques for modeling malaria diffusion, either based on scan statistics clustering (Coleman et al. 2009; Unkel et al. 2012), or by means of biologically modelling entomological inoculation rates and vectorial capacity (Gemperli et al. 2006; Ceccato et al. 2012), or by using a combination of epidemiological, meteorological, and demographic data through a fitting model such as time series analysis (Snow et al. 2005), cannot be applied to address this task because all of them are not designed to answer the questions of why so many cases are imported, how those cases are generated, and what socioeconomic factors dominate the generation process. Recently, some studies model infection diffusions based on cell phone data (Wesolowski et al. 2012; Frias-Martinez, Williamson, and Frias-Martinez 2011; Tatem et al. 2009), web search enquires (Ginsberg et al. 2009) or tweets (Gomide et al. 2011), from which the information of times, locations, even human mobilities can be extracted. However, such techniques cannot be used to solve our problem because mobile phone and Internet are not popular in most poor and remote regions.

In what follows, we first raise the problem of active surveillance planning, and then propose our solution to address it on the basis of predicting the spatiotemporal distribution of infection risks by sufficiently considering the influential factors discussed above. Finally, we perform a case study on real-world data to validate our solution.

Active surveillance planning method

Problem definition

The main objective of active surveillance planning is to answer the question about where and when to search infected cases so as to maximize the outcomes of available surveillance resources. Formally, we define such a planning task as a constrained optimization problem, as follows:

Out of all regions of interest (ROI), to select the minimum number of targets that are prioritized to scan, which would sufficiently guarantee to cover a large percentage (or a threshold predefined according to the limitation of available resources) of all potential incidences within a period of time in the future.

We propose an infection-risk-based planning method to address the above problem. Its main steps are as follows.

Step 1: for each region, predict its infection risk within a specified time window, say a quarter for a short-term planning, or more challenging, a year for a long-term planning;

Step 2: for each region, estimate the real number of infection incidences within the time window with the product of population and infection risk;

Step 3: rank all regions into a descending order according to their respective incidence numbers;

Step 4: select a minimum k so that the overall infection ratio of the top-k regions is above a predefined threshold;

Step 5: output the top-k regions as the high-priority targets to be searched for the given time window.

The first step is the foundation of the above method. We will elaborate it by presenting a model to represent the spatiotemporal patterns of infection risks as well as an algorithm to mine such patterns from heterogeneous data.

Modeling spatiotemporal pattern of infection risks

Human mobility has an important influence on the generation and distribution of infection incidences, which should be sufficiently considered during modeling. We characterize human mobility in terms of a transition matrix, in which entries depict the likelihoods of human going from one region to another. One can directly construct such a matrix based on the data recording the frequencies of human mobility between sources and targets. However, it is very difficult to get such detailed and sufficient information in practice especially for the migrations across international borders due to the reasons mentioned before. In view of this, our model will regard the transition matrix as a hidden variable to be estimated, which is collectively regulated by socioeconomic, geographical and transportation factors.

We now introduce our modeling method by proposing a new concept of *heterogeneous diffusion network* containing multiple types of nodes and links, as shown in Fig. 1. Without loss of generality, we take our empirical study, i.e. Tengchong County, as an example to illustrate the elements of the network for readily understanding.

