A Unified Framework for Epidemic Prediction based on Poisson Regression

Yu Zhang, *Member, IEEE*, William K. Cheung, *Member, IEEE*, and Jiming Liu, *Fellow, IEEE*

Abstract—Epidemic prediction is an important problem in epidemic control. Poisson regression methods are often adopted in existing works, mostly with only the (intra-)regional environmental factors considered. As the diffusion of epidemics is affected by not only the intra-regional factors but also inter-regional and external ones, a unified framework based on Poisson regression with the three types of factors incorporated is proposed for the prediction. Specifically, we propose a Poisson-regression-based model first with the intra-regional and inter-regional factors included. The intra-regional factor in a particular time interval is represented by one feature vector with the regionally environmental and social factors considered. The inter-regional factor is modeled by a diffusion matrix which describes the possibilities that the epidemics can spread from one region to another, which in turn accounts for the propagating effects of the infected cases. To learn the structure of the diffusion matrix, we propose two approaches—utilizing some a priori knowledge (e.g., transportation network) and estimating it from scratch via a sparse structure assumption. The resulting optimization problem of the maximum a posterior solution is a convex one and can be efficiently solved by the alternating direction method of multipliers (ADMM). In addition, we incorporate also the external factor, i.e., the imported cases. With one fact that the distribution of the number of infected cases over a year is (approximately) unimodal for most epidemics and one assumption that the importing rate has a small variance over the year, we can approximate the effect of the external factor with a parametric function (e.g., a quadratic function) over time. The resulting optimization problem is still convex and can be also solved by the ADMM algorithm. Empirical evaluations are conducted based on a real data set which records the 16-days-reported cases in the Yunnan province of China for seven years, from 2005 to 2011. The experimental results demonstrate the effectiveness of our proposed models.

Index Terms—Epidemic prediction, Poisson regression, ADMM algorithm

1 Introduction

Each outbreak of epidemics may bring much cost in both lives and dollars, e.g., Severe Acute Respiratory Syndrome (SARS) [16] and Influenza A virus subtype H1N1 [5]. So epidemic control attracts much attention of the government in each country and the World Health Organization. Among many problems in epidemics control, the epidemic prediction problem, which forecasts the prevalence of the epidemic in certain regions, is one of the important ones [20]. Accurate epidemic prediction is useful in saving lives and reducing unnecessary damage caused by the epidemic. For example, based on the prediction results, governments in hard-hit regions can take immediate actions, e.g., sterilizing in public places and giving warning to public, to prevent the outbreak or at least to reduce the consequence caused by the outbreak. Meanwhile, people can protect themselves from getting infection by, for example, avoiding to go to public areas and taking better care of personal hygiene. The epidemic prediction results can take some forms such as outbreak probabilities, the possible number of infected cases, and so on. The latter is adopted in this paper.

Pioneering studies in epidemic modeling and control are mostly based on meta-populations and using the model-based simulation approach [3]. Also, to achieve the early warning objective, Poisson-regression-based methods have been proposed for epidemic prediction based on the past reported cases where environmental factors [1], [17], [20], [21] are commonly considered. In order to make accurate prediction for epidemics, one key issue is to identify the factors affecting the epidemic propagation. Here we classify the factors into three types: *intra-regional factor*, *inter-regional factor*, and *external factor*. The intra-regional factor includes all region-specific elements that affect the propagation of the epidemics within one region, e.g., environmental [1], [17], [20], [21] and social elements [2]. The environmental elements can consist of temperature, rainfall, humidity, elevation and so on. Since the virus or bacterial triggering a disease normally can only survive under certain environmental conditions, the environmental elements become prerequisites for epidemic diffusion. The social elements can be, for example, average income and population size. A poor region typically has a bad medical condition and is not capable of controlling infectious diseases well. Moreover, if there are many people living in a region, then it is obvious that the chance that this region has infected cases is higher than another region with a smaller population size. The inter-regional factor refers to the elements that facilitate the epidemics to spread to different regions, for example, human mobility [9], [23]. The rapid development of traffic systems leads to easy and fast traveling between regions,

- *Y. Zhang is with the Department of Computer Science, Hong Kong Baptist University and the Institute of Research and Continuing Education, Hong Kong Baptist University (Shenzhen). E-mail: yuzhang@comp.hkbu.edu.hk.*
- *W. Cheung and J. Liu are with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China. E-mail: {william, jiming}@comp.hkbu.edu.hk.*

Manuscript received 13 Feb. 2014; revised 16 Apr. 2015; accepted 4 May 2015. Date of publication 21 May 2015; date of current version 2 Oct. 2015. Recommended for acceptance by H. Xiong. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TKDE.2015.2436918

making the propagation of epidemics through the traveling of disease-carriers highly possible. The SARS is a representative example which originated in Guangdong province of China and then propagated to Southeast Asia and thus the whole world via disease-carriers who were travelling around. The external factor denotes the elements which are due to the regions outside the scope of investigation. One example is the imported cases. Usually we can only make observations in some targeted regions and the behavior of the outside regions can only be treated as imported cases. Moreover, there are some studies to apply the biological epidemic models to understand the propagation of computer viruses, e.g., [24].

Existing works on epidemic prediction mainly focus on the investigation of the intra-regional factor. For example, Woodruff et al. [21], Teklehaimanot et al. [20], Abellana et al. [1], and Nkurunziza et al. [17], respectively, studied the effect of weather on the diffusion process of Ross River virus epidemics and malaria. Denoeud et al. [10] used Poisson regression [8] to predict the pneumonia and influenza mortality based on morbidity data. Achcar et al. [2] employed Poisson regression to investigate the effect of human development index and the size of inhabitants, population and the number of doctors, which corresponds to the environmental and social elements of the intra-regional factor, on the disease diffusion. Sharmin and Rayhan [18] used the negative binomial model, which is an extension of the Poisson regression model, to predict monthly infected cases for measles. Recently, the investigation on the effect of human mobility to epidemics spreading, which belongs to the inter-regional factor, has attracted much attention, e.g., [9], [23]. However, there is no work to investigate how to use the inter-regional factor for epidemic prediction.

To the best of our knowledge, there is also no work to investigate the combination of the intra-regional and inter-regional factors. In addition, the study of the effect of the external factor on the epidemic prediction is lacking in the literature. In this paper, we aim to fill this gap and study how to integrate the three factors within one framework based on the Poisson regression model for epidemic prediction. First we extend Poisson regression to combine the intra-regional and inter-regional factors. Specifically, the intra-regional factor is represented by a feature vector that encodes the environmental and social features. The inter-regional factor is modeled by a diffusion matrix, which describes the possibilities that the epidemics spread from one region to another, to determine the propagating effect of the infected cases at all regions. Then by summing up the effects of the intra-regional and inter-regional factors, we can define the parameters of the Poisson likelihood function corresponding to a region in a certain time interval. Since the structure of the diffusion matrix is mostly unknown, we can either utilize some a priori knowledge from other auxiliary networks (e.g., transportation network) or estimate it from scratch by assuming that the diffusion network is sparse. The optimization problem of the resulting model is found to be convex. In order to deal with the high-dimensional model parameters and also the constraints for the model parameters, we adopt the alternating direction method of multipliers (ADMM) [6].

Other than considering only the intra-regional and inter-regional factors, we add also the external factor by utilizing one fact that the distribution of the number of infected cases in different time intervals over a year is (approximately) unimodal for most epidemics. Based on this fact and one additional assumption that the importing rate has small variance over a year, we can approximate the effect of the external factor with a parametric function (e.g., a quadratic function) over time. The resulting optimization problem is still convex and hence the ADMM algorithm can be applied. The experiments are conducted on one malaria dataset which records the 16-days-reported number of infected cases in 62 counties of the Yunnan province in China. Experimental results on the epidemic prediction demonstrate the effectiveness of our proposed models.

The remainder of this paper is organized as follows. In Section 2, we give an overview on Poisson regression. The first model that combines the intra-regional and inter-regional factors is presented in Section 3 which will set the stage for the introduction of our second model that considers all the three factors in Section 4. In Section 5, we report the experimental results and Section 6 concludes the paper.

2 Poisson Regression with ADMM

Poisson regression and its extension (e.g., negative binomial regression) are natural choices for epidemic prediction, e.g., [2], [10], [18], since the number of infected cases is essentially an integer. In this section, we review the basic Poisson regression model. Moreover, we will discuss how to utilize the ADMM to solve the resulting optimization problem.

Suppose we are given a training dataset consisting of n pairs of data point $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where the ith data point $\mathbf{x}_i \in \mathbb{R}^d$ lies in a d-dimensional real space and its label $y_i \in \mathbb{Z}$ is a nonnegative integer.

Since the outputs here are integer values, it is very natural to use the Poisson distribution as the likelihood function:

$$y_i|\mathbf{x}_i \sim \mathcal{P}(\mu_i)$$
$$\ln \mu_i = \boldsymbol{\alpha}^T \mathbf{x}_i,$$

where $k!$ is the factorial of an integer k, and $\mathcal{P}(\mu)$ denotes a Poisson distribution with its probability density function formulated as $p(x) = \mu^x \exp\{-\mu\}/x!$. In order to penalize the complexity of $\boldsymbol{\alpha}$, we add a normal prior on it as

$$\boldsymbol{\alpha} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\lambda}\mathbf{I}\right),$$

where $\mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$ denotes a multivariate (or univariate) normal distribution with mean as $\mathbf{m}$ and covariance matrix (or variance) as $\boldsymbol{\Sigma}$, $\mathbf{0}$ denotes a zero vector or matrix with appropriate size, and $\mathbf{I}$ denotes an identity matrix with the size depending on the context. Moreover, we suppose there is some constraint for the model parameters $\boldsymbol{\alpha}$, i.e., $\boldsymbol{\alpha} \in \mathcal{C}$ where $\mathcal{C}$ denotes a constraint set.

The maximum a posterior (MAP) solution seeks to solve the following optimization problem as

$$\min_{\boldsymbol{\alpha} \in \mathcal{C}} \sum_{i=1}^{n} \left(\exp\{\boldsymbol{\alpha}^T \mathbf{x}_i\} - y_i \boldsymbol{\alpha}^T \mathbf{x}_i\right) + \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_2^2, \tag{1}$$

where $\|\cdot\|_2$ denotes the two-norm of a vector. The Newton-Raphson method, which is frequently used to solve the objective function of the Poisson regression model, has the scalability problem with respect to the high-dimensional data since it needs to compute the inverse of the Hessian matrix and cannot handle the constraints imposed on the model parameters. To deal with the high-dimensional data and also the constraints, we adopt the ADMM algorithm to solve problem (1). By utilizing ADMM, the model parameters can be decoupled from the constraints by creating new variables as their copies and the resulting unconstrained problem can enable the use of the conventional first-order gradient method such as the conjugate gradient method. Specifically, we introduce a new variable $\boldsymbol{\beta}$ as a copy of the model parameter $\boldsymbol{\alpha}$ as reflected in constraints that $\boldsymbol{\alpha} = \boldsymbol{\beta}$ and $\boldsymbol{\beta} \in \mathcal{C}$. The reformulated objective function is defined as

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \sum_{i=1}^{n}\big(\exp\{\boldsymbol{\alpha}^T\mathbf{x}_i\} - y_i\boldsymbol{\alpha}^T\mathbf{x}_i\big) + \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_2^2, \\ \text{s.t.} \quad & \boldsymbol{\alpha} = \boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathcal{C}. \end{aligned} \tag{2}$$

We define the augmented Lagrangian as

$$\begin{aligned} L_\rho(\boldsymbol{\alpha},\boldsymbol{\beta},\theta) = & \sum_{i=1}^{n}\big(\exp\{\boldsymbol{\alpha}^T\mathbf{x}_i\} - y_i\boldsymbol{\alpha}^T\mathbf{x}_i\big) + \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_2^2 \\ & + \boldsymbol{\theta}^T(\boldsymbol{\alpha}-\boldsymbol{\beta}) + \frac{\rho}{2}\|\boldsymbol{\alpha}-\boldsymbol{\beta}\|_2^2, \end{aligned}$$

where θ acts as Lagrange multipliers and ρ is a penalty parameter. Based on the above notations, the ADMM algorithm is described in Algorithm 1. According to [12], its convergence rate is at least $O(1/k)$ with k as the number of iterations. Moreover, the use of the ADMM algorithm can enable the development of distributed algorithms as hinted in step 6 of Algorithm 1, which however goes beyond the focus of this paper and will be investigated in our future study.

