Learning from panel data of dengue incidence and meteorological factors in Jakarta, Indonesia

Wijaya Karunia Putra, Aldila Dipo, Erandi K. K. W. Hashita, Fakhruddin
Muhammad, Amadi Miracle, Ganegoda Naleen

This is a Author's accepted manuscript (AAM) version of a publication
published by Springer Nature
in Stochastic Environmental Research and Risk Assessment

DOI: 10.1007/s00477-020-01887-w

Copyright of the original publication:

© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Please cite the publication as follows:

Wijaya, K.P., Aldila, D., Erandi, K.K.W.H. et al. Learning from panel data of dengue incidence and meteorological factors in Jakarta, Indonesia. Stoch Environ Res Risk Assess 35, 437–456 (2021). <https://doi.org/10.1007/s00477-020-01887-w>

This is a post-peer-review, pre-copyedit version of an article published in Stochastic Environmental Research and Risk Assessment. The final authenticated version is available online at: <http://dx.doi.org/10.1007/s00477-020-01887-w>

**This is a parallel published version of an original publication.
This version can differ from the original published article.**

Learning from panel data of dengue incidence and meteorological factors in Jakarta, Indonesia

Karunia Putra Wijaya · Dipo Aldila · K. K. W. Hashita Erandi · Muhammad Fakhruddin · Miracle Amadi · Naleen Ganegoda

Received: date / Accepted: date

Abstract Medical statistics collected by WHO indicates that dengue fever is still ravaging developing regions with climates befitting mosquito breeding amidst moderate-to-weak health systems. This work initiates a study over 2009–2017 panel data of dengue incidences and meteorological factors in Jakarta, Indonesia to bear particular understanding. Using a panel random-effect model joined by the pooled estimator, we show positively significant relationships between the incidence level and meteorological factors. We ideate a clustering strategy to decompose the meteorological datasets into several more datasets such that more ex-

planatory variables are present and the zero-inflated problem from the incidence data can be handled properly. The resulting new model gives good agreement with the incidence data accompanied by a high coefficient of determination and normal zero-mean error in the prediction window. A risk measure is characterized from a one-step vector autoregression model relying solely on the incidence data and a threshold incidence level separating the low-risk and high-risk regime. Its magnitude greater than unity and the weak stochastic convergence to the endemic equilibrium mark a persistent cyclicity of the disease in all the five districts in Jakarta. Moreover, all districts are shown to co-vary profoundly positively in terms of epidemics occurrence, both generally and timely. We also show that the peak of incidences propagates almost periodically every year on the districts with the most to the least recurrence: Central, South, West, East, and North Jakarta.

Keywords Dengue · Clustering-integrated multiple panel regression · Risk measure · Spatial correlation · Outbreak propagation

1 Introduction

Until today, dengue fever and its debilitating variants remain to be notable health burdens in Indonesia. Factors influencing the widespread of dengue diseases in big cities like Jakarta are manyfold. Environmental support for mosquito breeding (Arcari et al., 2007; Morin et al., 2013), compromised healthy lifestyle (Haryanto, 2016), the presence of all four dengue serotypes (Kusriastuti and Sutomo, 2005), migration and populousness (Suwandono et al., 2006), close proximity to shores (Haryanto, 2016; Wijaya et al., 2019), lack of indoor visibility in typical households (Spiegel et al., 2005), lack of community participation (Haryanto, 2016), and pro and contra on dengue vaccines (Deng et al., 2020;

K. P. Wijaya
Mathematical Institute, University of Koblenz, DE-56070 Koblenz, Germany

D. Aldila
Department of Mathematics, University of Indonesia, ID-16424 Depok, Indonesia
E-mail: aldiladipo@ui.sci.ac.id

K. K. W. H. Erandi
Department of Mathematics, University of Colombo, SL-00700 Colombo, Sri Lanka

M. Fakhruddin
Department of Mathematics, Bandung Institute of Technology, ID-13240 Bandung, Indonesia
and
Department of Mathematics, Indonesia Defense University, 16810 Bogor, Indonesia