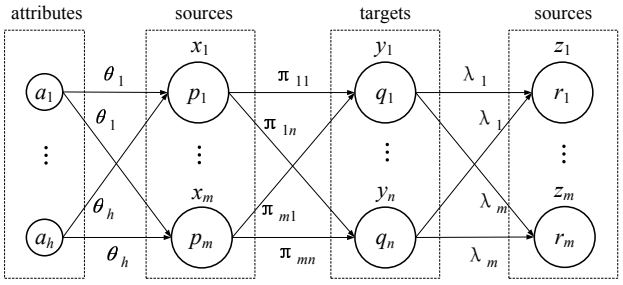

Figure 1: A heterogeneous diffusion network

In the network, node a_i denotes one of h socioeconomic attributes driving villagers to go to Myanmar to work. Node x_i denotes one of m locations (villages or towns) of Tengchong, and node weight p_i depicts the likelihood of its villagers going out to work. For example, according to our investigations from surveys, in most of the imported cases, their main purposes of going to Myanmar are for doing work such as logging and mining. Hence, p_i will be driven by a nonlinear combination of h socioeconomic attributes in terms of a set of weights $\theta = (\theta_1, \cdots, \theta_h)$, 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 villager of x_i going to y_j, one of n locations in Myanmar. Node weight q_i depicts the risk of malaria infection within y_i. Link weight λ_i depicts the temporal probability distribution of a villager going back to his/her source location z_i from Myanmar after working at different times. More specifically, λ_i is a t-dimension vector $(\lambda_{i1}, \cdots, \lambda_{it})$, where $\lambda_{it} (1 \leq t \leq T)$ denotes the probability of going back to source location z_i at time interval t. The number of time intervals is set according to the time scale of planning as well as the granularity of available surveillance data. For examples, one can set $T = 4$ for seasonal planning or set $T = 12$ for monthly planning. Node z_i denotes the same 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 the variations of input data monthly or annually. That is why we say vector r (actually the whole heterogeneous diffusion network) represents not only spatial but also temporal patterns of infection risks.

Noticeably, besides the expected spatiotemporal pattern of infection risks provided by vector r, from this network, 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 worked out, simultaneously.

Mining spatiotemporal pattern of infection risks

The task of mining the spatiotemporal pattern of infection risks is to determine or infer all parameters of the heterogeneous diffusion network from available data.

The task is not trivial in that most of parameters cannot be directly figured out due to lack of data. Before presenting detailed formulations, we briefly introduce our basic strategies of parameter estimations. Vector p can be calculated by means of a regression model taking the input of socioeconomic data and weights θ. We suggest to use the logistic regression to represent the causality between attributes a and p, as in our preliminary studies we have empirically found quite many of socioeconomic attributes are in an approximate log-linear scale to the actual risks of malaria infection. Once villagers make their decisions to work, where will they go? In real data, only a very small portion of infected cases reported their working places in Myanmar, and thus it is not reliable to statistically infer vector π from such insufficient information. Instead, a more reasonable π can be estimated by an economical job-finding model (Simini et al. 2012; Masucci et al. 2013), such as the radiation model taking input of the demographical, geographical, and transportation data of both source node x and target node y. Without the surveillance data of Myanmar, one cannot directly count the infection ratios in different locations. We turn to biological and epidemiological models, such as vector capacity model (Ceccato et al. 2012), to estimate vector q by taking input of environmental and meteorological data of Myanmar including temperature, rainfall and humidity. In practice, we have not data explicitly recording the distribution of λ_i, i.e. for each source location, how many people will return to it from Myanmar during a certain period. Alternatively, we estimate the distribution based on surveillance data by assuming that the ratio of villagers going back to their source location z_i from Myanmar 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 big enough to cover the incubation of infectious disease. For example, in our case study, the interval is set to 3 months (or $T = 4$) to safely cover the longest incubation interval of vivax. Thereafter, vector r can be represented as a multiplication of p, π, q and λ. Finally, we infer weights θ from surveillance data by means of a statistical inference method

such as maximum likelihood estimation (MLE).

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

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

where $X_i = (x_{i1}, \cdots, x_{ih})^{\mathrm{T}}$ 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 the same impact on different locations. However, this may not be reasonable due to the variations of socioeconomic levels of distinct locations. To characterize such variations, we relax this constraint by introducing a cluster based logistic regression model. In this model, source locations are clustered so that within the same cluster locations share a common θ to regulate respective p_i. Otherwise, θ will be different. This more flexible model enable us to provide a more detailed causal analysis, i.e. which socioeconomic factors will dominate the malaria infection of a specific location, and then plan a more targeted intervention for the location, accordingly.

