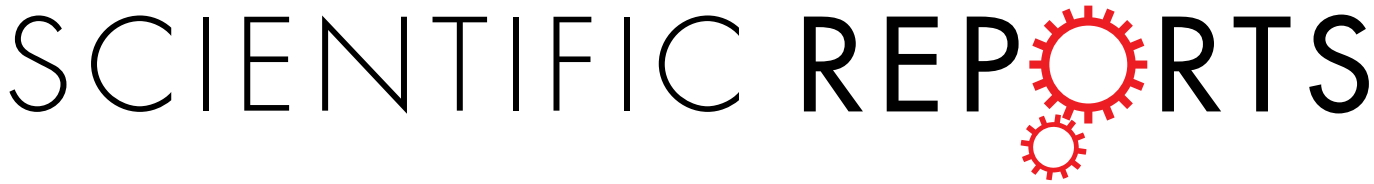
View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by HKU Scholars Hub

The HKU Scholars Hub The University of Hong Kong 香港大學學術庫

| | |
|---|---|
| Title | Analysis of heterogeneous dengue transmission in Guangdong in 2014 with multivariate time series model |
| Author(s) | Cheng, Q; Lu, X; Wu, JTK; Liu, Z; Huang, J |
| Citation | Scientific Reports, 2016, v. 6, p. 3375:1-9 |
| Issued Date | 2016 |
| URL | http://hdl.handle.net/10722/235702 |
| Rights | This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. |



OPEN

# Analysis of heterogeneous dengue transmission in Guangdong in 2014 with multivariate time series model

Received: 15 April 2016
Accepted: 02 September 2016
Published: 26 September 2016

Qing Cheng[1,2], Xin Lu[2,3,4,5], Joseph T. Wu[6], Zhong Liu[1,2] & Jincai Huang[1,2]

**Guangdong experienced the largest dengue epidemic in recent history. In 2014, the number of dengue cases was the highest in the previous 10 years and comprised more than 90% of all cases. In order to analyze heterogeneous transmission of dengue, a multivariate time series model decomposing dengue risk additively into endemic, autoregressive and spatiotemporal components was used to model dengue transmission. Moreover, random effects were introduced in the model to deal with heterogeneous dengue transmission and incidence levels and power law approach was embedded into the model to account for spatial interaction. There was little spatial variation in the autoregressive component. In contrast, for the endemic component, there was a pronounced heterogeneity between the Pearl River Delta area and the remaining districts. For the spatiotemporal component, there was considerable heterogeneity across districts with highest values in some western and eastern department. The results showed that the patterns driving dengue transmission were found by using clustering analysis. And endemic component contribution seems to be important in the Pearl River Delta area, where the incidence is high (95 per 100,000), while areas with relatively low incidence (4 per 100,000) are highly dependent on spatiotemporal spread and local autoregression.**

Dengue fever has spread rapidly within countries and across regions in the past few decades, resulting in an increased frequency of epidemics and severe dengue disease, hyperendemicity of multiple dengue virus serotypes in many tropical countries, and autochthonous transmission in Europe and the USA. Today, dengue is regarded as the most prevalent and rapidly spreading mosquito-borne viral disease among human beings[1]. Prior to 1970, only nine countries experienced dengue epidemics; however, the disease is now endemic in more than 120 countries in Africa, America, the Eastern Mediterranean, Southeast Asia and the Western Pacific[1]. The incidence of dengue has increased 30-fold in the past 50 years, and the geographic range of the virus and its vectors has expanded[2], with a recent study estimating that there are now 390 million (95% credible interval 284–528) dengue infections per year, of which 96 million (67–136) manifest apparently (any level of disease severity)[3].

In mainland China, the first outbreak of dengue occurred in Guangdong Province in 1978. Since then, dengue outbreaks have been recorded sequentially in Hainan, Guangxi, Fujian and Zhejiang provinces[4]. From 1990 to 2014, 69,321 cases of dengue including 11 deaths were reported in mainland China, equating to 2.2 cases per one million residents. The highest number was recorded in 2014 (47,056 cases). The number of provinces affected has increased, from a median of three provinces per year (range: 1 to 5 provinces) between 1990 and 2000 to a median of 14.5 provinces per year (range: 5 to 26 provinces) in the period 2001–2014[5]. Guangdong province has had the highest incidence of dengue in China (about 94.3% was reported in Guangdong from 2006 to 2014)[6,7]. dengue fever is a mosquito-borne viral disease with a strong potential for spatial variation[3,8] and varying transmission[9]. There are a number of reasons why incidences and transmission of dengue vary in time and space[10]. Dengue transmission is highly dependent on environmental factors and human movement. Environmental factors, such as temperature, rainfall and relative humidity, play a significant role in the transmission as well[11–14]. Limited

[1]Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, 410073 Changsha, China. [2]College of Information System and Management, National University of Defense Technology, 410073 Changsha, China. [3]Flowminder Foundation, 17177 Stockholm, Sweden. [4]Department of Public Health Sciences, Karolinska Institutet, 17177 Stock-holm, Sweden. [5]Division of Infectious Disease, Key Laboratory of Surveillance and Early-Warning on Infectious Disease, Chinese Centre for Disease Control and Prevention, Beijing 102206, P. R. China. [6]School of Public Health, Li Kashing Faculty of Medicine, Hong Kong University, Hong Kong Special Administrative Region, China. Correspondence and requests for materials should be addressed to Q.C. (email: sgggps@163.com)