M. Amadi
Department of Mathematics and Physics, Lappeenranta University of Technology, FI-53851 Lappeenranta, Finland

N. Ganegoda
Department of Mathematics, University of Sri Jayewardenepura, SL-10250 Nugegoda, Sri Lanka

Wilder-Smith, 2020) are conspicuous factors, just to name. High population density also helps increase the incidence risk given the nature of *Aedes* mosquitoes as *multi-biters*, i.e., that a single vector can bite multiple humans (Jansen and Beebe, 2010; Harrington et al., 2014). Difficulties in suppressing dengue incidences set a tremendous challenge for the government. Besides uncontrolled urbanization flows, lack of perseverance, motivation, and community participation in anti-dengue actions give another problem. Among the enhancers of dengue widespread, meteorological factors have been a special subject of discussions due to their learnable patterns. In tropical countries, epidemics mainly occur in the rainy season and reach the highest peak given some time lags after meteorological factors favor mosquitoes to breed (Wang et al., 2019; Kakarla et al., 2019). Indonesia's Ministry of Health even combines climate and early warning system as marked fulcrums of intervention against dengue diseases (Haryanto, 2016). Furthermore, finding a meticulous early warning system that integrates meteorological factors has been an active research area.

A number of studies have been conducted to study the effects of meteorological factors on the dengue incidence level in the context of multiple regression models. Lowe et al. (2011) use a generalized linear model (GLM) and the generalized linear mixed model (GLMM) counterpart pronouncing highly endemic areas to model large panel data of monthly incidence levels in Brazil, associated meteorological factors (rainfall, temperature, and El Niño), and a series of relevant factors reflecting zones and their interaction. Rainfall and temperature, each of lags 1 and 2 months, were shown to significantly increase, whereas El Niño 3.4 index decreases the incidence level. Several previous studies have also proposed different forecasting techniques and decision-support systems for Indonesia. Ramadona et al. (2015) assess the relation between seasonal incidence level and meteorological factors (rainfall, temperature, humidity) in Yogyakarta, Indonesia during 2001–2013 using a GLM in which humidity data were first adjusted according to rainfall and temperature variation, i.e., due to multicollinearity. For all meteorological factors, all the data of time lags 0 to 3 months were used as showing positive correlations with the incidence level. An autoregression model was done separately, however returning weaker predictability as compared to the model involving also lagged meteorological factors. Astuti et al. (2019) apply a GLM to assess the short-term effect of climate on the incidence level and the results showed that temperature of lag 4 months, rainfall of lag 2 months, and humidity of lag 0 month were associated with high incidence levels in children. Spatial analyses were conducted by using Moran's autocorrelation coefficient and local indicator of spatial association (LISA) to explore geographical clustering in the incidences and to identify high-risk

villages. As a result, they identified the total of 38 high-risk villages in Cirebon, Indonesia for dengue. Chien and Yu (2014) combine a quasi-Poisson distributed lag nonlinear model (DLNM) with a spatial function adapting spatial autocorrelations and describing spatial heterogeneity to identify the relation between weekly minimum temperature, maximum 24-hour rainfall, and dengue incidence level in southern Taiwan during 1998–2011 in the panel sense. With a characterized measure Relative Risk (RR), they were able to show that the increase of both meteorological factors rose RR while certain levels of maximum 24-hour rainfall determine the lasting time of the associated RR. Another study for southern Taiwan during the highest epidemic year 2007 was done by Yu et al. (2011) via a log-link Poisson model incorporating panel incidence level, seasonal lag meteorological factors, El Niño Southern Oscillation Index (SOI), Breteau index, and panel population sizes. They found that most included meteorological factors are positively correlated with the incidence level while, surprisingly, most lag SOI (associated with warmer winter) and Breteau index (mosquito-positive containers) are negatively correlated with the incidence level. A set of multiple linear regression models integrating the least absolute shrinkage and selection operator (LASSO) method was proposed by Shi et al. (2016) to assess the relation between weekly incidence level, mean temperature, maximum hourly temperature, number of hours for high temperature ($> 27.8^{\circ}\text{C}$), humidity, and breeding percentage (BP) describing the ratio between *Ae. aegypti* and total breeding sites in Singapore during 2001–2012. Having shown good fitting and strong predictability via the mean absolute percentage error, the method entails caution for possible lack of interpretation due to complexity in selecting different lags among explanatory variables. Bouzid et al. (2014) attempt to project dengue incidences in 27 European Union members by assessing the incidences in Mexico during 1985–2007 and their relation with minimal and maximal temperature, rainfall, humidity, GDP, percentage of the urban population, and population density. A generalized adaptive model (GAM), which is a GLM involving transformed explanatory variables, was used for the fitting. Under short-term, medium-term, and long-term projection, areas around Mediterranean and Adriatic coasts and in northern Italy were forecasted to exhibit higher risk of infection. Similar related seasonal models with the same attempt to relate dengue incidence level with meteorological factors in Sri Lanka, India, and Guangzhou, China, can be found in (Withanage et al., 2018), (Ramachandran et al., 2016), and (Xu et al., 2017), respectively.