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

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

where Z_i corresponds to the i-th row 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, they use the same weights $\theta_i = (\theta_{i1,} \cdots, \theta_{ih})$ 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 (Simini et al. 2012) to estimate π, as follows:

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

where pop_i and pop_j are the populations of source location x_i and target location y_j, respectively. Let r_{ij} be the distance between x_i and y_j, s_{ij} is the total population in the circle of radius r_{ij} centered as x_i by excluding the target population.

The infection risk of malaria is determined by the ability of the mosquitoes to transmit Plasmodium, generally referred to as vectorial capacity, which can be formally expressed by the following VCAP model (Ceccato et al. 2012):

$$V = \frac{-(\mu \alpha^2)\rho^{\tau}}{ln(\rho)} \tag{4}$$

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 of a mosquito surviving through one whole day, and τ is the extrinsic incubation period of malaria parasites or the time taken for completion of the extrinsic cycle.

The parameters of the VCAP can be determined by temperature and rainfall (Paaijmans, Read, and Thomas 2009), two of major environmental and meteorological factors triggering malaria epidemics in warm semiarid and altitude areas. More 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} denotes the rainfall and the lowest temperature of an area during a time interval. Furthermore, we can estimate the infection risk of location y according to the following model (Smith and McKenzie 2004):

$$q = \frac{\beta V - \sigma}{\beta V + \sigma \frac{\alpha}{\eta}} \tag{5}$$

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

Note that, all parameters in the above models, expect for 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 according to their studies on the malaria infection in Myanmar.

Based on the above analysis, the infection risk of source location z_i at time interval t in a certain year u is as follows:

$$r_{it}^{(u)} = p_i^{(u)} (\sum_{j=1}^{n} \pi_{ij}^{(u)} q_{jt}^{(u)}) \lambda_{it}^{(u)} \tag{6}$$

The total surveillance data of Y years can be represented as a cube tensor denoted as $C = [c_{uit}]_{Y \times m \times T}$, where c_{uit} denotes the number of incidences reported at location x_i during the time interval t of year u. In terms of two parameters θ and Z, the likelihood of surveillance data C is as follows:

$$L(C; \theta, Z) = \Pi_{u=1}^{Y} \Pi_{i=1}^{m} \Pi_{t=1}^{T} \binom{pop_i^{(y)}}{c_{uit}} (r_{it}^{(u)})^{c_{uit}} (1 - r_{it}^{(u)})^{pop_i^{(y)} - c_{uit}} \tag{7}$$

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

$$\begin{aligned} &\max \log L(C; \theta, Z) \\ &s.t. \quad \forall i \ \sum_{j=1}^{\Omega} z_{ij} = 1, \quad \forall i, j \ \ z_{ij} \geq 0 \end{aligned} \tag{8}$$

We solved the problem by iteratively performing the gradient descents on θ and Z until convergence, as follows:

$$\begin{cases} \theta = \theta - \delta \cdot \frac{\partial J}{\partial \theta} \\ Z = Z - \delta \cdot \frac{\partial J}{\partial Z} \end{cases} \tag{9}$$

where δ is a given learning rate and we have:

$$\begin{aligned} J = &- \log L(C; \theta, Z) \\ &+ \phi_1 \sum_{i=1}^{m} (1 - \sum_{j=1}^{\Omega} z_{ij})^2 - \frac{1}{\phi_2} \sum_{i=1}^{m} \sum_{j=1}^{\Omega} \log z_{ij} \end{aligned} \tag{10}$$