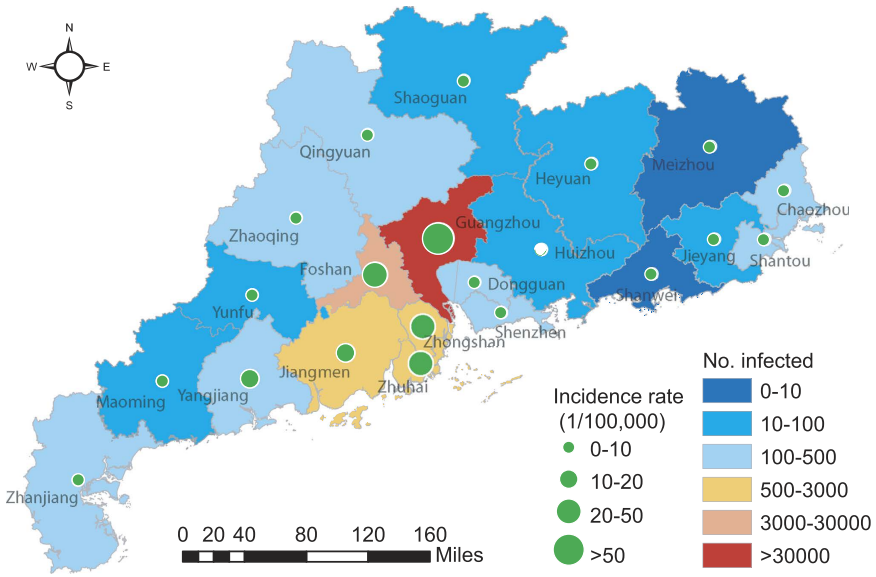



**Figure 1. Dengue map at the district level in Guangdong, China, 2014 (created with ArcGis Professional software version 10.2, http://www.esri.com/).**

dispersal distance of the dengue viruses[15] and daytime biting[16] imply that human movement should be the primary means by which the viruses spread spatially[17,18]. However, human movement and their spatial interaction change over space and time, as individuals vary considerably in the frequency, distance and nature of their movements[9,19]. In addition, the heterogeneous incidence levels might be influenced by other unobserved heterogeneity, such as underreporting and cultural differences at geographical scales. To implement efficient control measures, it is crucial to understand the heterogeneous incidence levels and varying transmission of dengue underlying this heterogeneity.

There was a great increase in the incidence of dengue fever in Guangdong Province in 2014. In 2014, the number of dengue fever cases in Guangdong reached a historically high level and exceeded the total number of cases over the previous 10 years. In this study, we perform an analysis of district-level time series of dengue transmission in Guangdong Province in 2014 using a multivariate time series model[20,21], which decomposes dengue risk additively into autoregressive, spatiotemporal and endemic components. The autoregressive and spatiotemporal components represent an autoregression on past counts in the same and in other districts, respectively, and should capture occasional outbreaks and dependencies across regions. The endemic component will describe the background risk of new events by external factors (independent of the history of the epidemic), which in the context of dengue may include seasonality/climate, population, immigration and sociodemographic variables. To account for heterogeneous incidence levels and varying transmission of dengue across districts, region-specific and possibly spatially correlated random effects are introduced in the model[22,23]. We identify the potential degree of heterogeneity on a district level through estimated random effect parameters and perform clustering analyses of model fitted value to uncover patterns driving dengue transmission. Understanding the characterization of these patterns might assist in the development of dengue control and prevention strategies in the province.

## Results

A total of 45,171 dengue cases were reported from 21 districts of Guangdong Province in 2014. The provincial capital, Guangzhou, has 37,394 cases (82.8% of all cases), followed by Foshan (7.8%) and Zhongshan (1.5%) (see Fig. 1). The incidence varies regionally with the highest incidence concentrated in the central districts of Guangdong (see Fig. 1). The majority of cases (42,538, 94.2%) were reported in September and October (weeks 35 to 44). The number of cases peaked in the $40^{th}$ week (9698 cases, about 21.5% of all cases), then decreased towards the end of the year, which indicates a clear seasonal pattern. In addition, the cumulative number of cases grows, similarly to a logistic growth curve (Fig. 2).

In this paper, a multivariate time series model is applied to data on weekly counts of dengue in 21 districts of Guangdong in 2014. In this model, the degree of heterogeneity in real situations is quantified by estimating the random effect value. For model selection in time series models, the comparison of successive one-step-ahead predictions with the actually observed data is used (here, the predictive quality of the models is assessed through one-step-ahead predictions of the last 8 weeks), i.e. fitted value is in turn used as initial value for model update and for calculating all subsequent predictions. In order to capture the effects of significant seasonality, we consider models that differ depending on seasonality parameters (let seasonal term $S = 1, 2, 3$, details shown in the Method section). Meanwhile, to describe dengue spread in space, we consider two different models (first order neighborhood (Fo) model and power law (PL) model) accounting for spatial interaction between districts. The Fo model assumes that an epidemic can only arrive from directly adjacent districts, and that all districts have the same coefficient for importing cases form neighboring districts. The PL model is considered as a description of

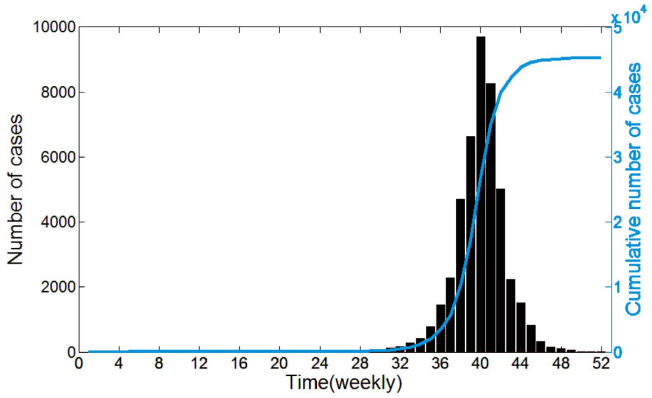



**Figure 2. Time series of weekly dengue cases reported, 2014.** The blue curve is the cumulative number of dengue cases.