Algorithm 1. The ADMM Algorithm for Problem (2)

1: Set ρ to be 10;
2: Initialize $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$;
3: $l := 0$;
4: **while** not converged **do**
5: $\quad\boldsymbol{\alpha}^{(l+1)} := \arg\min_{\boldsymbol{\alpha}} L_\rho(\boldsymbol{\alpha}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\theta}^{(l)})$;
6: $\quad\boldsymbol{\beta}^{(l+1)} := \arg\min_{\boldsymbol{\beta}\in\mathcal{C}} L_\rho(\boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\beta}, \boldsymbol{\theta}^{(l)})$;
7: $\quad\boldsymbol{\theta}^{(l+1)} := \boldsymbol{\theta}^{(l)} + \rho(\boldsymbol{\alpha}^{(l+1)} - \boldsymbol{\beta}^{(l+1)})$;
8: $\quad l := l + 1$;
9: **end while**

We need to solve the steps 5 and 6 in the ADMM algorithm. For step 5, the optimization problem is formulated as

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \ h(\boldsymbol{\alpha}) = & \sum_{i=1}^{n}(\exp\{\boldsymbol{\alpha}^T\mathbf{x}_i\} - y_i\boldsymbol{\alpha}^T\mathbf{x}_i) + \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_2^2 \\ & + (\boldsymbol{\theta}^{(l)})^T\boldsymbol{\alpha} + \frac{\rho}{2}\|\boldsymbol{\alpha}-\boldsymbol{\beta}^{(l)}\|_2^2. \end{aligned}$$

This is an unconstrained problem with no analytical solution due to the existent of the exponential function and we use some gradient method such as the conjugate gradient method to solve it. The gradient with respect to $\boldsymbol{\alpha}$ can be computed as

$$\begin{aligned} \frac{\partial h(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = & \sum_{i=1}^{n}\big(\exp\{\boldsymbol{\alpha}^T\mathbf{x}_i\} - y_i\big)\mathbf{x}_i + \lambda\boldsymbol{\alpha} + \boldsymbol{\theta}^{(l)} \\ & + \rho(\boldsymbol{\alpha} - \boldsymbol{\beta}^{(l)}). \end{aligned}$$

Step 6 solves an optimization problem formulated as

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \frac{\rho}{2}\|\boldsymbol{\beta} - \boldsymbol{\alpha}^{(l+1)}\|_2^2 - (\boldsymbol{\theta}^{(l)})^T\boldsymbol{\beta} \\ \text{s.t.} \quad & \boldsymbol{\beta} \in \mathcal{C}. \end{aligned}$$

The objective function is a quadratic function with respect to $\boldsymbol{\beta}$. If the constraint is convex with a simple structure, this problem will have an analytical solution. For example, if each element in $\boldsymbol{\beta}$ is only required to be nonnegative, the optimal $\boldsymbol{\beta}$ can be computed as

$$\boldsymbol{\beta} = \max\left(0, \boldsymbol{\alpha}^{(l+1)} + \frac{1}{\rho}\boldsymbol{\theta}^{(l)}\right),$$

where $\max(\cdot,\cdot)$ computes the maximum of the two input arguments in the elementwise manner.

In existing works on epidemic prediction such as [2], [10], [18], $\mathbf{x}_i$ represents a feature vector encoding some intra-regional elements and y_i is the corresponding number of infected cases. In the following two sections, we will show how to extend the basic Poisson regression model to take the inter-regional and external factors into consideration.

3 Combining Intra-Regional and Inter-Regional Factors

Suppose there are n regions with v_i denoting the ith region. Each year is divided into time intervals with fixed length (e.g., two weeks). Usually we focus on epidemic prediction year by year and hence without the loss of generality data are assumed to be sorted in years. Our observations, denoted by a set of y_{jk}^i, are the number of persons who got infected at v_j during the kth time interval of the ith year. Usually the observations are incomplete in one year and so the effective number of time intervals, i.e., those with observations, for the ith year denoted by l_i may be different in different years. Moreover, a vector $\mathbf{x}_{jk}^i$ encodes the features to represent the intra-regional factor of v_j in the kth time interval of the ith year.

We use Poisson distribution to model the likelihood function for y_{jk}^i as

$$y_{jk}^i \sim \mathcal{P}(\mu_{jk}^i) \tag{3}$$

$$\ln \mu_{jk}^i = \boldsymbol{\alpha}^T\mathbf{x}_{jk}^i + \sum_{s\in c_{jk}^i} w_{sj} \ln y_{s(k-1)}^i, \tag{4}$$

where $c_{jk}^i = \{s|\ y_{s(k-1)}^i > 0\}$ denotes the set of regions with infection cases reported in the $(k-1)$th interval of the ith year. According to Eq. (4), $\ln \mu_{jk}^i$ consists of two parts: the linear function of the intra-regional factor parameterized by $\boldsymbol{\alpha}$ which is identical to the basic model introduced in the previous section and the contribution of the infected cases of all regions at the last time interval parameterized by w_{sj}.

Here the inter-region factor is described by the matrix $\mathbf{W}$ whose (s,j)th element is w_{sj} and the interaction between v_i and v_j is reflected in w_{ij} and w_{ji}. Specifically, all regions form an epidemic network and w_{sj} represents the non-negative possibility of the disease propagating from v_s to v_j. In particular, w_{jj} can be non-zero and be viewed as an indicator for the unrecovered rate from the perspective of the widely used susceptible-infected-recovered (SIR) model [7] in epidemics. Here the matrix $\mathbf{W}$ can be viewed as a diffusion matrix which describes the diffusion possibility between each pair of the n regions. In Eq. (4), we make a first-order Markov assumption that the parameter μ^i_{jk} at current time interval k depends on the number of infected cases $y^i_{s(k-1)}$ of all regions at previous time interval $k-1$. The generalization to high-order Markov assumption is not difficult and will be investigated in our future study.

By defining $\ln 0 = 0$,[1] we can unify the two cases where $y^i_{s(k-1)}$ is equal to zero or not and simplify the second term in the right-hand-side of Eq. (4) as

$$\sum_{s\in c^i_{jk}} w_{sj} \ln y^i_{s(k-1)} = \sum_{s=1}^{n} w_{sj} \ln y^i_{s(k-1)}.$$

To penalize the complexity of $\boldsymbol{\alpha}$, we place a normal distribution as a prior on it as

$$\boldsymbol{\alpha} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\lambda_1}\mathbf{I}\right). \tag{5}$$

$\mathbf{W}$ contains n^2 entries. If we directly learn a dense $\mathbf{W}$, the model complexity may be high, leading to the overfitting problem, i.e., there may not be enough training data for accurate estimation of $\mathbf{W}$. To control the complexity of $\mathbf{W}$, two approaches are proposed to learn the structure of $\mathbf{W}$; the first approach is to use some a priori information (e.g., transportation network) while the second one directly learns from data based on a sparsity assumption without using a priori information. We will discuss those two approaches respectively in the following two sections.

3.1 Learning W Based on A Priori Information

Suppose from available resources like Google map, one can obtain some a priori information, e.g., transportation network $\mathbf{A}$ where the (i,j)th element of $\mathbf{A}$, denoted by a_{ij}, equals 1 if two regions v_i and v_j are connected via some trafficway and 0 otherwise. Usually the transportation network is very sparse and for example the sparsity of the transportation network used in our experiments is about 90 percent. Then by assuming that the structure of $\mathbf{W}$ is similar to that of the transportation network, we can define the structure of $\mathbf{W}$ as

$$\begin{aligned} w_{ij} &= 0 \text{ if } a_{ij} = 0 \\ w_{ij} &\geq 0 \text{ if } a_{ij} > 0. \end{aligned}$$

We also assume that v_i is self-connected and hence $a_{ii} = 1$ which makes $w_{ii} \geq 0$ and indicates the existent of the self-infection within one region. Note that the structure of $\mathbf{W}$ is not identical to that of $\mathbf{A}$ since $a_{ij} > 0$ can only imply $w_{ij} \geq 0$ but not $w_{ij} > 0$. Due to the non-negativeness of $\{w_{ij}\}$, we place a half-normal distribution on nonzero elements of $\mathbf{W}$ as

$$w_{ij} \sim \mathcal{HN}\left(0, \frac{1}{\lambda_2}\right) \text{ if } a_{ij} > 0. \tag{6}$$

The MAP solution leads to the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\mathbf{W}\in\mathcal{S}_W} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}(\mu^i_{jk} - y^i_{jk}\ln\mu^i_{jk}) + \frac{\lambda_1}{2}\|\boldsymbol{\alpha}\|_2^2 \\ & + \frac{\lambda_2}{2}\sum_{a_{ij}>0} w_{ij}^2, \end{aligned} \tag{7}$$

where $\mathcal{S}_W = \{\mathbf{W} | w_{ij} = 0 \text{ if } a_{ij} = 0; \text{otherwise } w_{ij} \geq 0\}$ and μ^i_{jk} is defined in Eq. (4).

It is easy to show that problem (7) is convex since $\ln\mu^i_{jk}$ is a linear function of $\boldsymbol{\alpha}$ and $\mathbf{W}$. Since there is one constraint for $\mathbf{W}$ and the number of parameters is not small, we use the ADMM algorithm to solve problem (7). By introducing a new variable $\mathbf{U}$ as a copy of $\mathbf{W}$, we can reformulate problem (7) as

$$\begin{aligned} \min_{\boldsymbol{\alpha},\mathbf{W},\mathbf{U}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}(\mu^i_{jk} - y^i_{jk}\ln\mu^i_{jk}) + \frac{\lambda_1}{2}\|\boldsymbol{\alpha}\|_2^2 \\ & + \frac{\lambda_2}{2}\sum_{a_{ij}>0} w_{ij}^2 \\ \text{s.t.} \quad & \mathbf{U} = \mathbf{W}, \mathbf{U} \in \mathcal{S}_W. \end{aligned} \tag{8}$$

Since there is no constraint placed on $\boldsymbol{\alpha}$, we need not create a copy for it. In order to use the ADMM algorithm, we define the augmented Lagrangian as

$$\begin{aligned} & L_\rho(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{U}, \boldsymbol{\Theta}) \\ & = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}(\mu^i_{jk} - y^i_{jk}\ln\mu^i_{jk}) + \frac{\lambda_1}{2}\|\boldsymbol{\alpha}\|_2^2 + \frac{\lambda_2}{2}\sum_{a_{ij}>0} w_{ij}^2 \\ & \quad + \operatorname{tr}\big(\boldsymbol{\Theta}^T(\mathbf{U}-\mathbf{W})\big) + \frac{\rho}{2}\|\mathbf{U}-\mathbf{W}\|_F^2, \end{aligned}$$

where $\operatorname{tr}(\cdot)$ denotes the trace of a square matrix and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. To use Algorithm 1, we optimize $\boldsymbol{\alpha}$ and $\mathbf{W}$ together as in step 4 and $\mathbf{U}$ in step 5. Specifically, we need to solve two subproblems: $\min_{\boldsymbol{\alpha},\mathbf{W}} L_\rho(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{U}, \boldsymbol{\Theta})$ and $\min_{\mathbf{U}\in\mathcal{S}_W} L_\rho(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{U}, \boldsymbol{\Theta})$. Here for notational simplicity, we omit the superscripts which indicate the number of iterations. In the following we discuss how to solve those two problems.