Note that, a penalty function and a barrier function are introduced corresponding to two constraints of Eq.8, respectively. ϕ_1 and ϕ_2 are penalty weights and ϕ_2 should be set to a very small number close to 0. Moreover, we have:

$$\begin{cases} \frac{\partial J}{\partial \theta} = & -\sum_{u=1}^{Y}\sum_{i=1}^{m}\sum_{t=1}^{T} f(Z_i^{\mathrm{T}} X_i^{\mathrm{T}}) \\ \frac{\partial J}{\partial z_{ij}} = & -\sum_{u=1}^{Y}\sum_{t=1}^{T} f(X_i^{\mathrm{T}}\theta_j^{\mathrm{T}}) \\ & -2\phi_1(1-\sum_{j=1}^{\Omega} z_{ij}) - \frac{1}{\phi_2 z_{ij}} \end{cases} \quad (11)$$

where $f(S) = \frac{c_{uit}\cdot S\cdot e^{Z_i\theta X_i}}{1+e^{Z_i\theta X_i}} - \frac{(pop_i^{(u)}-c_{uit})\cdot S\cdot e^{Z_i\theta X_i}}{(1+e^{Z_i\theta X_i}-\omega)(1+e^{Z_i\theta X_i})}\cdot\omega$ and $\omega = \lambda_{it}^{(u)}\sum_{j=1}^{n}\pi_{ij}^{(u)}q_{jt}^{(u)}$.

Finally, we adopt the cross-validation method to address the model selection issue: how to determine a reasonable Ω, i.e. the number of clusters of source locations. In this method, we first bipartition whole training data into two. One is used to estimate θ and Z for a given $\Omega(1 \le \Omega \le m)$. Another is used to test the performance of such estimations in terms of the accuracy of infection risks predictions based on Eq.6. From all m candidates, we select the one with the best performance as the real number of clusters. With the selected Ω and corresponding θ and Z, one can predict infection risks and accordingly plan active surveillance for a given time interval in the future.

An empirical study in Tengchong County

Data collection and description

Collecting sufficient and accurate data from multiple heterogeneous sources is filled with challenges. The data sources explored by us to mine the heterogeneous diffusion network are summarized here.

For Tengchong County, the surveillance data on monthly malaria cases on the village level for seven years (2005-2011) were obtained from the annual reports of National Institute of Parasitic Disease, Chinese CDC. The data contains total 7,835 incidences distributed in 221 villages, as illustrated in Fig. 2(c). The annual demographic data on the village level were obtained from Chinese Natural Resources Database. The socioeconomic data on the town level were obtained from the annual reports issued by Tengchong government, which contains total 22 socioeconomic attributes, denoted by $a_1, \cdots, a_{22}$, respectively. Note that, the surveillance data is quite sparse on the village level particularly for the recent years of 2010 and 2011, as illustrated by Figs. 2(c)(d). From 2005 to 2011, the average infection of each village per year is about 5. As a result, big bias of parameter estimation will be introduced if the model (Eq.7) directly fits surveillance data on the village level. Additionally, the annual socioeconomic data on the village level is not available. Due to these reasons, in our empirical study we take 18 towns (consisting of 221 villages) as source locations, denoted by $x_1, \cdots, x_{18}$ as shown in Fig.2(b).

Comparatively speaking, it is difficult to directly obtain Myanmar data since the official data opened by the country is still limited. From the supplement information of surveillance data, we determined total 72 places in Myanmar where the imported cases of Tengchong have ever been to before. We found that most of these places are distributed in 10

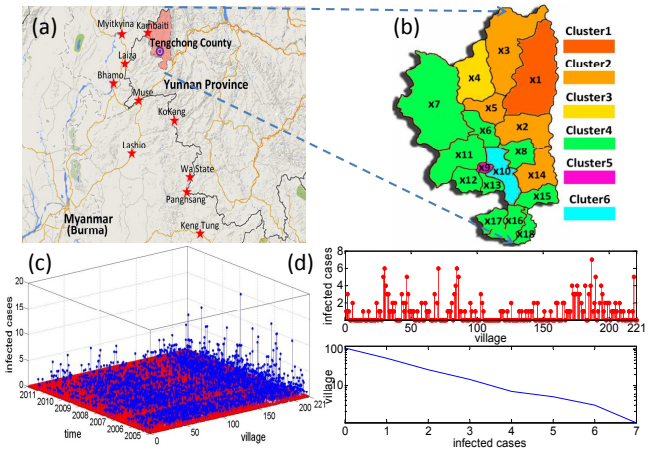

Figure 2: The spatiotemporal distribution of surveillance data. (a) The map of Yunnan-Myanmer border. (b) The map of Tengchong County consisting of 18 towns. (c) The landscape of total cases in 221 villages during 7 years, in which blue and red denote nonzero and zero entries, respectively. (d) Up: The accumulated cases of respective village in 2011. Bottom: The x is the number of cases and the y is the number of villages having a certain number of cases in 2011. This approximate power law distribution indicates that it is possible for us to find the majority of cases by just targeting a small number of villages.