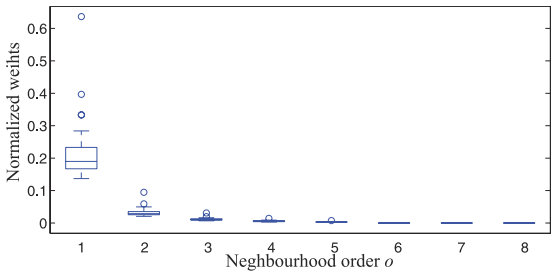



**Figure 3. Normalized weights in the multivariate time series model with "PL + pop." weight.**

| Model | $\widehat{\alpha^{(\nu)}}$ (se) | $\widehat{\alpha^{(\lambda)}}$ (se) | $\widehat{\alpha^{(\phi)}}$ (se) | $\widehat{\sigma_\nu^2}$ | $\widehat{\sigma_\lambda^2}$ | $\widehat{\sigma_\phi^2}$ | $\widehat{\psi}$ (se) | $\widehat{d}$ (se) |
|---|---|---|---|---|---|---|---|---|
| D3 ("PL + pop.", S = 3) | −2.934 (0.811) | −0.899 (0.101) | 1.769 (3.324) | 8.166 | 0.022 | 3.331 | 0.130 (0.022) | 2.745 (0.672) |

**Table 1. Estimated model parameters (with standard errors).**

spatial interaction as motivated by human travel behavior can be well described by a decreasing power law of the distance or neighborhood order, which assumes the form $w_{ji} = o_{ji}^{-d}$, for $j \neq i$ and $w_{jj} = 0$, where $w_{ji}$ is the weight that describe the strength of transmission form district $j$ to district $i$, $o_{ji}$ is the order of neighborhood, and $d$ is the decay parameter (the details of both models are described in Method section). According to Table S1 (Supplementary Table S1), all models including PL perform better than models including Fo with respect to all scores. Further improvement of the PL model's description of human mobility can be achieved by accounting for the district-specific population in the spatiotemporal component, i.e. "PL + pop." model. Table S1 shows that the "PL + pop." model with seasonal term S = 3 (denoted as Model D3) outperforms all other models. Therefore, we apply model D3 with random effects to account for district heterogeneity. Parameter estimates for D3 are shown in Table 1.

From the Table 1, the decay parameter estimate is $\widehat{d} = 2.745$ (0.672), which represents a strong decay of spatial interaction for higher-order neighbors because the higher the decay parameter $d$, the less important are higher-order neighbors. Moreover, Fig. 3 shows neighborhood weights $w_{ij}$ against neighborhood order $o_{ij}$, it is obvious that the spatiotemporal component effects mainly account for nearest neighbors dependence.

Note that the heterogeneity of the dengue incidence can thus be captured adequately according to random effect parameter $\sigma_\lambda^2$, $\sigma_\phi^2$, $\sigma_\nu^2$ estimated. Little variation in the autoregressive component among districts is found since the variance $\sigma_\lambda^2 = 0.022$ is estimated to be quite small. In contrast, there is considerable spatial variation concerning the endemic coefficient with $\sigma_\nu^2 = 8.166$ and spatiotemporal coefficient with $\sigma_\phi^2 = 3.331$. We thus believe that there is significant spatial heterogeneity in the endemic and spatiotemporal component and spatial homogeneity in the autoregressive component.

In particular, for the endemic component in Fig. 4(a), there is a pronounced heterogeneity between the Pearl River Delta area (Guangzhou, Foshan, Zhongshan, Jiangmen and Zhuhai, shown in pink) and the remaining districts. Similarly, there is considerable heterogeneity across districts for the spatiotemporal component as shown in Fig. 4(b). However, no clear spatial pattern can be seen. Moreover, all districts exhibit a very low random effect value in the autoregressive component according to the bar chart in Fig. 4(c). There seems to be no significant

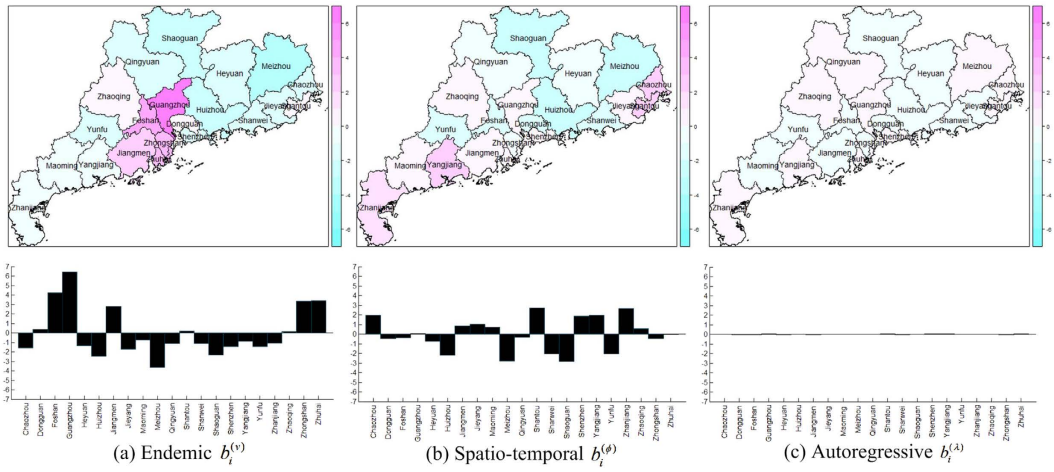



**Figure 4. Estimated district-specific random effects in the multivariate time series model.** There is considerable variation concerning the endemic coefficient and spatiotemporal coefficient, and there seems to be little variation in the autoregressive coefficient (created with R version 3.3.3, https://www.r-project.org/).

difference among districts for the random effects in the autoregressive component, thus we infer that the autoregressive effect accounting for dengue transmission in all districts might be homogeneous.

For each district, the relative contributions of endemic, autoregressive and spatiotemporal factors in driving the dengue prevalence with time is called the "patterns driving dengue transmission" in this district. An intuitive way of quantifying the relative contributions of the three components is provided by Fig. 5. It shows the fitted component means along with the observed time series for the 20 districts with at least one case. Figure 5 also demonstrates that dengue transmission appears to be synchronous in Guangdong, peaking at the same times of the year in different districts (between weeks 35 and 44). In order to further understand different aspects of dengue transmission patterns and their drivers, we focus on the outbreak period (weeks 35 to 44) and use the Fdp clustering method (see Method Section) to find patterns driving dengue transmission.