To solve $\min_{\boldsymbol{\alpha},\mathbf{W}} L_\rho(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{U}, \boldsymbol{\Theta})$ which is an unconstrained problem, we use the conjugate gradient method with the gradients with respect to $\boldsymbol{\alpha}$ and the (r,s)th element w_{rs} of $\mathbf{W}$ computed as

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\alpha}} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}(\mu^i_{jk} - y^i_{jk})\mathbf{x}^i_{jk} + \lambda_1\boldsymbol{\alpha} \\ \frac{\partial l}{\partial w_{rs}} &= \sum_{i=1}^{m}\sum_{k=2}^{l_i}(\mu^i_{sk} - y^i_{sk})\ln y^i_{r(k-1)} + \lambda_2 w_{rs} - \theta_{rs} \\ & \quad + \rho(w_{rs} - u_{rs}), \end{aligned} \tag{9}$$

1. This definition only works in the formulation of $\ln\mu^i_{jk}$.

where l denotes the objective function, μ_{jk}^i is computed according to Eq. (4), and u_{ij} and θ_{ij} are the (i,j)th elements of $\mathbf{U}$ and $\mathbf{\Theta}$ respectively.

The second subproblem can be expressed as

$$\min_{\mathbf{U}} \ \mathrm{tr}\big(\mathbf{\Theta}^T\mathbf{U}\big) + \frac{\rho}{2}\|\mathbf{U}-\mathbf{W}\|_F^2 \quad \text{s.t. } \mathbf{U} \in \mathcal{S}_W. \tag{10}$$

This problem has an analytical solution as

$$u_{ij} = \begin{cases} 0 & \text{if } a_{ij} = 0 \\ \max\Big(0, w_{ij} - \frac{1}{\rho}\theta_{ij}\Big) & \text{otherwise.} \end{cases} \tag{11}$$

Those two subproblems are iteratively solved until some termination criteria is satisfied. In our experiments, we find that the problem normally converges in less than 10 iterations and so the convergence is fast.

3.2 Learning W Based on Sparsity Assumption

In this section, we discuss how to learn $\mathbf{W}$ from scratch where no information is available about the structure of $\mathbf{W}$.

We assume that each region is reachable from any other regions and so each element in $\mathbf{W}$ can be non-negative. To restrict the complexity of $\mathbf{W}$ and to discover the hot spots for epidemic prediction, we assume $\mathbf{W}$ is sparse and hence place a Laplace prior, which corresponds to the l_1 regularization, on non-negative $\mathbf{W}$ as

$$w_{ij} \sim \mathcal{L}\left(0, \frac{1}{\lambda_2}\right), \tag{12}$$

where w_{ij} is the (i,j)th element of $\mathbf{W}$, $|\cdot|$ denotes the absolute value of a scalar, and $\mathcal{L}(a,b)$ denotes a Laplace distribution with its probability density function as $p(x) = \frac{1}{b}\exp\{-\frac{1}{b}|x-a|\}$. Then we can formulate the objective function of the MAP solution as

$$\begin{aligned} \min_{\boldsymbol{\alpha},\mathbf{W}\geq\mathbf{0}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}(\mu_{jk}^i - y_{jk}^i \ln \mu_{jk}^i) + \frac{\lambda_1}{2}\|\boldsymbol{\alpha}\|_2^2 \\ & + \lambda_2 \sum_{i,j} w_{ij}. \end{aligned} \tag{13}$$

Obviously problem (13) is a convex optimization problem with respect to $\boldsymbol{\alpha}$ and $\mathbf{W}$ and we again use the ADMM algorithm to solve it. Similar to the previous section, we reformulate problem (13) as

$$\begin{aligned} & \min_{\boldsymbol{\alpha},\mathbf{W}} \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}(\mu_{jk}^i - y_{jk}^i \ln \mu_{jk}^i) + \frac{\lambda_1}{2}\|\boldsymbol{\alpha}\|_2^2 + \lambda_2 \sum_{i,j} w_{ij} \\ & \text{s.t. } \mathbf{U} = \mathbf{W}, \mathbf{U} \geq \mathbf{0}. \end{aligned} \tag{14}$$

The augmented Lagrangian is defined as

$$\begin{aligned} & L_\rho(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{U}, \mathbf{\Theta}) \\ & = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}(\mu_{jk}^i - y_{jk}^i \ln \mu_{jk}^i) + \frac{\lambda_1}{2}\|\boldsymbol{\alpha}\|_2^2 + \lambda_2 \sum_{i,j} w_{ij} \\ & \quad + \mathrm{tr}\big(\mathbf{\Theta}^T(\mathbf{U}-\mathbf{W})\big) + \frac{\rho}{2}\|\mathbf{U}-\mathbf{W}\|_F^2. \end{aligned}$$

The first subproblem with respect to $\boldsymbol{\alpha}$ and $\mathbf{W}$ is similar to that of the problem in the previous section with the difference lying in the regularizor of $\mathbf{W}$. The gradient method is used to solve it and the gradient with respect to $\boldsymbol{\alpha}$ is the same as Eq. (9). The gradient with respect to w_{rs} for $r,s = 1,\ldots,n$ can be computed as

$$\begin{aligned} \frac{\partial l}{\partial w_{rs}} = & \sum_{i=1}^{m}\sum_{k=2}^{l_i}(\mu_{sk}^i - y_{sk}^i)\ln y_{r(k-1)}^i + \lambda_2 - \theta_{rs} \\ & + \rho(w_{rs} - u_{rs}). \end{aligned}$$

The objective function of the second subproblem which minimizes $L_\rho(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{U}, \mathbf{\Theta})$ with respect to $\mathbf{U}$ can be simplified as

$$\begin{aligned} \min_{\mathbf{U}} \quad & \mathrm{tr}\big(\mathbf{\Theta}^T\mathbf{U}\big) + \frac{\rho}{2}\|\mathbf{U}-\mathbf{W}\|_F^2 \\ \text{s.t.} \quad & \mathbf{U} \geq \mathbf{0}, \end{aligned} \tag{15}$$

which has an analytical solution as

$$u_{ij} = \max\left(0, w_{ij} - \frac{1}{\rho}\theta_{ij}\right). \tag{16}$$

4 Incorporating External Factor

In the previous section, we study how to combine the intra-regional and inter-regional factors together. In this section, we further investigate the combination of all three factors together based on the preceding model.

Here the external factor mainly denotes the imported cases from the regions outside the n regions under investigation. The main challenge here is that we have little information about the external factor. It is well known that the epidemic curves of most epidemics, which record the infected cases in one region for successive time intervals in one year, are (approximately) unimodal distributions. By assuming that the importing rate is fixed or the variance on the importing rates over time is very small, the curve of the imported cases over time can also be viewed as an (approximately) unimodal distribution. We propose to use a parametric function of the time information to reflect this property. One choice for the parametric function is a quadratic function since a quadratic function with a negative leading coefficient is unimodal and it is very simple and intuitive. Specifically, the parameter in the Poisson likelihood μ_{jk}^i is defined as

$$\begin{aligned} \ln \mu_{jk}^i = & \boldsymbol{\alpha}^T\mathbf{x}_{jk}^i + \sum_{s\in c_{jk}^i} w_{sj} \ln y_{s(k-1)}^i + \big((\boldsymbol{\beta}^{(2)})^T\hat{\mathbf{z}}_j^i\big)k^2 \\ & + \big((\boldsymbol{\beta}^{(1)})^T\hat{\mathbf{z}}_j^i\big)k + (\boldsymbol{\beta}^{(0)})^T\hat{\mathbf{z}}_j^i, \end{aligned}$$

or equivalently

$$\ln \mu_{jk}^i = \boldsymbol{\alpha}^T\mathbf{x}_{jk}^i + \sum_{s\in c_{jk}^i} w_{sj} \ln y_{s(k-1)}^i + \boldsymbol{\beta}^T\mathbf{z}_{jk}^i, \tag{17}$$

where $\hat{\mathbf{z}}_j^i$ record some features (e.g., the elevation, the income level, and the population), which belongs to the intra-regional factor, and is assumed to affect disease importing for the jth region in the ith year, $\boldsymbol{\beta} = ((\boldsymbol{\beta}^{(2)})^T, (\boldsymbol{\beta}^{(1)})^T, (\boldsymbol{\beta}^{(0)})^T)^T$, and $\mathbf{z}_{jk}^i = (k^2(\hat{\mathbf{z}}_j^i)^T, k(\hat{\mathbf{z}}_j^i)^T, (\hat{\mathbf{z}}_j^i)^T)^T$.

In order to guarantee that the distribution of the external factor has one peak at one of the time intervals, it is better to add constraints that $(\boldsymbol{\beta}^{(2)})^T\hat{\mathbf{z}}_j^i$ is negative for all i and j and $(\boldsymbol{\beta}^{(1)})^T\hat{\mathbf{z}}_j^i$ is positive for all i and j as a quadratic function with a negative leading coefficient and a positive linear coefficient has only one peak at some positive scalar. However, by adding those constraints, the optimization procedure to solve the MAP solution will become very complicated. In the following, we do not consider the incorporation of those constraints but instead use those constraints as a criterion to test whether our assumption matches the real-world data used in the experiments.

In the following two sections, similar to the previous section we respectively discuss two approaches to combine three factors together according to the situations whether some a priori information on the network structure $\mathbf{W}$ is available or not.

4.1 Learning W Based on A Priori Information

In this section, suppose we have a priori information, e.g., the transportation network $\mathbf{A}$, on the structure of $\mathbf{W}$, which corresponds to the situation in Section 3.1.