cities or towns of Myanmar near Yunnan-Myanmar international border, as marked by asterisks in Fig.2(a), which are taken as the target locations denoted by $y_1, \cdots, y_{10}$, respectively. We get the temperature and rainfall data of these targets by integrating three sources including IRI/LDEO Climate Data Library, TRMM (Tropical Rainfall Measuring Mission) and MODIS (MODerate-resolution Imaging Spectroradiometer). The last two datasets are provided by NASA, in which useful data can be extracted by a remote sense image processing software ENVI (ENvironment for Visualizing Images). 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. Furthermore, we obtained the geographical and transportation data about the source and target locations from Google Earth.

Validations and analysis

We use the surveillance data of 2005-2009 for learning and those of 2010-2011 for testing. Specifically, we will first analyze the socioeconomic factors dominating imported incidences based on the estimated θ, and then test the accuracy of infection risks prediction in terms of vector r and the effectiveness of active surveillance planning under different coverage thresholds.

The results are presented in Fig.3. According to the estimated clustering indicator Z, the 18 towns of Tengchong are clustered into 6 groups, as shown in Fig.2(b), in which different colors represent different clusters. Accordingly, six weight vectors, $\theta_1, \cdots, \theta_6$, are obtained. As an example,

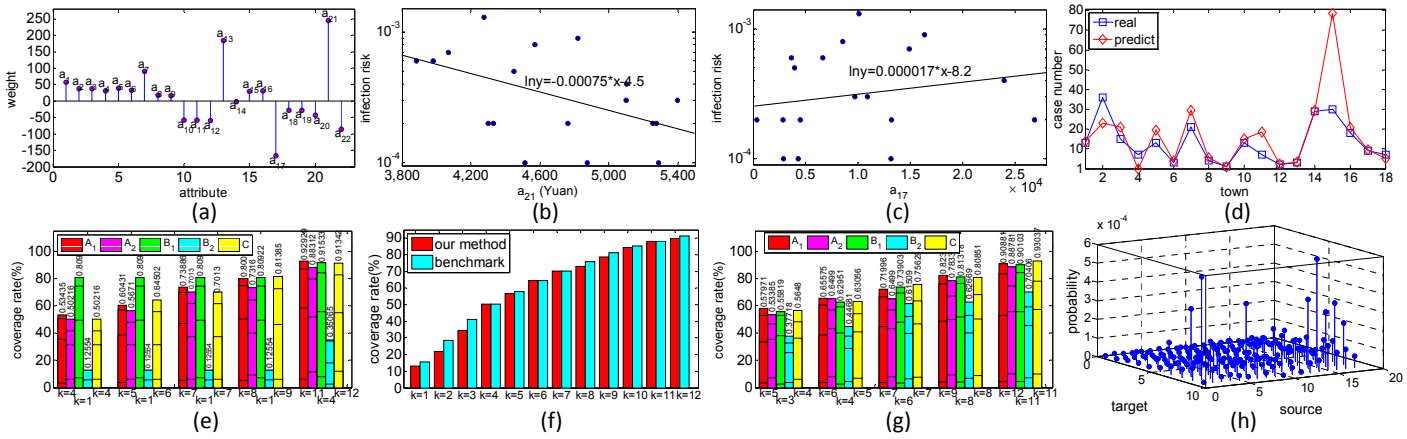

Figure 3: The experimental results

Fig. 3(a) plots out θ_4, the weights of 22 socioeconomic attributes for the largest cluster consisting of 10 towns, in which there are 13 positive weights and 9 negative weights. From it one can find three dominant factors. Two of them are "rural per capita net income (a_{21})" and "food production per capita (a_{13})" with maximum positive weighs and one of them is " amount of heavy livestock on hand at year's end (a_{17})" with a maximum negative weight. Note that, according to Eq.2, the attributes with positive or negative weights will suppress or promote the probability of out-going for work, respectively. It implies that one can cut down the infection ratios of the 10 towns by increasing their levels of income and food production while decreasing their amounts of heavy livestock at the end of year. Moreover, we found that these attributes are in an approximate log-linear scale to actual risks of malaria infection, as shown in Figs.3(b) and (c), in which the exponent of fitting function is either negative for the attribute with a positive weight or positive for the attribute with a negative weight, respectively.

Figs.3(d) shows the distributions of infected cases of 18 towns in 2011 for testing the accuracy of infection risks prediction. The blue line gives actual numbers and the red line gives the prediction of our method. One can observe that the prediction fits the truth quite well expect for town No.15.