It is obvious that three cluster centers and four noises can be found in the decision graph, as shown in Fig. 6. Thus, we hypothesize that there are three patterns driving dengue transmission in Guangdong, denoted as Patterns A, B, C, respectively (see Fig. 7, red districts belong to Pattern A, green districts belong to Pattern B and blue districts belong to Pattern C). It is worthwhile noting that autoregressive component trends in all districts are similar because there no significant differences among districts for the random effects in the autoregressive component as shown in Fig. 4(c). According to the Fig. 7. It is obvious that the autoregressive's percentage increase during the dengue outbreak period, which means the local autoregression factor plays an increasingly important role in dengue progress in Guangdong. But the endemic and spatiotemporal components play different roles for different patterns.

Jiangmen, Zhongshan, Guangzhou and Foshan belong to Pattern A (they are in the Pearl River Delta area), have been mainly affected by the endemic component and also have the highest overall number of cases. Especially in the early outbreak period, more than 70% of cases account for endemic component which implies that the socioeconomic, climate and environment might be major factors attribute to dengue outbreak in these districts[24] because the endemic component describe the background risk of new events by external factors (independent of the history of the epidemic) in our model. In contrast, these districts are estimated to have a relatively low spatiotemporal contribution. Maoming, Dongguan, Zhaoqing and Shantou belong to Pattern B, are mainly influenced by endemic and spatiotemporal components in the early outbreak period, and then the endemic proportion declines gradually, but the spatiotemporal proportion always maintains a relatively high level, i.e. the factors accounting for dengue outbreak in these districts change from endemic and spatiotemporal components to autoregressive and spatiotemporal components. Chaozhou, Shaoguan, Huizhou, Jieyang, Heyuan, Shenzhen, Zhanjiang, Qingyuan and Yangjiang belong to Pattern C. In these districts, the incidence is clearly dominated by the spatiotemporal component, meaning that a great amount of cases is explained via transmission from neighboring districts.

The Fdp clustering method classifies districts independently of the geographic area taking into account only the patterns driving dengue transmission. The majority of cases are clustered in the Pearl River Delta area (with more than 93% of all cases) and the incidence in these districts is relatively high (incidence is about 95 per 100,000). These districts belong to Pattern A and their incidence is clearly dominated by the endemic and autoregressive component, while districts that belong to Patterns B and C with relatively low incidence (incidence is about 4 per 100,000) are highly dependent on spatiotemporal spread and local autoregression.

## Discussion

Our analysis of dengue case report data characterizes considerable heterogenous incidence levels and transmission across districts in the dengue outbreak in Guangdong Province in 2014. The degree of heterogeneity is quantified through random effect parameter estimates; moreover, by using the Fdp clustering method, we explore the

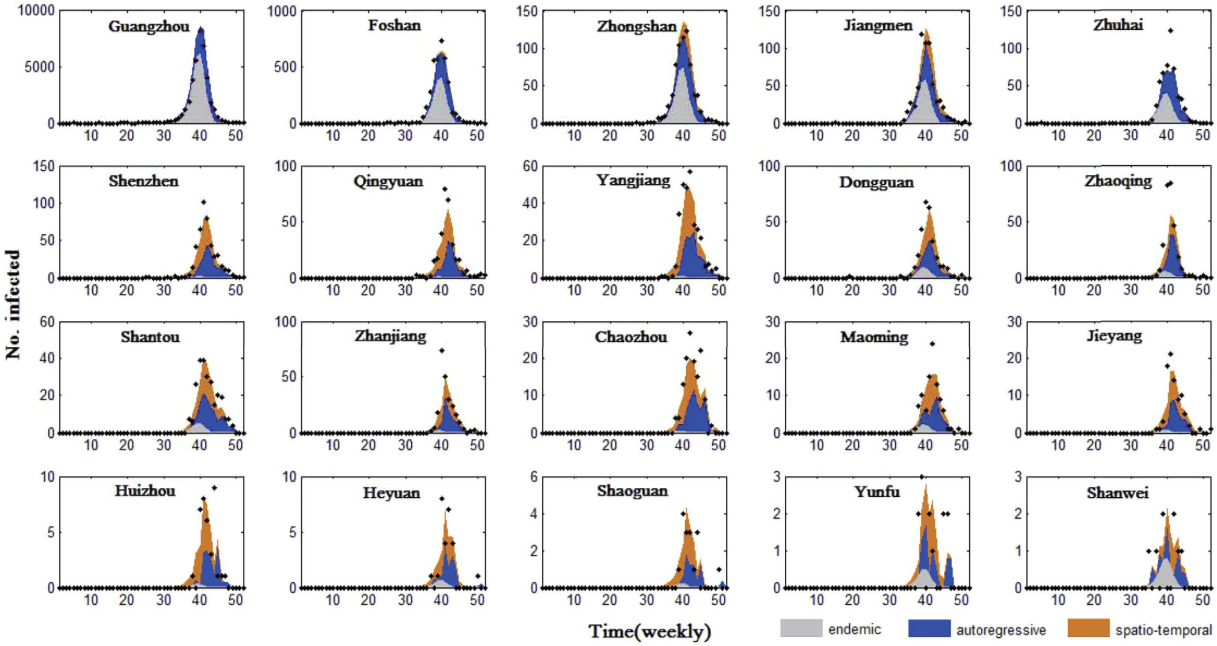



**Figure 5. Fitted components in the multivariate time series model for 20 districts with more than 0 cases.** Black dots are drawn for weekly counts, the light gray area shows the estimated endemic component, the blue area corresponds to the autoregressive contribution and the orange area corresponds to the spatiotemporal contribution.