We also assign normal priors on $\boldsymbol{\alpha}$, w_{ij} and $\boldsymbol{\beta}$ as

$$\begin{aligned}\boldsymbol{\alpha} &\sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\lambda_1}\mathbf{I}\right)\\ w_{ij} &\sim \mathcal{HN}\left(0, \frac{1}{\lambda_2}\right) \text{ if } a_{ij} > 0\\ \boldsymbol{\beta} &\sim \mathcal{N}\left(0, \frac{1}{\lambda_3}\mathbf{I}\right).\end{aligned}$$

By computing the MAP solution, we get the following optimization problem as

$$\begin{aligned}\min_{\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}\left(\mu_{jk}^i - y_{jk}^i \ln \mu_{jk}^i\right) + \frac{\lambda_1}{2}\|\boldsymbol{\alpha}\|_2^2 + \frac{\lambda_3}{2}\|\boldsymbol{\beta}\|_2^2\\ & + \frac{\lambda_2}{2}\sum_{a_{ij}>0} w_{ij}^2\\ \text{s.t. } & \mathbf{W} \in \mathcal{S}_W,\end{aligned} \tag{18}$$

where μ_{jk}^i is defined in Eq. (17). It is easy to show that problem (18) is convex since $\ln \mu_{jk}^i$ is a linear function of the model parameters, and hence we can adopt the ADMM algorithm to solve the problem. By introducing a new variable $\mathbf{U}$, we can reformulate problem (18) as

$$\begin{aligned}\min_{\mathbf{W},\mathbf{U},\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}\left(\mu_{jk}^i - y_{jk}^i \ln \mu_{jk}^i\right) + \frac{\lambda_1}{2}\|\boldsymbol{\alpha}\|_2^2\\ & + \frac{\lambda_3}{2}\|\boldsymbol{\beta}\|_2^2 + \frac{\lambda_2}{2}\sum_{a_{ij}>0} w_{ij}^2\\ \text{s.t. } & \mathbf{U} = \mathbf{W}, \mathbf{U} \in \mathcal{S}_W.\end{aligned} \tag{19}$$

The augmented Lagrangian is defined as

$$\begin{aligned}&L_\rho(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}, \mathbf{U}, \boldsymbol{\Theta})\\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}\left(\mu_{jk}^i - y_{jk}^i \ln \mu_{jk}^i\right) + \frac{\lambda_1}{2}\|\boldsymbol{\alpha}\|_2^2 + \frac{\lambda_2}{2}\sum_{a_{ij}>0} w_{ij}^2\\ &\quad + \frac{\lambda_3}{2}\|\boldsymbol{\beta}\|_2^2 + \operatorname{tr}\left(\boldsymbol{\Theta}^T(\mathbf{U}-\mathbf{W})\right) + \frac{\rho}{2}\|\mathbf{U}-\mathbf{W}\|_F^2.\end{aligned}$$

For the subproblem in step 4 of the ADMM algorithm, we need to minimize $L_\rho(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}, \mathbf{U}, \boldsymbol{\Theta})$ with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{W}$ by using some gradient method with the gradients computed as

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\alpha}} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}\left(\mu_{jk}^i - y_{jk}^i\right)\mathbf{x}_{jk}^i + \lambda_1\boldsymbol{\alpha}\\ \frac{\partial l}{\partial w_{rs}} &= \sum_{i=1}^{m}\sum_{k=2}^{l_i}\left(\mu_{sk}^i - y_{sk}^i\right)\ln y_{r(k-1)}^i + \lambda_2 w_{rs} - \theta_{rs}\\ &\quad + \rho(w_{rs} - u_{rs})\\ \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}\left(\mu_{jk}^i - y_{jk}^i\right)\mathbf{z}_{jk}^i + \lambda_3\boldsymbol{\beta}.\end{aligned}$$

The subproblem in step 5 solves the same problem as problem (10) with the solution as in Eq. (11).

4.2 Learning W Based on Sparsity Assumption

When there is no a priori information, we will learn $\mathbf{W}$ from data directly. Here we assume that $\mathbf{W}$ is sparse and place a Laplace prior on non-negative $\mathbf{W}$ as

$$w_{ij} \sim \mathcal{L}\left(0, \frac{1}{\lambda_2}\right), \tag{20}$$

where w_{ij} is the (i,j)th element of $\mathbf{W}$. The objective function of the MAP solution is formulated as

$$\begin{aligned}\min_{\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=2}^{l_i}\left(\mu_{jk}^i - y_{jk}^i \ln \mu_{jk}^i\right) + \frac{\lambda_1}{2}\|\boldsymbol{\alpha}\|_2^2\\ & + \lambda_2\sum_{i,j} w_{ij} + \frac{\lambda_3}{2}\|\boldsymbol{\beta}\|_2^2\\ \text{s.t. } & \mathbf{W} \geq \mathbf{0}.\end{aligned} \tag{21}$$

We still use the ADMM method to solve problem (21) with the difference lying in the regularization term on $\mathbf{W}$ compared to problem (18), leading to a different gradient with respect to $\mathbf{W}$ in the first subproblem of the ADMM algorithm as

$$\begin{aligned}\frac{\partial l}{\partial w_{rs}} &= \sum_{i=1}^{m}\sum_{k=2}^{l_i}\left(\mu_{sk}^i - y_{sk}^i\right)\ln y_{r(k-1)}^i + \lambda_2 - \theta_{rs}\\ &\quad + \rho(w_{rs} - u_{rs}).\end{aligned}$$

The gradients with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ remain identical to those of the problem in the previous section. The second subproblem is identical to problem (15) with the solution as in Eq. (16).

TABLE 1
The Number of the Time Intervals with Reported Cases for Different Years in the Malaria Dataset

Year	The Number of Time Intervals l_i
2005	23
2006	10
2007	23
2008	23
2009	23
2010	23
2011	23

5 Experiments

In this section, we conduct empirical experiments to test the performance of our proposed methods.

5.1 Data

To our knowledge, there exists no public epidemic data. We use a malaria dataset provided by the National Institute of Parasitic Diseases in the Chinese Center for Disease Control and Prevention. This dataset was collected from the endemic areas in Yunnan province of China and containing 16-days-reported malaria cases from 62 counties in Yunnan province from years 2005 to 2011. So the length of each time interval is 16 (in days) and the number of years under investigation m is 7. Due to the time-consuming data collection process, the available data are usually incomplete. Table 1 depicts the number of the time intervals found in the dataset. We can see that the data corresponding to 13 times intervals are missing for year 2006. With each region is defined as one of the 62 counties in Yunnan province, the number of regions n for our experiments is equal to 62.

We have manually collected the environmental information[2] corresponding to each county in Yunnan province during the period of years 2005-2011, including the population denoted by p_j of region v_j, the temperature denoted by t^i_{jk} during the kth time interval of the ith year at region v_j, the rainfall denoted by r^i_{jk} during the kth time interval of the ith year at v_j, and the elevation denoted by e_j of region v_j. Then we define the feature representation for the intra-regional factor $\mathbf{x}^i_{jk}$ as $\mathbf{x}^i_{jk} = (\ln p_j, \ln t^i_{jk}, t^i_{jk}, \ln r^i_{jk}, r^i_{jk}, \ln e_j)^T$.

The Yunnan province is close to the boundary between China and Myanmar. The medical condition for malaria treatment in Myanmar is not very good, resulting in many infected cases. Many business activities between China and Myanmar happening in Yunnan province leads to many imported cases where the subjects got infected in Myanmar before accessing the Yunnan province. In general, the imported case has been considered to be one of the most important factors for malaria diffusion in Yunnan province, making the investigation of the external factor crucial. We define $\hat{\mathbf{z}}^i_j$, the feature vector for the external factor, as $\hat{\mathbf{z}}^i_j = (p_j, \ln p_j, e_j, \ln e_j, b_j)^T$ where b_j is a binary feature to indicate whether the corresponding region v_j has a direct connection to the boundary or not.

2. The MODIS data for temperature can be found at http://iridl.ldeo.columbia.edu/expert/SOURCES/.USGS/.LandDAAC/.MODIS/.1km/.8day/.version_005/.Aqua/.CN/.Day/and the TRMM data for rainfall is from http://iridl.ldeo.columbia.edu/expert/SOURCES/.NASA/.GES-DAAC/.TRMM_L3/.TRMM_3B42/.v6/.daily/.precipitation/.

5.2 Experimental Settings

We abbreviate the classical Poisson regression model introduced in Section 2 to 'PR'. We name the model that combines the intra-regional and inter-regional factors with priori structural information of the diffusion matrix, which is depicted in Section 3.1, as the PR-2-p model and that with the sparsity assumption in Section 3.2 as the PR-2-s model. Their counterparts with all three factors combined are called the PR-3-p and PR-3-s models respectively. Moreover, we also compare with the semi-supervised Gaussian process ordinary regression (SSGPOR) [19], which uses all the information including the transportation network to learn the kernel matrix.

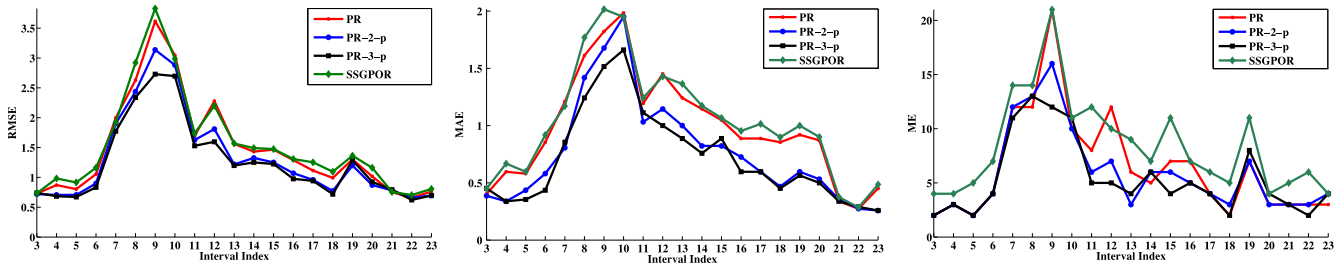
In order to test the performance of all the proposed models, we adopt the notation denoted by (i, j) which indicates an experiment setting using the malaria cases from the first time interval of the first year (i.e., year 2005) to the $(j-1)$th time interval of the ith year as the training data to predict the number of infected cases in the jth time interval of the ith year. In order to obtain the inter-regional factor, we need the training data to contain at least two preceding time intervals, which implies $j - 1 \geq 2$ or equivalently $j \geq 3$. Moreover, in order to make sure that the training dataset contains sufficient data for accurate estimation, we further require that $i \geq 4$, i.e., there should be at least three years of data for training. As a result, there are four sets of experiments, including the settings $(4, j)$, $(5, j)$, $(6, j)$, and $(7, j)$ with the range of j between 3 and 23.

We use the Bayesian regularization method in [11] for the setting of the regularization parameters (i.e., λ_1 and λ_2) in all models. The core idea of [11] is to first place a Gamma prior on each of the regularization parameters and then integrate out the regularization parameters. By using the majorization-minimization (MM) algorithm [14], [15], [22], the objective function in each iteration is similar to the original problem with the regularization parameters inversely depending on the solution in the previous iteration. One advantage of the Bayesian regularization method over the traditional cross-validation-style methods is that we only need to optimize the problem once instead of solving it for multiple times which can grow exponentially in terms of the number of regularization parameters. Hence the Bayesian regularization method is efficient especially for models with multiple regularization parameters.

5.3 Evaluation Measure

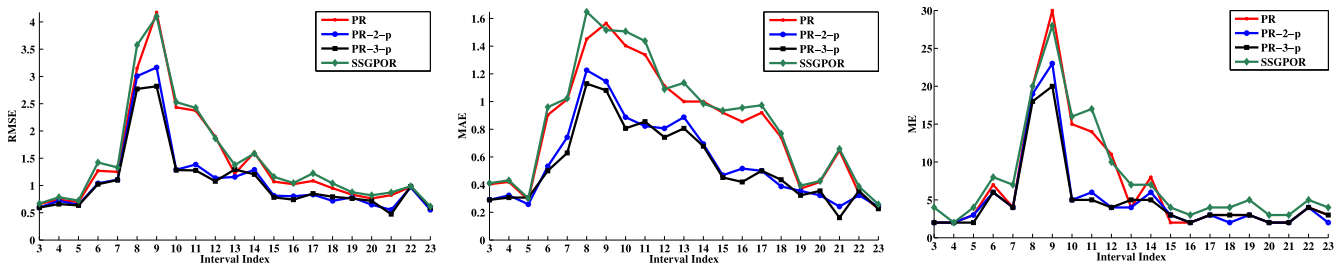
Three performance measures are used in our experiments:

- Root mean square error (RMSE). $\mathrm{RMSE}(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, where $\mathbf{y}$ denotes the ground truth on the number of infected cases for all n regions under investigation in a particular time interval, $\hat{\mathbf{y}}$ denotes the estimation of the number of the infected cases for all n regions in investigation in the same time interval, y_i and $\hat{y}_i$ correspond to the ith elements of $\mathbf{y}$ and $\hat{\mathbf{y}}$, representing the ground truth and



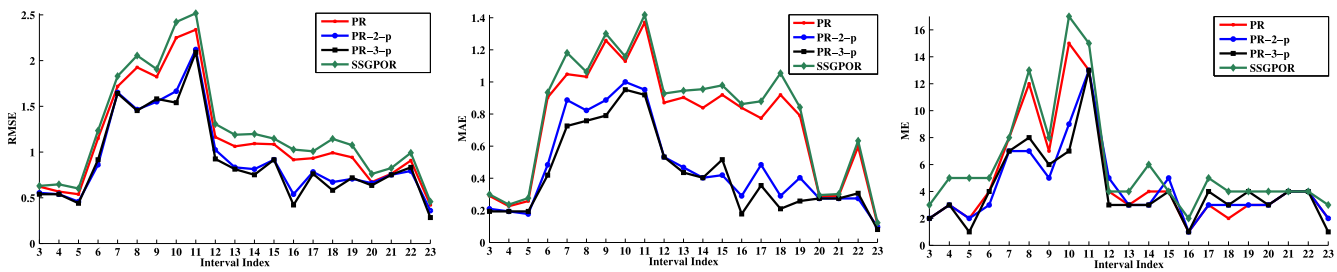
(a) RMSE for setting $(4, j)$ (b) MAE for setting $(4, j)$ (c) ME for setting $(4, j)$

Fig. 1. Performance comparison (in terms of RMSE, MAE, and ME) between the SSGPOR, PR, PR-2-p, and PR-3-p models under settings $(4, j)$ where $3 \leq j \leq 23$.