Fig.3(e) shows the plans of active surveillance (i.e. the top-k towns selected to search) for 2011 under five coverage thresholds (0.5, 0.6, 0.7, 0.8, 0.9 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. Plan B is given by the proposed method integrating a basic logistic model (Eq.1). 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. Note that, each bar of coverage rate consists of 4 portions, denoting the contributions of four quarters of a year from bottom to top, respectively. One can see, 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 under all five cases. We further compare the top-k towns selected by our method against the benchmark in terms of the coverage rates of real cases, as shown in Fig.3(f). One can see the coverage rates achieved by our method are still very close to the benchmark from top 1 to top 12. Similarly, Fig.3(g) provides the plans of A, B and C for 2010 under five coverage thresholds.

The hidden distribution of cross-border migration estimated by p and π can be served as a reference for local government to prevent illegal stowaways. As an example, Fig.3(h) shows the distribution in 2010.

Conclusion

In summary, the contribution of this work is four-fold: (a) raise the problem of active surveillance planning with broad applications in disease control, especially for poor regions; (b) propose a framework to address the important real-world problem, in which sparse resources to be planned can be human resources (as studied here) or others such as vaccine; (c) propose a novel representation of disease propagation (heterogeneous diffusion network) to model and infer spatiotemporal patterns of infection risks based on multiple and heterogeneous data; (d) for the first time, comprehensively explore the causes of epidemiological situations from a socioeconomic perspective by means of artificial intelligence methodologies such as learning and mining.

Although our empirical study focuses on malaria, we note that the problems and ideas proposed and demonstrated here are general and can readily be extended to address other vector-borne diseases, such as dengue, cholera and avian influenza, if imported cases are dominant during an outbreak or a stage of an outbreak. For instances, most dengue cases in Guangzhou are imported from southeast Asia through traveling; all cholera cases of Caracas reported in 2011 are imported from Dominica; in the early stage of H1N1 in 2009 and H7N9 in 2013, reported cases of many cities in China are imported. Our method can flexibly address such diseases by replacing the VCAP model with others, which calculate their infection risks in terms of corresponding factors.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grants 61133011, 61373053, 61300146 and 81273192, the Program for New Century Excellent Talents in University under grant NCET-11-0204, and Hong Kong Research Grants Council under grant RGC/HKBU211212.

References

Ceccato, P.; Vancutsem, C.; Klaver, R.; Rowland, J.; and Connor, S. J. 2012. A vectorial capacity product to monitor changing malaria transmission potential in epidemic regions of africa. *Journal of Tropical Medicine* 2012(595948):1–6.

Coleman, M.; Coleman, M.; Mabuza, A. M.; Kok, G.; Coetzee, M.; and Durrheim, D. N. 2009. Using the satscan method to detect local malaria clusters for guiding malaria control programmes. *Malaria Journal* 8(1):68–73.