In comparison to the studies mentioned earlier, our base contributions lie in the following aspects. We develop a clustering-integrated multiple regression model for aggregated weekly panel data of incidence level and meteorological factors (rainfall, average temperature, humidity) for the

instance of Jakarta, Indonesia. A particular motivation of the clustering is the zero-inflated problem in the incidence data and limited accessible supporting data in Indonesia, while at the same time, a model with proper fitting and strong predictability is highly required to help decision-makers build up an early warning system. As the incidence data are clustered into low, medium, and high level, the associated meteorological factors are classified following the places of the incidence levels. After sequential tests, we opt for a random-effect model optimized using the pooled estimator. As varying the clustering barriers returns different modeling results, we then attempt to find the optimal barriers in the sense that the mean squared error is minimized. The fact that constant coefficients in the model are used for different districts prompts a unified model where the coefficients represent the marginal effects of the meteorological factors to the incidence level in the general sense, i.e., district-independent. Unlike previous studies, we are only using the lags for the meteorological factors associated with the maximal Spearman's correlation coefficients for computational efficiency. A framework for estimating risk is proposed via a one-step vector autoregression model involving only incidence data. The resulting risk measure can be used to study both disease persistence and which districts correspond to the low and high risk. As far as spatial competition is concerned, we perform global and local spatial autocorrelation analysis using Moran's and Geary's autocorrelation coefficient, respectively. The results show how all the districts either simultaneously raise the incidence level or exhibit significant discrepancy without given clear evidence, which district should be of the priority at a specific time point. For this reason, we determine how the highest peak of incidences from all the districts propagates over time through spatial periodicity investigation.

2 Preliminaries

2.1 Study area and data

The city of Jakarta is divided into six districts, among which the central part serves as the administrative regency. The districts are North Jakarta (approximated area 142 km², share 21.48%), East Jakarta (187 km², 28.3%), South Jakarta (146 km², 22.1%), West Jakarta (126 km², 19.1%), the Thousand Islands (12 km², 1.82%), and Central Jakarta (48 km², 7.2%). To the north stretches the coast along 35 km, where 13 rivers and two canals disembody. Due to a tiny share and almost-zero incidence level, we skip the Thousand Islands from our investigation. The means of the population shares from 2009 to 2017 (see Tab. 1) impart that North, East, South, West, and Central Jakarta share 16.99%, 27.89%, 21.63%, 23.89%, and 9.36% of the total population in the city, respectively. From the preceding calculation, we can

extract further information regarding the population-to-area ratio, where the quotient returns 0.79, 0.99, 0.98, 1.25, and 1.3 for the respective districts. The numbers clearly indicate that North, East, and South Jakarta are *relatively* underpopulated, whereas West and Central Jakarta are *relatively* overpopulated. Jakarta has the highest population density in Indonesia with the total population of 10,374,235 inhabitants (661 km² area) as of 2017, which is far above the second-largest city, Surabaya, with the corresponding population of 3,057,766 inhabitants (327 km² area). The weather condition of the city generally is warm and humid with a maximum temperature of 35.4°C during the day, and a minimum temperature of 21.5°C at night. The average rainfall throughout the year during the period 2009–2017 is 6.48 mm, and the air humidity falls within 47.5% – 99.5% on a daily basis.