(a) RMSE for setting $(5, j)$ (b) MAE for setting $(5, j)$ (c) ME for setting $(5, j)$

Fig. 2. Performance comparison (in terms of RMSE, MAE, and ME) between the SSGPOR, PR, PR-2-p, and PR-3-p models under settings $(5, j)$ where $3 \leq j \leq 23$.



(a) RMSE for setting $(6, j)$ (b) MAE for setting $(6, j)$ (c) ME for setting $(6, j)$

Fig. 3. Performance comparison (in terms of RMSE, MAE, and ME) between the SSGPOR, PR, PR-2-p, and PR-3-p models under settings $(6, j)$ where $3 \leq j \leq 23$.

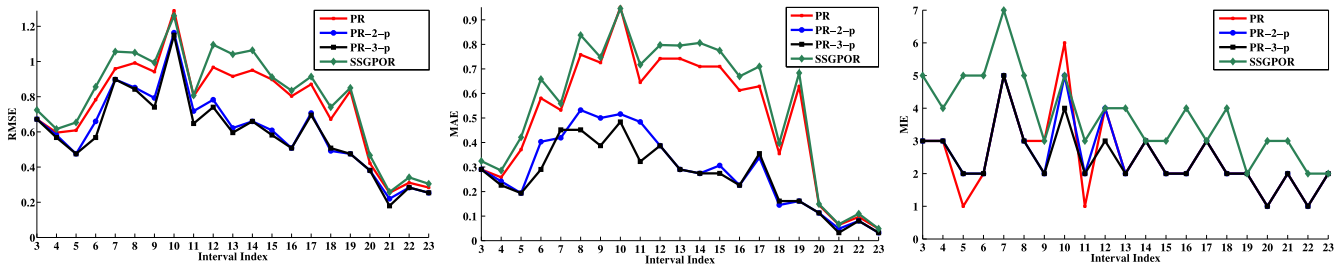
estimation of the number of the infected cases in the ith region respectively.

- Mean absolute error (MAE). $\text{MAE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$, where $|\cdot|$ denotes the absolute value of a scalar.
- Maximum absolute error (ME). $\text{ME}(\mathbf{y}, \hat{\mathbf{y}}) = \max_i |y_i - \hat{y}_i|$.

RMSE and MAE measure the prediction error on average but under different metric spaces (i.e., spaces with ℓ_2 and ℓ_1 distance). Different from those two measures, ME measures the prediction error for the worst case by reporting the maximum of the absolute error among all n regions. Those three measures evaluate different aspects of a learning model and we will use all the three measures to fully analyze each learning model.

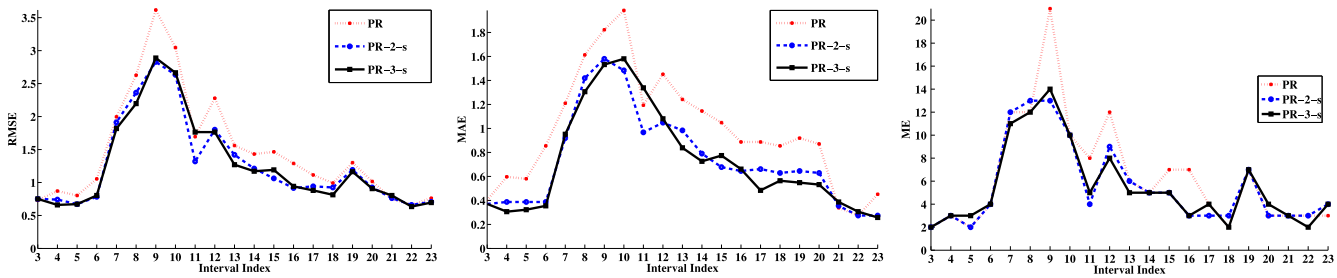
5.4 Experimental Results

The results for the four experimental settings are depicted in Figs. 1, 2, 3, 4, 5, 6, 7, and 8. For clear presentation, the comparison results for the PR, SSGPOR, PR-2-p and PR-3-p models with a priori information on the structure of the diffusion matrix assumed are shown in Figs. 1, 2, 3, and 4 and those for the PR, PR-2-s, and PR-3-s models without a priori information in Figs. 5, 6, 7, and 8. Since the SSGPOR model performs comparably with the PR model as shown in Figs. 1, 2, 3, and 4, leading to the inferiority to the PR-2-s and PR-3-s models, we do not plot the results of the SSGPOR model in Figs. 5, 6, 7, and 8 for clear illustration. From the results, we can see at the beginning of the year (i.e., intervals 3-5), our proposed models have comparable performance with the PR model based on some measure (e.g., the ME measure). One reason is that the epidemic has not outbroken due to unsuitable weather condition (e.g., low temperature) and the number of infected cases is very small, making the consideration of the inter-regional and external factors not very important. This phenomenon also exists near the end of each year (i.e., intervals 20-23), another inactive period for the epidemic. During the outbreak (i.e., intervals 6-19), we can see that the performance of the PR-2-p and PR-2-s models is better than that of the PR model under the four settings, which demonstrates that the incorporation of the inter-regional factor is useful for epidemic prediction. By adding the external factor, the PR-3-p and PR-3-s models give further enhancement in performance compared to their counterparts, i.e., the PR-2-p and



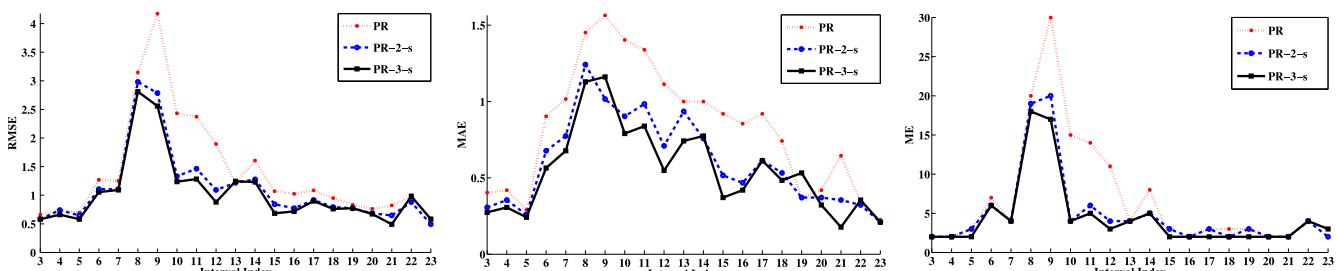
(a) RMSE for setting $(7, j)$ (b) MAE for setting $(7, j)$ (c) ME for setting $(7, j)$

Fig. 4. Performance comparison (in terms of RMSE, MAE, and ME) between the SSGPOR, PR, PR-2-p, and PR-3-p models under settings $(7, j)$ where $3 \leq j \leq 23$.



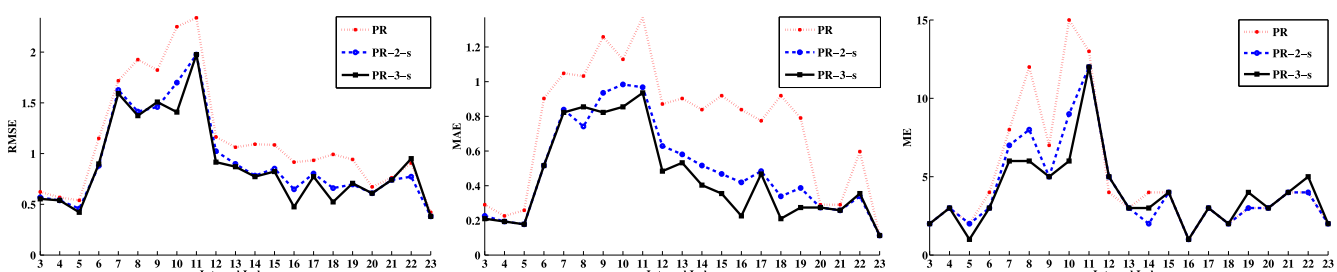
(a) RMSE for setting $(4, j)$ (b) MAE for setting $(4, j)$ (c) ME for setting $(4, j)$

Fig. 5. Performance comparison (in terms of RMSE, MAE, and ME) between the PR, PR-2-s, and PR-3-s models under settings $(4, j)$ where $3 \leq j \leq 23$.



(a) RMSE for setting $(5, j)$ (b) MAE for setting $(5, j)$ (c) ME for setting $(5, j)$

Fig. 6. Performance comparison (in terms of RMSE, MAE, and ME) between the PR, PR-2-s, and PR-3-s models under settings $(5, j)$ where $3 \leq j \leq 23$.

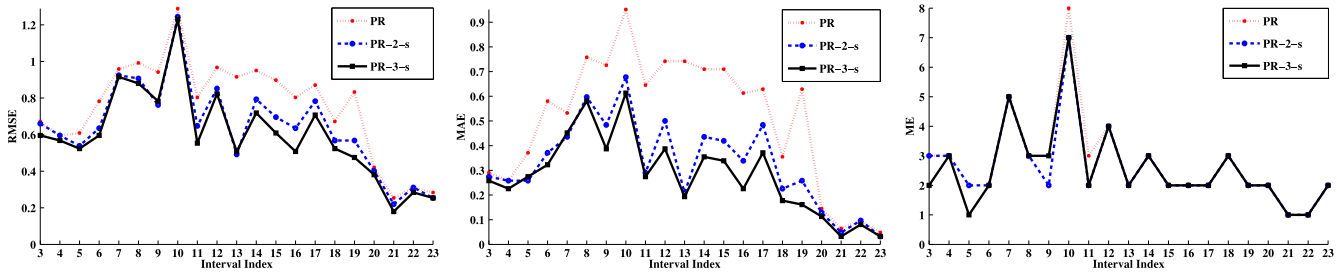


(a) RMSE for setting $(6, j)$ (b) MAE for setting $(6, j)$ (c) ME for setting $(6, j)$

Fig. 7. Performance comparison (in terms of RMSE, MAE, and ME) between the PR, PR-2-s, and PR-3-s models under settings $(6, j)$ where $3 \leq j \leq 23$.

PR-2-s models. For example, as shown in Fig. 1c, the ME of the PR model is 21 and the ME's for our PR-2-p and PR-3-p models are 16 and 12, with the error reduction ratio being 23.80 and 42.85 percent respectively. The significant error reduction can also be observed in Figs. 2c, 5c, 6c and so on.