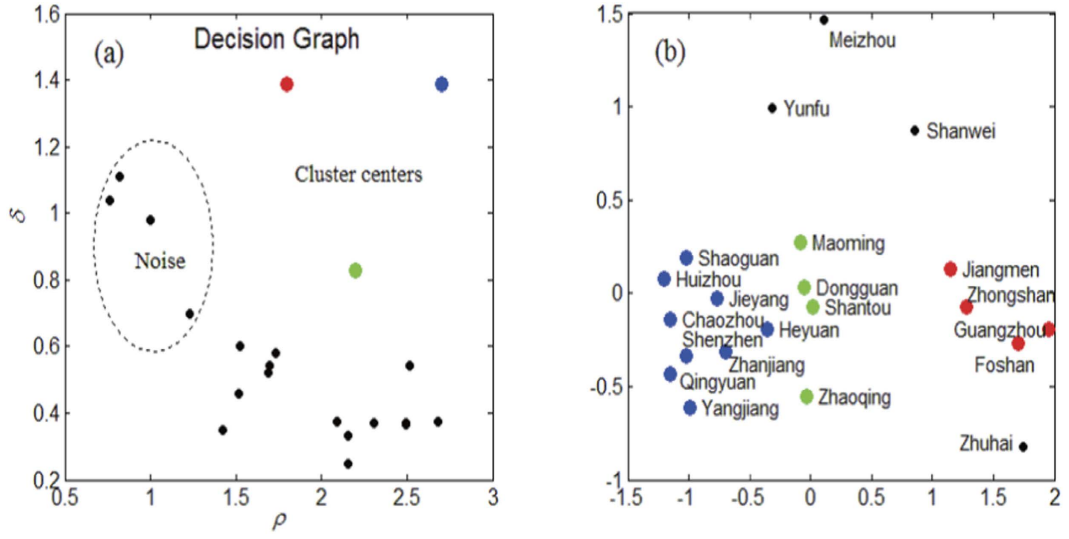



**Figure 6. Using Fdp clustering method to find spread patterns.** District-specific patterns driving dengue transmission regarded as points: (**a**) Decision graph for the district-specific pattern driving dengue transmission, three cluster centers and four noises are identified; (**b**) Point distribution, different colors correspond to different clusters.

characterization of patterns driving dengue transmission, which might be useful for improving our understanding of heterogeneous dengue transmission.

To analyze the spatial and temporal occurrence of dengue and its association with the heterogeneity of environmental characteristics, we fit a multivariate time series model of dengue virus transmission to spatial time series data from Guangdong and compare maximum-likelihood random effect estimates to account for unobserved heterogeneity. To assess the potential for dengue spread in space, both models, the first-order neighbor model and the power law model, accounting for spatial interaction between districts, have been incorporated in the multivariate time series model. We find that the power law model accounting for spatial interaction between

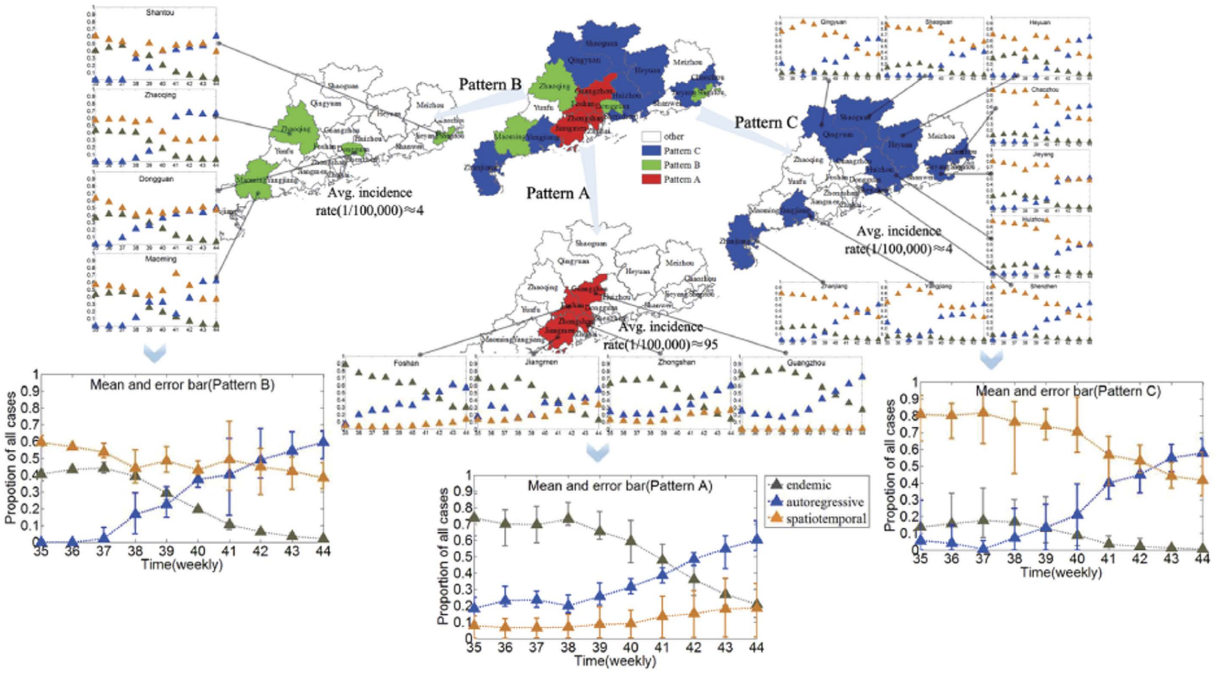



**Figure 7. Three patterns and their characteristics.** The autoregressive component trends are similar, and their percentages increase in all districts during the dengue outbreak period. In addition, red districts (Jiangmen, Zhongshan, Guangzhou and Foshan, Pattern A) have been mainly affected by the endemic component and are also the ones with the highest overall number of cases. Green districts (Maoming, Dongguan, Zhaoqing and Shantou, Pattern B) are mainly influenced by endemic and spatiotemporal components in the early period, and then the endemic proportion declines gradually, but the spatiotemporal proportion always maintains a relatively high level. The spatiotemporal component of incidence is the most powerful in blue districts (Chaozhou, Shaoguan, Huizhou, Jieyang, Heyuan, Shenzhen, Zhanjiang, Qingyuan and Yangjiang, Pattern C), meaning that a great number of cases are explained via transmission from neighboring districts (created with ArcGis Professional software version 10.2, http://www.esri.com/).