2.2 Incidence data and scaling

All datasets undertaken in this study cover the time window January 6, 2009 – September 25, 2017, defined on a daily basis. The data of dengue hospitalized cases were collected from Jakarta Health Office, while the meteorological data were supplied by the Indonesian Agency for Meteorology, Climatology, and Geophysics. The dengue data are compiled records of dengue fever inpatients from 154 hospitals in Jakarta, irrespective of variants of illness. Note that 154 represents the total number of public and private hospitals that accommodate dengue patients and participate in dengue surveillance. Unlike meteorological data that are taken from unique regional weather stations and are ranging in comparable orders of magnitude, the incidence data suffer from bias due to the unmentioned population sizes. A region might appear to present large outliers in epidemics while serving as the most populated region. Another region might also show uniformity in the incidence levels with the others, however serving as the least populated region. To avoid further correction due to such bias, we utilize the density data. The required population data are taken from the annual book *Jakarta Dalam Angka* (Jakarta in Numbers) X+1 released by Statistics Indonesia, where X and X+1 denote the corresponding year in which the data were collected and the publication year, respectively. The data are presented in Tab. 1.

2.3 Reasons for data selection

Dengue virus is estimated to expand in the propagation that is marked by high rainfall, temperature, and humidity. An increment in the air temperature alters water temperature at the breeding place of mosquitoes that, in turn, hastens the egg

	2009	2010	2011	2012	2013	2014	2015	2016	2017
North	1471663	1645659	1716345	1715564	1711036	1729444	1747315	1764614	1781316
East	2448653	2693896	2926732	2801784	2791072	2817994	2843816	2868910	2892783
South	2159638	2062232	2135571	2148261	2141941	2164070	2185711	2206732	2226830
West	2221243	2281945	2260341	2395130	2396585	2430410	2463560	2496002	2528065
Central	902216	902973	1123670	908829	906601	910381	914182	917754	921344

Table 1 The total population of the five districts of Jakarta. Source: *Jakarta Dalam Angka* (Jakarta in Numbers) X+1, where X denotes the corresponding year the data were recorded.

hatching (Christophers, 1960; Byttebier et al., 2014). Additionally, such increments also shorten the extrinsic incubation period of the virus in mosquitoes (Ogden and Lindsay, 2016). A previous study on cross-correlation between monthly dengue incidence, rainfall, and temperature in eight provinces in Indonesia reveals that rainfall steers the geographical distribution while temperature intensifies the epidemics (Arcari et al., 2007). Other empirical studies confirm, however, that only under a proper combination of temperature and humidity can *Aedes aegypti* perform two-fold survival and produce evermore eggs (Juliano et al., 2002; de Almeida Costa et al., 2010). In fact, a certain range of humidity values leads to the accelerated larval growth and may decrease the proportion of the population surviving extrinsic incubation period (Christophers, 1960). Therefore, the role of humidity in the context of dengue widespread is inasmuch as that of temperature. Rainfall is a weather element whose every datum is measured using a rain gauge so that the amount can be known in millimeters. Stagnant water in abandoned bottles and cans can become natural mosquito breeders since *Aedes* mosquitoes favor to breed in standing water that is not in direct contact with the ground (Jansen and Beebe, 2010).

2.4 Data transformation

A few aspects shall be mentioned here regarding data pre-processing. The first problem appears as the incidence data contain lattice values, each of which has a relatively high recurrence. The resulting histogram merely shows a geometric distribution, which is of discrete type. We mitigate such incompatibility with the meteorological data by accumulating all the daily datasets into weekly datasets. The yielded cardinality is $454 \text{ weeks} + 6 \text{ days} \approx 455 \text{ weeks}$. The resulting datasets yet show another incompatibility in the context of linear regression. We will see that the incidence and rainfall dataset exhibit exponential distributions, whereas temperature and humidity dataset exhibit normal distributions. The eligibility of linear regression requires all datasets to be identically and independently distributed (*i.i.d.*). For wider compatibility with dummy or coded variables, we normalize the incidence and rainfall dataset. The