We also report the mean and the standard deviation of the six methods over each year in Tables 2, 3, and 4. Since the number of infected cases changes a lot in different time intervals of a year, to make the measure in different time intervals comparable, we use the relative RMSE (rRMSE), relative MAE (rMAE), and relative ME (rME), which equal the ratio of the corresponding measure (RMSE, MAE, or ME) over the average of the true number of infected cases during a time interval, as the performance measures. Specifically, to measure the difference between the ground truth $\mathbf{y}$ on the number of infected cases for all n regions in a time



(a) RMSE for setting $(7, j)$ (b) MAE for setting $(7, j)$ (c) ME for setting $(7, j)$

Fig. 8. Performance comparison (in terms of RMSE, MAE, and ME) between the PR, PR-2-s, and PR-3-s models under settings $(7, j)$ where $3 \leq j \leq 23$.

interval and the corresponding estimation $\hat{\mathbf{y}}$, the rRMSE, rMAE, and rME are defined as

$$\text{rRMSE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\text{RMSE}(\mathbf{y}, \hat{\mathbf{y}})}{\text{avg}(\mathbf{y})}$$
$$\text{rMAE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\text{MAE}(\mathbf{y}, \hat{\mathbf{y}})}{\text{avg}(\mathbf{y})}$$
$$\text{rME}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\text{ME}(\mathbf{y}, \hat{\mathbf{y}})}{\text{avg}(\mathbf{y})},$$

where $\text{avg}(\mathbf{y})$ gives the average of the elements in $\mathbf{y}$. By conducting significance t-test with 95 percent confidence on those results, we find that the SSGPOR and PR methods are comparable, the PR-2-p and PR-2-s methods, which are comparable in performance, perform significantly better than the SSGPOR and PR methods, and the PR-3-p and PR-3-s methods are among the best methods.

To compare the two strategies to learn the diffusion matrix incorporated in the inter-regional factor, i.e., presetting the structure of $\mathbf{W}$ based on the given transportation network or learning $\mathbf{W}$ from scratch with the sparsity assumption, Table 5 records the win/tie/loss results in terms of different measures by comparing the PR-2-p and PR-2-s models as well as the PR-3-p and PR-3-s models respectively. Based on the RMSE measure, we can see those two strategies perform comparably due to the comparable win and loss counts. In terms of the MAE measure, the second strategy, which learns the diffusion matrix from scratch based on the sparsity assumption, performs better than the first one. However, the situation is different for the ME measure in which the first strategy has slightly better performance. In summary, according to the performance on the three measures, those two strategies perform comparably and each strategy has its own favor. That is, the first strategy favors the ME measure while the second one is better based on the MAE measure.

One reason that using a priori information (i.e., the transportation network) cannot bring performance improvement is that the priori information may be inaccurate or incomplete. The transportation information provided by the Google map may not be up-to-date and the transportation network, which is manually recorded by ourselves from the Google map, may contain some noise, e.g., missing a connection between two regions or wrongly adding a nonexistent connection. Another possible reason is that the transportation network only gives the direct connection between a few pairs of regions but the epidemic can propagate between two regions, which have no direct connection in the transportation network, via some intermediate regions through the traveling of disease-carriers, which may limit the expressive power of the proposed models. In our future studies, we will try other ways to design the structure of $\mathbf{W}$ based on the transportation network $\mathbf{A}$.

Moreover, to see the functional shape of the external factor, we compute $(\boldsymbol{\beta}^{(2)})^T \hat{\mathbf{z}}_j^i$ and $(\boldsymbol{\beta}^{(1)})^T \hat{\mathbf{z}}_j^i$ for $i \geq 4$ and

TABLE 2
Comparison of the rRMSE among the Six Methods for Years 4 to 7

Method	4th Year	5th Year	6th Year	7th Year
PR	2.1332	2.3045	2.6821	3.6210
	0.2883	0.2314	0.3199	0.3719
PR-2-p	1.8899	1.8758	2.1863	2.9170
	0.1825	0.2236	0.2215	0.2749
PR-2-s	1.8625	1.8971	2.1962	3.0596
	0.1576	0.2656	0.2730	0.2741
PR-3-p	1.5245	1.5391	1.8867	2.8081
	0.1851	0.1988	0.2717	0.1677
PR-3-s	1.5319	1.4829	1.8057	2.8256
	0.1831	0.1107	0.2203	0.1591
SSGPOR	2.2571	2.4279	2.9460	3.8621
	0.3923	0.3726	1.0730	0.3813

The first line for each method records the mean of the rRMSE over 21 time intervals in a year and the second one is for the standard deviation

TABLE 3
Comparison of the rMAE among the Six Methods for Years 4 to 7

Method	4th Year	5th Year	6th Year	7th Year
PR	1.4520	1.4322	1.7589	2.0029
	0.3302	0.3502	0.3232	0.4021
PR-2-p	1.0635	0.9497	1.0051	1.0902
	0.2412	0.2043	0.1443	0.2279
PR-2-s	1.0511	1.0367	1.0932	1.2505
	0.2871	0.2087	0.1325	0.2698
PR-3-p	0.8024	0.7998	0.7956	0.7057
	0.1056	0.1627	0.1435	0.1789
PR-3-s	0.7997	0.8202	0.8476	0.8509
	0.1258	0.1401	0.1681	0.2014
SSGPOR	1.5419	1.5027	1.8813	2.1783
	0.3776	0.4695	0.3329	0.4905

The first line for each method records the mean of the rMAE over 21 time intervals in a year and the second one is for the standard deviation.

TABLE 4
Comparison of the rME among the Six Methods for Years 4 to 7

Method	4th Year	5th Year	6th Year	7th Year
PR	9.5447	9.1064	10.2107	14.4832
	0.7236	0.8212	1.5181	1.5806
PR-2-p	7.1757	7.4974	9.7589	11.4743
	0.8959	0.8520	0.7792	1.6310
PR-2-s	6.9698	7.3224	9.3292	11.7802
	0.7690	0.7961	0.6377	1.2723
PR-3-p	5.7640	5.5334	7.3502	9.2735
	1.2970	0.8964	0.9382	1.7077
PR-3-s	5.8707	4.6468	7.2367	9.4219
	1.1477	0.8381	0.7297	1.3784
SSGPOR	12.5642	11.5655	12.0074	16.7220
	2.2644	1.1325	2.9733	3.9954

The first line for each method records the mean of the rME over 21 time intervals in a year and the second one is for the standard deviation.

$3 \leq j \leq 23$. We find that $(\boldsymbol{\beta}^{(2)})^T \hat{z}_j^i$'s are all negative and $(\boldsymbol{\beta}^{(1)})^T \hat{z}_j^i$'s all positive, which implies that the peak value takes place around some time interval. So without placing constraints on all $\boldsymbol{\beta}^{(2)}$'s and $\boldsymbol{\beta}^{(1)}$'s, we can learn the expected functional form, which verifies that the distribution of the external factor matches our assumption.

5.5 Comparison between Diffusion Matrices

We are also interested in the comparison of the diffusion matrices learned by different models. We plot the adjacency matrix of the transportation network obtained from Google map and the diffusion matrices learned by our four models, which are trained on the whole dataset, in Fig. 9. In all the subfigures, a cell with a brighter color implies that the corresponding weight is larger. The transportation network with 62 nodes has 394 edges (including self-loop), implying the 89.75% $((62^2 - 394)/62^2 \times 100\%)$ sparsity. From Table 6, we can see the diffusion matrices learned by the PR-2-p and PR-3-p models have higher sparsity (i.e., 90.14 and 92.04 percent respectively) than that of the transportation network, which implies that some edges in the transportation network have zero weights in the learned diffusion matrices and hence the structures of the transportation network and the diffusion matrices are similar but not exactly identical. Table 6 records the proportion of the elements at different levels (i.e., equal to 0, smaller than 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2} and 10^{-1}) in the diffusion matrices learned by different models. It shows that the diffusion matrices learned from the PR-2-s and PR-3-s models have lower sparsity (i.e., 43.47 and 55.65 percent) corresponding to the proportion of the elements being exactly 0. Diffusion matrices with lower sparsity can have higher expressive power and this may be one reason why the PR-2-s and PR-3-s models are superior to the PR-2-p and PR-3-p models under some measure as shown in the previous section. Moreover, we find that most elements in the diffusion matrices learned by the PR-2-s and PR-3-s models are not very large and the proportion of elements taking large values is smaller than that in the diffusion matrices learned by the PR-2-p and PR-3-p models at some levels (e.g., 10^{-3}, 10^{-2} and 10^{-1}). This can be verified based on Fig. 9 where the figures corresponding to the PR-2-p and PR-3-p models have more cells with bright colors than those of the PR-2-s and PR-3-s models. This indicates that the complexity of the learned diffusion matrices in terms of the l_1 or the Frobenius norms by the PR-2-s and PR-3-s models is not very high even though their sparsity, which corresponding to the level 0 where the elements are exactly zero, is higher.

We compare the values of the corresponding elements in different diffusion matrices learned by the PR-2-p and PR-3-p models as well as the comparison between the PR-2-s and PR-3-s models respectively in Table 7. From Table 7, we can see that the models utilizing the external factor (i.e., the PR-3-p and PR-3-s models) are likely to have elements of smaller value in $\mathbf{W}$ than their counterparts (i.e., the PR-2-p and PR-2-s models) without utilizing the external factor. This observation verifies the effectiveness of the modeling of the external factor to some extent.

We also compare the structure of the diffusion matrices learned based on the sparsity assumption with that of the given transportation network. Specifically, we count the number of the corresponding elements in the diffusion matrix and the transportation network where both weight values are greater than different thresholds. The results are shown in Table 8. The number of mutual non-zero elements

TABLE 5
Performance Comparison between Different Strategies to Learn the Diffusion Matrix $\mathbf{W}$

RMSE

	win	tie	loss
Year 4	13	1	7
Year 5	9	2	10
Year 6	11	3	7
Year 7	5	2	14
Sum	38	8	38

MAE

	win	tie	loss
Year 4	11	3	7
Year 5	4	2	15
Year 6	5	4	12
Year 7	5	2	14
Sum	25	11	48

ME

	win	tie	loss
Year 4	6	13	2
Year 5	3	18	0
Year 6	4	16	1
Year 7	2	16	3
Sum	15	63	6

RMSE

	win	tie	loss
Year 4	10	2	9
Year 5	12	2	7
Year 6	13	1	7
Year 7	3	7	11
Sum	38	12	34