Frias-Martinez, E.; Williamson, G.; and Frias-Martinez, V. 2011. An agent-based model of epidemic spread using human mobility and social network information. In *Proceedings of IEEE 3rd International Conference on Social Computing*, 57–64. IEEE.

Gemperli, A.; Vounatsou, P.; Sogoba, N.; and Smith, T. 2006. Malaria mapping using transmission models: application to survey data from mali. *American Journal of Epidemiology* 163(3):289–297.

Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski, M. S.; and Brilliant, L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014.

Gomide, J.; Veloso, A.; Meira Jr, W.; Almeida, V.; Benevenuto, F.; Ferraz, F.; and Teixeira, M. 2011. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd International Web Science Conference*, number 3, 1–8. ACM.

Hui, F.; Xu, B.; Chen, Z. W.; Cheng, X.; Liang, L.; Huang, H.; Fang, L.; Yang, H.; Zhou, H.; Yang, H.; et al. 2009. Spatio-temporal distribution of malaria in yunnan province, china. *The American Journal of Tropical Medicine and Hygiene* 81(3):503–509.

Liu, J.; Yang, B.; Cheung, W. K.; and Yang, G. 2012. Malaria transmission modelling: a network perspective. *Infectious Diseases of Poverty* 1(11):1–8.

Masucci, A. P.; Serras, J.; Johansson, A.; and Batty, M. 2013. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E* 88(2):022812–022819.

Na-Bangchang, K., and Congpuong, K. 2007. Current malaria status and distribution of drug resistance in east and southeast asia with special focus to thailand. *The Tohoku Journal of Experimental Medicine* 211(2):99–113.

Paaijmans, K. P.; Read, A. F.; and Thomas, M. B. 2009. Understanding the link between malaria risk and climate. *Proceedings of the National Academy of Sciences* 106(33):13844–13849.

Queensland-Health. 2012. Malaria: Queensland health guidelines for public health units. available at: www.health.qld.gov.au/cdcg/index/malaria.asp.

Simini, F.; González, M. C.; Maritan, A.; and Barabási, A.-L. 2012. A universal model for mobility and migration patterns. *Nature* 484(7392):96–100.

Smith, D. L., and McKenzie, F. E. 2004. Statics and dynamics of malaria infection in anopheles mosquitoes. *Malaria Journal* 3(1):13–26.

Snow, R. W.; Guerra, C. A.; Noor, A. M.; Myint, H. Y.; and Hay, S. I. 2005. The global distribution of clinical episodes of plasmodium falciparum malaria. *Nature* 434(7030):214–217.

Tang, L. 2000. Progress in malaria control in china. *Chinese Medical Journal* 113(1):89–92.

Tatem, A. J.; Qiu, Y.; Smith, D. L.; Sabot, O.; Ali, A. S.; and Moonen, B. 2009. The use of mobile phone data for the estimation of the travel patterns and imported plasmodium falciparum rates among zanzibar residents. *Malaria Journal* 8(1):287–298.

Unkel, S.; Farrington, C.; Garthwaite, P. H.; Robertson, C.; and Andrews, N. 2012. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175(1):49–82.

Wesolowski, A.; Eagle, N.; Tatem, A. J.; Smith, D. L.; Noor, A. M.; Snow, R. W.; and Buckee, C. O. 2012. Quantifying the impact of human mobility on malaria. *Science* 338(6104):267–270.

WHO. 2010. World malaria report 2010. available at: www.who.int/malaria/world_malaria_report_2010/en/.

Zhou, S.; Wang, Y.; Fang, W.; and Tang, L. 2008. Malaria situation in the people's republic of china in 2007. *Chinese Journal of Parasitology & Parasitic Diseases* 26(6):401–403.

Zinszer, K.; Verma, A. D.; Charland, K.; Brewer, T. F.; Brownstein, J. S.; Sun, Z.; and Buckeridge, D. L. 2012. A scoping review of malaria forecasting: past work and future directions. *British Medical Journal Open* 2(6):e001992.