districts substantially improves model fit and predictions. We further note that in the best-fitted model D3, there is a strong decay of spatial interaction for higher-order neighbors, which indicates that the spatiotemporal component effects mainly account for nearest neighbor dependence. But this may be limited by the small number of districts in Guangdong: A second-order neighboring district would include all. A better way of accounting for spatial interaction effects would thus be to explicitly incorporate movement network data[25,26]. For instance, mobile phone data have been used as proxy for human mobility to achieve improved predictive performance in disease spreading[27,28]. But the power law approach to modeling spatial interaction is especially attractive if movement network data are not available[22].

In this study, we show that the estimated random effect parameters were able to capture the influence of heterogeneity at district level. In particular, there is significant endemic and spatiotemporal variation across districts, while no clear autoregressive heterogeneity is found. Another encouraging finding is the relative importance of the three components in each district and three patterns driving dengue transmission in Guangdong. The Pearl River Delta area in pattern C is highly exhibit a relatively high endemic incidence, which means that a great amount of cases is explained by external factors, such as seasonality/climate, socioeconomic and environment. It seems that the spread of dengue in Pearl River Delta area is more of an endemic than epidemic nature[29]. Our analysis also shows that the districts in Pattern C are highly dependent on the spatiotemporal component and these districts are mostly around the Pearl River Delta area, meaning that a large number of cases in these districts are explained via transmission from neighboring districts, and the Pearl River Delta area is thought to act as a reservoir for the virus from where it can spread to the neighboring districts.

To conclude, this paper underlines the varying transmission of dengue across districts and characteristics of three patterns driving dengue transmission by combing multivariate time series model and clustering method. However, when analyzing spatially stratified time series the assumption of equal transmission rates or incidence levels across all districts is question. For example, dengue transmission might be influenced by factors such as vaccination status in individuals, vector, or environmental factors, Such factors could be incorporated into the multivariate time series model as covariates if they are observable and available. Therefore, it would be necessary to further enrich the model by entering external processes such as mosquito density as covariates in the endemic and epidemic components, and a better way of accounting for spatial interaction would thus be to explicitly incorporate human mobility data.

## Methods

**Study area.** Guangdong Province is in Southeast China and has a population of more than 100 million people. It had the highest incidence of dengue in mainland China in 2014, accounting for more than 90% of all cases[5]. Guangdong covers about 180,000 $Km^2$, with Guangzhou city as the provincial capital; it is one of the most densely urbanized regions in the world and one of the main hubs of China's economic growth, and it comprises 21 districts (see Fig. 1. In addition, Guangdong is warm and damp all year round with average temperatures ranging from 19 to 26 °C and with a rainy season from April to September[30].

**Data collection.** In this study, we obtained data on dengue fever cases in Guangdong in 2014 from the Chinese Center for Disease Control and Prevention (China CDC). The data were aggregated to weekly counts. In China, all cases of dengue were diagnosed according to the unified diagnostic criteria issued by the Chinese Ministry of Health. Population data for every district in Guangdong in 2014 were retrieved from the Guangdong Statistical Yearbook[31].

**Multivariate time series model.** The multivariate time series model established by Held and Paul[20,23] is designed for spatially and temporally aggregated surveillance data. Let $Y_{i,t}$ denote the number of cases of a specific disease in region $i = 1, \ldots, I$ at time $t = 1, \ldots, T$. The counts are assumed to be negatively binomially distributed with conditional mean

$$Y_{i,t} | Y_{i,t-1} \propto NegBin(u_{it}, \psi), \tag{1}$$

where $u_{it} = e_i\nu_{it} + \lambda_i Y_{i,t-1} + \phi_i \sum_{j \neq i} w_{ji} Y_{j,t-1}$ and $\Psi$ is an overdispersion parameter such that the conditional variance of $Y_{i,t}$ is $u_{it}(1 + \Psi u_{it})$. $e_i\nu_{it}$ is the endemic component and parametrically models seasonal variation and trends. The other two components are observation-driven epidemic components: An autoregressive component $\lambda_i Y_{i,t-1}$ on the number of cases at the previous time point, and a spatiotemporal component $\phi_i \sum_{j \neq i} w_{ji} Y_{j,t-1}$ capturing transmission from other units. Each of $\nu_{it}$, $\lambda_i$, $\phi_i$ is a log-linear predictor of the form

$$\log(\nu_{it}) = \alpha^{(\nu)} + b_i^{(\nu)} + \sum_{s=1}^{S} \{\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t)\}, \tag{2}$$

$$\log(\lambda_i) = \alpha^{(\lambda)} + b_i^{(\lambda)}, \tag{3}$$

$$\log(\phi_i) = \alpha^{(\phi)} + b_i^{(\phi)}, \tag{4}$$

where $\alpha^{(\nu)}$, $\alpha^{(\lambda)}$, $\alpha^{(\phi)}$, are intercepts, $b_i^{(\nu)}$, $b_i^{(\lambda)}$, $b_i^{(\phi)}$ are regional random effects which account for heterogeneity between districts, and are assumed to follow independently a normal distribution with zero mean and covariance matrix $\Sigma = \mathrm{diag}(\sigma_\nu^2 \mathrm{I}, \sigma_\lambda^2 \mathrm{I}, \sigma_\phi^2 \mathrm{I})$, where $\sigma_\lambda^2$, $\sigma_\phi^2$, $\sigma_\nu^2$, and I is the identity matrix.