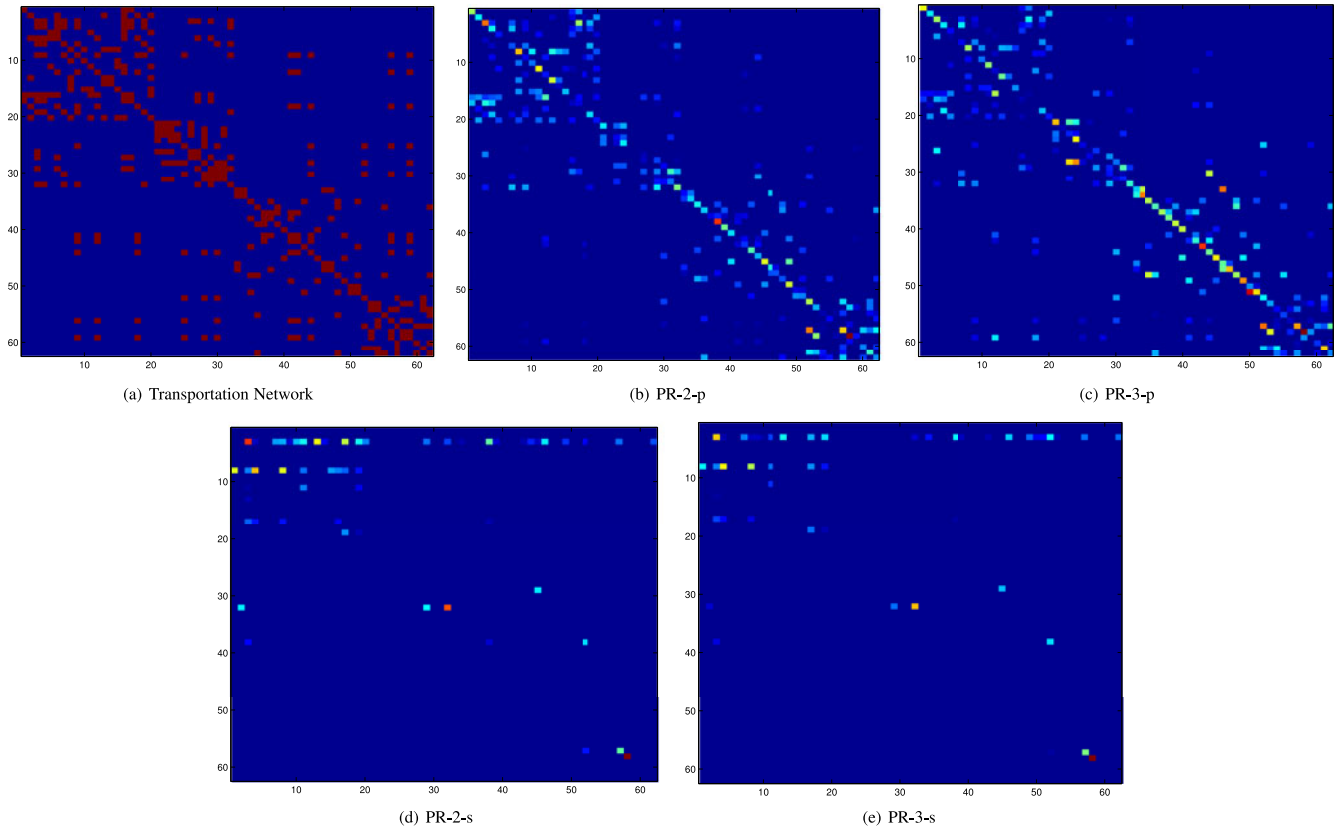
MAE

	win	tie	loss
Year 4	10	1	10
Year 5	9	4	8
Year 6	5	4	12
Year 7	3	10	8
Sum	27	19	38

ME

	win	tie	loss
Year 4	5	11	5
Year 5	8	13	0
Year 6	8	10	3
Year 7	4	12	5
Sum	25	46	13

The three tables in the first row record the win/tie/loss results between the PR-2-p and the PR-2-s models in terms of the RMSE, MAE, and ME, respectively. The three tables in the second row record the win/tie/loss results between the PR-3-p and the PR-3-s models in terms of the RMSE, MAE, and ME, respectively.



(a) Transportation Network (b) PR-2-p (c) PR-3-p

(d) PR-2-s (e) PR-3-s

Fig. 9. The given transportation network and the diffusion matrices learned by different models.

is not very small for the PR-2-s and PR-3-s models, which to some extent verifies the rationality of the diffusion matrices learned by both two models. Moreover, the number of the mutual elements with large values is not very large, which may suggest that the structure of the given transportation network contains some redundancy.

5.6 Convergence

We conduct experiments to demonstrate the ADMM method is a good choice for the proposed models.

The optimization methods we compared with are the ISTA method [4], a first-order gradient method, and the FISTA method [4] which is an accelerated gradient method. The algorithmic procedures for the ISTA and FISTA methods are described in Algorithms 2 and 3. The objective function that the ISTA and FISTA algorithms minimize takes a form as $F(\mathbf{\Theta}) = f(\mathbf{\Theta}) + g(\mathbf{\Theta})$ where $\mathbf{\Theta}$ denotes a set of model parameters and $f(\mathbf{\Theta})$ and $g(\mathbf{\Theta})$ are convex in terms of $\mathbf{\Theta}$. Both methods assume that the gradient of $f(\mathbf{\Theta})$ has Lipschitz continuity and $g(\mathbf{\Theta})$ has a simple and decomposable structure. The main idea of the ISTA and FISTA methods is first to find a surrogate function $Q_l(\mathbf{\Theta}, \hat{\mathbf{\Theta}})$ as

$$Q_l(\mathbf{\Theta}, \hat{\mathbf{\Theta}}) = g(\mathbf{\Theta}) + f(\hat{\mathbf{\Theta}}) + (\mathbf{\Theta} - \hat{\mathbf{\Theta}})^T \bigtriangledown_{\mathbf{\Theta}} f(\hat{\mathbf{\Theta}}) + \frac{l}{2}\mathcal{D}(\mathbf{\Theta}, \hat{\mathbf{\Theta}})^2,$$

where $\mathcal{D}(\mathbf{\Theta}, \hat{\mathbf{\Theta}})^2$ denotes the sum of squared Euclidean distances between each corresponding part in $\mathbf{\Theta}$ and $\hat{\mathbf{\Theta}}$, and $\bigtriangledown_{\mathbf{\Theta}} f(\hat{\mathbf{\Theta}})$ denotes the derivative of $f(\mathbf{\Theta})$ with respect to $\mathbf{\Theta}$ at $\mathbf{\Theta} = \hat{\mathbf{\Theta}}$, for $F(\mathbf{\Theta})$ based on the current solution $\hat{\mathbf{\Theta}}$ and then to optimize $Q_l(\mathbf{\Theta}, \hat{\mathbf{\Theta}})$ with respect to $\mathbf{\Theta}$ with the minimizor denoted by $q_l(\hat{\mathbf{\Theta}})$. According to [4], we have

$$F(\mathbf{\Theta}_k) - F(\mathbf{\Theta}^\star) \leq \frac{\eta L(f)\|\mathbf{\Theta}_0 - \mathbf{\Theta}^\star\|^2}{2k},$$

for the ISTA method, and

$$F(\mathbf{\Theta}_k) - F(\mathbf{\Theta}^\star) \leq \frac{2\eta L(f)\|\mathbf{\Theta}_0 - \mathbf{\Theta}^\star\|^2}{(k+1)^2},$$

TABLE 6
The Proportion of the Elements at Different Levels in the Diffusion Matrices Learned by Different Models

Level	PR-2-p	PR-2-s	PR-3-p	PR-3-s
$= 0$	90.14%	43.47%	92.04%	55.65%
$< 10^{-5}$	92.20%	57.60%	93.03%	63.94%
$< 10^{-4}$	92.20%	90.27%	93.08%	91.02%
$< 10^{-3}$	92.25%	98.36%	93.11%	98.62%
$< 10^{-2}$	92.79%	98.57%	93.24%	98.73%
$< 10^{-1}$	95.42%	98.99%	94.59%	99.22%

TABLE 7
The Proportion of the Elements in Diffusion Matrices $\mathbf{W}$ of the PR-3-p and PR-3-s Models Which are Larger than, or Equal to, or Smaller than the Corresponding Ones in the PR-2-p and PR-2-s Models, Respectively

	PR-2-p vs. PR-3-p	PR-2-s vs. PR-3-s
Larger	4.21%	44.72%
Equal	89.95%	27.03%
Smaller	5.93%	28.25%

TABLE 8
The Number of the Mutual Elements at Different Levels Which Exist in Both the Adjacency Matrix of the Transportation Network and the Diffusion Matrices Learned by the PR-2-s and PR-3-s Models, Respectively

Level	PR-2-s	PR-3-s
> 0	227	168
$> 10^{-5}$	178	138
$> 10^{-4}$	38	53
$> 10^{-3}$	23	19
$> 10^{-2}$	21	18
$> 10^{-1}$	16	12

for the FISTA method. Here $\mathbf{\Theta}^{\star}$ denotes the optimal solution and $L(f)$ denotes a Lipschitz constant of the gradient of $f(\cdot)$. From the above results, we can see theoretically the order of the convergence rate for the ADMM algorithm is the same as that of the ISTA algorithm but inferior to that of the FISTA algorithm.

Algorithm 2. The ISTA Algorithm

Initialize l_0, $\eta > 1$, and $\mathbf{\Theta}_0$;
$k := 1$;
while not converged **do**
  Find the smallest nonnegative integers i_k such that $F(q_{\hat{l}}(\mathbf{\Theta}_{k-1})) \leq Q_{\hat{l}}(q_{\hat{l}}(\mathbf{\Theta}_{k-1}), \mathbf{\Theta}_{k-1})$ where $\hat{l} = \eta^{i_k} l_{k-1}$;
  Set $l_k = \eta^{i_k} l_{k-1}$ and compute $\mathbf{\Theta}_k = q_{l_k}(\mathbf{\Theta}_{k-1})$;
  $k := k + 1$;
end while

To apply the ISTA and FISTA methods to our models, for all the proposed models we define $f(\mathbf{\Theta})$ as $f(\mathbf{\Theta}) = \sum_{i,j,k} \mu_{jk}^{i}$ and use $g(\mathbf{\Theta})$ to denote the rest parts in the objective function. Here $g(\mathbf{\Theta})$, a linear or quadratic function with respect to all model parameters, has a simple and decomposable structure, which satisfies the requirement for the function $g(\mathbf{\Theta})$. We find that the ISTA and FISTA methods are also suitable for solving the proposed objective functions with high-dimensional model parameters and constraints on the model parameters since in each step the subproblem has a close-form solution which requires only simple operations such as the addition and substraction.

Algorithm 3. The FISTA Algorithm

Initialize l_0, $\eta > 1$, and $\mathbf{\Theta}_0$;
$\mathbf{\Lambda}_1 := \mathbf{\Theta}_0$;
$k := 1$;
$t_1 := 1$;
while not converged **do**
  Find the smallest nonnegative integers i_k such that with $\hat{l} = \eta^{i_k} l_{k-1}$, $F(q_{\hat{l}}(\mathbf{\Lambda}_k)) \leq Q_{\hat{l}}(q_{\hat{l}}(\mathbf{\Lambda}_k), \mathbf{\Lambda}_k)$;
  $l_k := \eta^{i_k} l_{k-1}$;
  $\mathbf{\Theta}_k := q_{l_k}(\mathbf{\Lambda}_k)$;
  $t_{k+1} := \frac{1+\sqrt{1+4t_k^2}}{2}$;
  $\mathbf{\Lambda}_{k+1} := \mathbf{\Theta}_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{\Theta}_k - \mathbf{\Theta}_{k-1})$
  $k := k + 1$;
end while

We use all of the data for training to test the performance of the ISTA, FISTA, and ADMM algorithms by solving the PR, PR-2-p, PR-2-s, PR-3-p, and PR-3-s models. The results are recorded in Fig. 10. From the results, the FISTA algorithm converges faster than the ISTA algorithm which matches the theoretical analysis in [4]. Moreover, the ADMM algorithm has a much faster convergence than that of the ISTA and FISTA algorithms. This observation seems conflicting with the theoretical analysis. One reason for this is that the Lipschitz constant $L(f)$ in our model is very large (i.e., 10^8) due to the exponential function defined on the high-dimensional model parameters, making the convergence of the ISTA and FISTA not as fast as shown by the theoretical results.