The endemic log-linear predictor $\nu_{it}$ incorporates a sinusoidal wave of frequency ($\omega_s$ are Fourier frequencies, let $\omega_s = 2\pi/52$ for weekly data in this paper), and $S$ is the seasonal parameters. As a basic district-specific measure of disease incidence, the population fraction $e_i$ is included as a multiplicative offset.

The weights $w_{ji}$ of the spatiotemporal component describe the strength of transmission from region $j$ to region $i$, and the neighborhood-based approach to model spatial interaction is especially attractive if movement network data are not available. Thus we consider four different neighborhood-based approaches to measure neighborhood weights.

- The first-order neighborhood model (Fo) assumes that an epidemic can only arrive from directly adjacent districts 20, and that all districts have the same coefficient for importing cases from neighboring districts;
- To reflect commuter-driven spread in our model, we scale the district's susceptibility according to its population fraction by multiplying $\phi$ by $e_i^{\beta_{pop}}$ where $e_i$ is the population fraction and $\beta_{pop}$ is to be estimated (**Fo + pop**.)[20];
- To account for long-range case transmission, a power law model (**PL**) is suggested, which assumes the form $w_{ji} = o_{ji}^{-d}$, for $j \neq i$ and $w_{jj} = 0$, where $o_{ji}$ is the order of neighborhood, the decay parameter $d$ is to be estimated[32];
- Based on the power law model used to measure neighborhood weights, we scale the district's susceptibility according to its population fraction by multiplying $\phi$ by $e_i^{\beta_{pop}}$ where $e_i$ is the population fraction and $\beta_{pop}$ is to be estimated (**PL + pop**.).

The estimation of parameters involves integration of the likelihood with respect to the random effects which cannot be obtained analytically. Paul and Held[23] suggest a penalized likelihood approach for inference, where variance components are treated as known when estimating the fixed and random effects. The variance components themselves are estimated through maximizing the approximated marginal likelihood obtained via a Laplace approximation. However, classical model choice criteria such as Akaike's Information Criterion (AIC) cannot be used straightforwardly for models with random effects. Therefore, the performance of the power law models and the first-order formulations is compared by one-step-ahead forecasts assessed with strictly proper scoring rules: the logarithmic score (logS) and the ranked probability score (RPS)[33], and lower scores correspond to better predictions.

**Clustering method.** Based on the multivariate time series model, for each district, endemic, autoregressive and spatiotemporal components are different over time. For each district, the relative contributions of endemic, autoregressive and spatiotemporal factors in driving the dengue prevalence with time are called the "pattern driving dengue transmission". Let $end_{i,t}$, $ar_{i,t}$ and $ne_{i,t}$ denote endemic, autoregressive and spatiotemporal components accounting for the proportion of number of cases in district $i = 1, \ldots, I$ at time $t = 1, \ldots, T$ respectively, i.e. $end_{i,t} = e_i v_{it}/Y_{i,t}$, $ar_{i,t} = \lambda_i Y_{i,t-1}/Y_{i,t}$ and $ne_{i,t} = \phi_i \sum_{j \neq i} w_{ji} Y_{j,t-1}/Y_{j,t}$. Then the pattern driving dengue transmission of district $i$ from $t = s$ to $t = e$ can be denoted as $P_i = \{\{end_{i,t}\}, \{ar_{i,t}\}, \{ne_{i,t}\}: t \in [s, e]\}$. In fact, the pattern driving dengue transmission is a multivariate time series.

First, we define the distance between two patterns driving dengue transmission by using the Euclidean distance

$$d(P_i, P_j) = \sum_{t=s}^{e} (end_{i,t} - end_{j,t})^2 + \sum_{t=s}^{e} (ne_{i,t} - ne_{j,t})^2 + \sum_{t=s}^{e} (ar_{i,t} - ar_{j,t})^2. \quad (5)$$

Then, we revise the Fdp clustering method proposed by Alex Rodriguez *et al.*[34] as below: The pattern driving dengue transmission in each district is regarded as a data point. The local density $\rho_i$ of data point $i$ is defined as

$$\rho_i = 1/\sum_{j=1}^{k} d(i, Ne_j(i)), \quad (6)$$

where $d(i, j)$ represents the distance between data point $i$ and $j$, and $Ne_j(i)$ represents the $j^{\text{th}}$-nearest neighbor of data point $i$. In this paper, let $k = 3$.

The distance $\delta_i$ of data point $i$ is measured by computing the minimum distance between the data point $i$ and any other data point with higher density[34]:

$$\delta_i = \min_{j:\rho_j > \rho_i} (d(i, j)). \quad (7)$$

If the data point is with the highest density, we conventionally take $\delta_i = \max_j(d(i, j))$. Note that $\delta_i$ is much larger than the typical nearest neighbor distance only for points that are local or global maxima in the density. Thus, cluster centers are recognized as points for which the value of $\delta_i$ is anomalously large. After the cluster centers have been found, each remaining point is assigned to the same cluster as its nearest neighbor of higher density. Some points have a relatively high $\delta$ and a low $\rho$ because they are isolated; they can be considered clusters composed of a single point, namely noise.