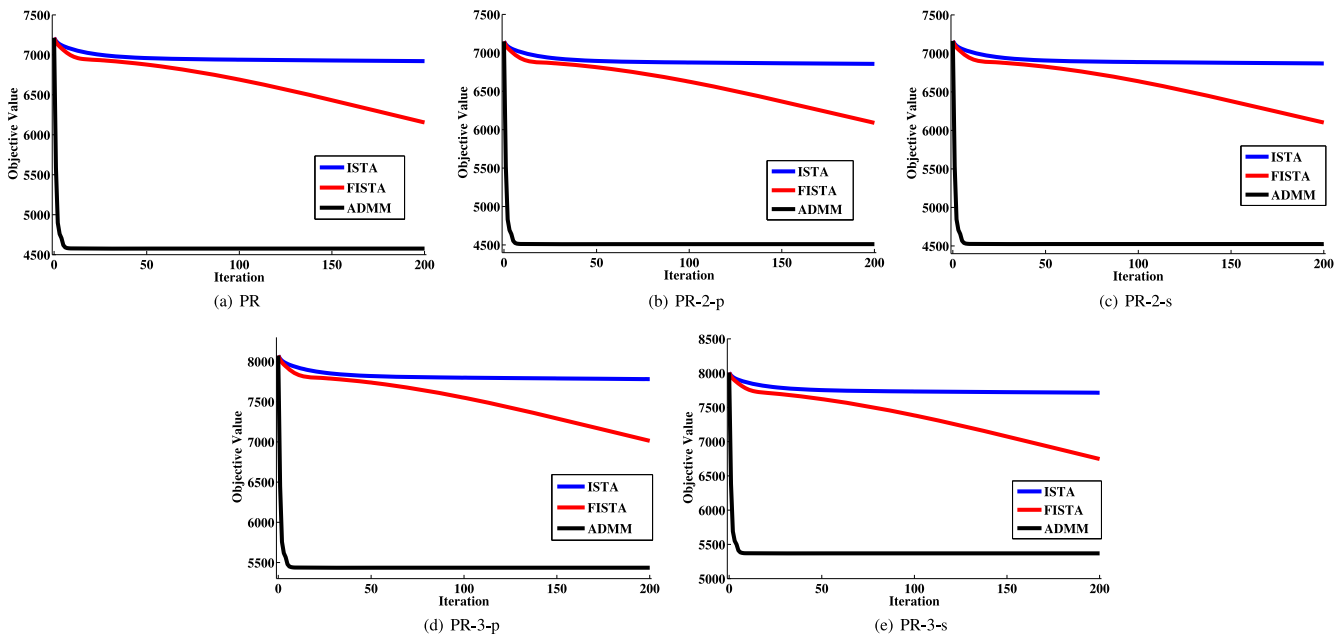


Fig. 10. Comparison on the convergence of the ISTA, FISTA, and ADMM algorithms in terms of the objective function values for different models.

TABLE 9
The Running Time of Different Methods

Method	Running Time (Second)
PR	0.3760
SSGPOR	20.5120
PR-2-p	1.6230
PR-2-s	1.4157
PR-3-p	2.2334
PR-3-s	2.2760

5.7 Analysis on Running Time

In this section, we analyze the running time of the proposed models. We use the Matlab R2011b as the testing environment and run the different models on a ThinkPad notebook with Intel i7 CPU and 8 GB RAM. By using all the data as the training data, the average running time over 100 repetitions for different models is shown in Table 9. From the results, we can see that by incorporating more factors, the running time becomes slightly longer, which matches our intuition since more model parameters are needed to learn. The running time of our proposed models is very short (less than 3 seconds for all the cases we tested), implying that the learning process is very efficient. Moreover, the SSGPOR model requires a much longer running time since it needs to do a costly matrix inverse in each iteration.

6 Conclusions

In this paper, we investigate how to combine three types of factors, the intra-regional, inter-regional and external factors, for epidemic prediction in a unified framework based on Poisson regression. The intra-regional factor is modeled by a linear function of region-specific features, the inter-regional one is modeled by a diffusion matrix learned from the data, and the effect of the external factor can be approximated by a parametric function (e.g., a quadratic function) of the time information.

Usually the epidemic diffusion pattern changes over years, making the model for epidemic prediction differ in different years. One extension of our work is to learn year-specific models for epidemic prediction. Moreover, the Poisson regression model has the overdispersion problem in which the expected prediction equals its variance. This problem seems more restrictive for some applications and we are interested in employing some extension of the Poisson regression model such as the negative-binomial regression model [13] for epidemic prediction based on the combination of those three factors.

Acknowledgments

This work is supported by Natural Science Foundation of China under Grant No. 61305071 and General Research Fund 211212 from the Research Grants Council of Hong Kong. The authors would like to thank Prof. Xiao-Nong Zhou of the National Institute of Parasitic Diseases (NIPD), Chinese Center for Disease Control and Prevention, for research collaborations.

References

[1] R. Abellana, C. Ascaso, J. Aponte, F. Saute, D. Nhalungo, A. Nhacolo, and P. Alonso, "Spatio-seasonal modeling of the incidence rate of malaria in mozambique," *Malaria J.*, vol. 7, p. 228, 2008.

[2] J. A. Achcar, E. Z. Martinez, A. D. Souza, V. M. Tachibana, and E. F. Flores, "Use of poisson spatiotemporal regression models for the brazilian amazon forest: Malaria count data," *Revista da Sociedade Brasileira de Medicina Tropical*, vol. 44, no. 6, pp. 749–754, 2011.

[3] R. M. Anderson, R. M. May, and B. Anderson, "*Infectious Diseases of Humans: Dynamics and Control*. New York, NY, USA: Oxford Univ. Press, 1992.

[4] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[5] P. J. Birrell, G. Ketsetzis, N. J. Gay, B. S. Cooper, A. M. Presanis, R. J. Harris, A. Charlett, X. S. Zhang, P. J. White, R. G. Pebody, and D. D. Angelis, "Bayesian modeling to unmask and predict influenza A/H1N1 PDM dynamics in London," *Proc. Nat. Acad. Sci. United States Am.*, vol. 108, no. 45, pp. 18238–18243, 2011.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[7] F. Brauer and C. Castillo-Chavez, *Mathematical Models in Population Biology and Epidemiology*. New York, NY, USA: Springer, 2001.

[8] A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[9] B. D. Dalziel, B. Pourbohloul, and S. P. Ellner, "Human mobility patterns predict divergent epidemic dynamics among cities," in *Proc. Roy. Soc. B*, vol. 280, p. 20130763, 2013.

[10] L. Denoeud, C. Turbelin, S. Ansart, A.-J. Valleron, A. Flahault, and F. Carrat, "Predicting pneumonia and influenza mortality from morbidity data," *PLoS ONE*, vol. 2, no. 5, p. e464, 2007.

[11] C.-S. Foo, C. B. Do, and A. Y. Ng, "A majorization-minimization algorithm for (multiple) hyperparameter learning," in *Proc. 26th Int. Conf. Mach. Learn.*, Montreal, Quebec, Canada, 2009, pp. 321–328.

[12] B. S. He and X. M. Yuan, "On the $o(1/t)$ convergence rate of alternating direction method," *Working paper*, 2011.

[13] J. M. Hilbe, *Negative Binomial Regression*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[14] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The Am. Statist.*, vol. 58, no. 1, pp. 30–37, 2004.

[15] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *J. Comput. Graphical Statist.*, vol. 9, no. 1, pp. 1–59, 2000.

[16] M. Lipsitch, T. Cohen, B. Cooper, J. M. Robins, S. Ma, L. James, G. Gopalakrishna, S. K. Chew, C. C. Tan, M. H. Samore, D. Fisman, and M. Murray, "Transmission dynamics and control of severe acute respiratory syndrome," *Science*, vol. 300, no. 5627, pp. 1966–1970, 2003.

[17] H. Nkurunziza, A. Gebhardt, and J. Pilz, "Bayesian modelling of the effect of climate on malaria in Burundi," *Malaria J.*, vol. 9, p. 114, 2010.

[18] S. Sharmin and M. I. Rayhan, "A stochastic model for early identification of infectious disease epidemics with application to measles cases in Bangladesh," *Asia-Pacific J. Public Health*, vol. 27, no. 2, pp. NP816–NP823, 2015.

[19] P. K. Srijith, S. K. Shevade, and S. Sundararajan, "Semi-supervised gaussian process ordinal regression," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Prague, Czech Republic, 2013, pp. 144–159.

[20] H. D. Teklehaimanot, J. Schwartz, A. Teklehaimanot, and M. Lipsitch, "Weather-based prediction of plasmodium falciparum malaria in epidemic-prone regions of Ethiopia II. weather-based prediction systems perform comparably to early detection systems in identifying times for interventions," *Malaria J.*, vol. 4, Article 44, 2004.

[21] R. E. Woodruff, C. S. Guest, M. G. Garner, N. Becker, J. Lindesay, T. Carvan, and K. Ebi, "Predicting ross river virus epidemics from regional weather data," *Epidemiology*, vol. 13, no. 4, pp. 384–393, 2002.

[22] T. T. Wu and K. Lange, "The MM alternative to EM," *Statist. Sci.*, vol. 25, no. 4, pp. 492–505, 2010.

[23] Y. Xiao, S. Tang, Y. Zhou, R. J. Smith, J. Wu, and N. Wang, "Predicting the HIV/AIDS epidemic and measuring the effect of mobility in mainland China," *J. Theoretical Biol.*, vol. 317, pp. 271–285, 2013.
[24] H. Yuan, G. Chen, J. Wu, and H. Xiong, "Towards controlling virus propagation in information systems with point-to-group information sharing," *Decision Support Syst.*, vol. 48, no. 1, pp. 57–68, 2009.

Yu Zhang received the BSc and MEng degrees in the Department of Computer Science and Technology at Nanjing University in 2004 and 2007, respectively, and the PhD degree in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. He is a research assistant professor in the Department of Computer Science of Hong Kong Baptist University. His research interests mainly include machine learning and data mining, especially in multi-task learning, transfer learning, dimensionality reduction, metric learning, and semi-supervised learning. He is a reviewer for various journals such as *Journal of Machine Learning Research*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Neural Networks and Learning Systems*, and *Public Relations*, and a program committee member for several conferences including NIPS, ICML, IJCAI, UAI, SDM, and so on. He has received the best paper award in the 26th Conference on Uncertainty in Artificial Intelligence (UAI) 2010 and the best student paper award in the 2013 IEEE/WIC/ACM International Conference on Web Intelligence. He is a member of the IEEE.

William K. Cheung received the BSc and MPhil degrees in electronic engineering from the Chinese University of Hong Kong and the PhD degree in computer science in 1999 from the Hong Kong University of Science and Technology. He is an associate professor in the Department of Computer Science, Hong Kong Baptist University. He has served as the co-chair and a program committee member for a number of international conferences, as well as a guest editor of journals on areas including artificial intelligence, web intelligence, data mining, web services, and e-commerce technologies. He has been on the editorial board of the *IEEE Intelligent Informatics Bulletin* since 2002. His recent research interests include collaborative information filtering, social network analysis and mining, and data mining applications in healthcare. He is a member of the IEEE.

Jiming Liu received the MEng and PhD degrees from McGill University, Canada. He is a chair professor in computer science and an associate dean of the Faculty of Science (Research) at Hong Kong Baptist University. His current research focuses on data mining and data analytics, health informatics, computational epidemiology, complex systems, multi-agent computing, and collective intelligence. He has served as the editor-in-chief of *Brain Informatics* (Springer) and *Web Intelligence* (IOS), and an associate editor of *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Cybernetics, Big Data and Information Analytics* (AIMS), *Computational Intelligence* (Wiley), and *Neuroscience and Biomedical Engineering* (Bentham), among others. He is a fellow of the IEEE.