## References


1. Geneva. Global strategy for dengue prevention and control. Tech. Rep., World Health Organization (2012).
2. Maria, G. & Guzman, E. H. Dengue. *The Lancet* **385,** 453–465 (2015).
3. Bhatt, S. *et al.* The global distribution and burden of dengue. *Nature* **496,** 504–507 (2013).
4. Wu, J.-Y., Lun, Z.-R., James, A. A. & Chen, X.-G. Review: Dengue fever in mainland china. *The American Journal of Tropical Medicine and Hygiene* **83,** 664–671 (2010).
5. Lai, S. *et al.* The changing epidemiology of dengue in china, 1990–2014: a descriptive analysis of 25 years of nationwide surveillance data. *BMC medicine* **13,** *100* (2015).
6. Sang, S. *et al.* Dengue is still an imported disease in china: A case study in guangzhou. *Infection, Genetics and Evolution* **32,** 178–190 (2015).
7. Wang, C. *et al.* Spatial and temporal patterns of dengue in guangdong province of china. *Asia-Pacific Journal of Public Health* 1010539513477681 (2013).
8. Brady, O. J. *et al.* Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Negl Trop Dis* **6,** e1760 (2012).
9. Perkins, T. A., Scott, T. W., Le Menach, A. & Smith, D. L. Heterogeneity, mixing, and the spatial scales of mosquito-borne pathogen transmission. *PLoS Comput Biol* **9,** e1003327 (2013).
10. Kraemer, M. U. G. *et al.* Big city, small world: density, contact rates, and transmission of dengue across pakistan. *Journal of The Royal Society Interface* **12** (2015).
11. Bai, L., Morton, L. C., Liu, Q. *et al.* Climate change and mosquito-borne diseases in china: a review. *Global Health* **9,** 1–22 (2013).
12. SHEN, J. C. *et al.* The impacts of mosquito density and meteorological factors on dengue fever epidemics in guangzhou, china, 2006–2014: a time-series analysis. *Biomedical and Environmental Sciences* **28,** 321–329 (2015).
13. Wongkoon, S., Jaroensutasinee, M. & Jaroensutasinee, K. Climatic variability and dengue virus transmission in chiang rai, thailand. *Biomedica* **27,** 5–13 (2011).
14. Wu, P.-C., Guo, H.-R., Lung, S.-C., Lin, C.-Y. & Su, H.-J. Weather as an effective predictor for occurrence of dengue fever in taiwan. *Acta Tropica* **103**, 50 – 57 (2007).
15. Harrington, L. C. *et al.* Dispersal of the dengue vector aedes aegypti within and between rural communities. *The American Journal of Tropical Medicine and Hygiene* **72,** 209–220 (2005).
16. Akram, W. *et al.* Seasonal distribution and species composition of daytime biting mosquitoes. *Entomological Research* **39,** 107–113 (2009).
17. Stoddard, S. T. *et al.* The role of human movement in the transmission of vector-borne pathogens. *PLoS Negl Trop Dis* **3,** e481 (2009).
18. Teurlai, M. *et al.* Can human movements explain heterogeneous propagation of dengue fever in cambodia? *PLoS Negl Trop Dis* **6,** e1957 (2012).
19. Reiner, R. C., Stoddard, S. T. & Scott, T. W. Socially structured human movement shapes dengue transmission despite the diffusive effect of mosquito dispersal. *Epidemics* **6**, 30–36 (2014).
20. Meyer, S., Held, L. & Höhle, M. Spatio-temporal analysis of epidemic phenomena using the r package surveillance. *arXiv preprint arXiv:1411.0416* (2014).
21. Paul, M., Held, L. & Toschke, A. M. Multivariate modelling of infectious disease surveillance data. *Statistics in medicine* **27,** 6250–6267 (2008).
22. Geilhufe, M., Held, L., Skrϕvseth, S. O., Simonsen, G. S. & Godtliebsen, F. Power law approximations of movement network data for modeling infectious disease spread. *Biometrical Journal* **56,** 363–382 (2014).

23. Paul, M. & Held, L. Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine* **30,** 1118–1136 (2011).
24. Qi, X. *et al.* The effects of socioeconomic and environmental factors on the incidence of dengue fever in the pearl river delta, china, 2013. *PLoS Negl Trop Dis* **9,** e0004159 (2015).
25. Wang, L., Wang, Z., Zhang, Y. & Li, X. How human location-specific contact patterns impact spatial transmission between populations? *Scientific Reports* **3,** 1468 (2013).
26. Marshall, J. M. *et al.* Key traveller groups of relevance to spatial malaria transmission: a survey of movement patterns in four sub-saharan african countries. *Malaria Journal* **15,** 1–12 (2016).
27. Bengtsson, L. *et al.* Using mobile phone data to predict the spatial spread of cholera. *Scientific Reports* **5,** 8923 (2015).
28. Wesolowski, A. *et al.* Impact of human mobility on the emergence of dengue epidemics in pakistan. *Proceedings of the National Academy of Sciences* **112,** 11887–11892 (2015).
29. Shen, S.-Q. *et al.* Multiple sources of infection and potential endemic characteristics of the large outbreak of dengue in guangdong in 2014. *Scientific Reports* **5,** 16913 (2015).
30. Li, Z. *et al.* Spatiotemporal analysis of indigenous and imported dengue fever cases in guangdong province, china. *BMC Infect Dis* 132–132 (2012).
31. Yeabook, G. S. http://www.gdstats.gov.cn/tjnj/2014/directory.html. Tech. Rep., Statistics Bureau of Guangdong Province (2014).
32. Meyer, S. & Held, L. Power-law models for infectious disease spread. *The Annals of Applied Statistics* **8,** 1612–1639 (2014).
33. Czado, C., Gneiting, T. & Held, L. Predictive model assessment for count data. *Biometrics* **65,** 1254–1261 (2009).
34. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks. *Science* **344,** 1492–1496 (2014).


## Acknowledgements

The authors would like to thank J.H., Yu, S. and J.L. from China CDC for helpful discussions. X.L. acknowledges the Natural Science Foundation of China under Grant Nos 71301165 and 71522014. Q.C. acknowledges the Hunan Provincial Innovation Foundation for Postgraduate under Grant No. CX2013B024.

## Author Contributions

Q.C. and X.L. contributed equally to this work. Conceived and designed the experiments: Q.C. and X.L. Performed the experiments: Q.C. and X.L. Analyzed the data: Q.C. and X.L. Contributed reagents/materials/analysis tools: Q.C., X.L., J.T.W., Z.L and J.H. Wrote the paper: Q.C., X.L., Z.L. and J.H.

## Additional Information

**Supplementary information** accompanies this paper at http://www.nature.com/srep

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article**: Cheng, Q. *et al.* Analysis of heterogeneous dengue transmission in Guangdong in 2014 with multivariate time series model. *Sci. Rep.* **6**, 33755; doi: 10.1038/srep33755 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2016