transformation idea follows from the so-called *Gaussianization* (Chen and Gopinath, 2000). For a random seasonal variable x of an exponential distribution, the first transformation into the cumulative distribution function (CDF) of the uniform distribution can be done via $\Theta(x) = 1 - \exp(-x/\bar{x})$, where $\bar{x}$ denotes the mean of x . Then, the second transformation $\Phi^{-1}(\Theta)$, where Φ denotes the CDF of the standard normal distribution, gives an approximation to the normal distribution. The preceding formula contains a flaw as the inverse of the error function returns $-\infty$ at 0. Consequently, overwhelming zero values obstruct the normalization process, which is the case in our incidence and rainfall dataset. Here, we ideate an approximation to the inverse of the error function, $\Psi : [0, 1] \rightarrow \mathbb{R}$. The function Ψ should behave similarly as Φ^{-1} and is flexible in terms of boundary conditions $\Psi(0)$, $\Psi(1)$, a mid-point M , and a stiffness measure κ . We come across, but not limited to, the following formula

$$\Psi = \kappa \cdot \tan(\eta \Theta + \nu) + M \quad (1)$$

where

$$\nu = \tan^{-1} \left(\frac{\Psi(0) - M}{\kappa} \right), \quad \eta = \tan^{-1} \left(\frac{\Psi(1) - M}{\kappa} \right) - \nu.$$

Since both Θ and Ψ are monotonically increasing, so is the composite form $\Psi \circ \Theta$. For both incidence and rainfall dataset, we use $\Psi(0) = 0$, $\Psi(1) = 1$, $M = 0.5$, and $\kappa = 0.1$. The corresponding transformation results can be seen in Fig. 1.

2.5 Calculation of true marginal effects

A critical viewpoint to be considered after data transformation is how the result of linear regression can be interpreted. For a simple seasonal model $y = \beta x + \varepsilon$, the marginal effect of x to y is defined as $dy/dx = \beta \approx \Delta y / \Delta x$. In this case, for any current value of the explanatory variable x , the increase of it by $(\Delta x / x) \times 100\%$ produces the increase of the corresponding response variable at y by $(\Delta y / y) \times 100\% \approx (\beta \Delta x / y) \times 100\%$.

In the sequel, the data are to be approached by (panel) linear regression models. When the pooled estimator is used,

Fig. 1 Distribution of the weekly incidence, rainfall, temperature, and humidity dataset in panel (a), (c), (e), (f), respectively. Those of the transformed incidence and rainfall datasets are shown in panel (b) and (d), respectively.

the computation of the coefficients is based on rewriting the panel model into a seasonal model by piling up district datasets into single-district datasets. In a first representation of the seasonal model, the interrelationship among datasets can be written as $d = \beta_0 + \beta_r r + \beta_t t + \beta_h h + \varepsilon$, where $d = (d_i)$, $r = (r_i)$, $t = (t_i)$, $h = (h_i)$ with $i = 1, \dots, IJ$ denote the concatenated dengue incidence, rainfall, temperature, and humidity dataset, respectively. Here, I and J denote the number of time points and that of districts, respectively. Under Gaussianization of the incidence and rainfall dataset, the model changes to

$$\Psi \circ \Theta(d) = \beta_0 + \beta_r \Psi \circ \Theta(r) + \beta_t t + \beta_h h + \varepsilon.$$

The marginal effects thereupon do not represent the true marginal effects. In case of the marginal effect of rainfall to dengue incidence, we obtain the true marginal effect β_r^{true} by the following identity

$$\beta_r = \frac{d\{\Psi \circ \Theta(d_i)\}}{d\{\Psi \circ \Theta(r_i)\}} = \frac{\Psi' \Theta'|_{d_i} \beta_r^{\text{true}}}{\Psi' \Theta'|_{r_i}}, \quad (2)$$

where the apostrophes in Ψ' and Θ' denote the first derivative with respect to the argument, and d_i, r_i any taken values from the datasets. A caveat here is, that β_r^{true} as in (2) returns different values when using d_i and r_i for different i 's.

Two notions mitigate the problems of having variable true marginal effects (Cameron and Trivedi, 2010). The first notion is the so-called *marginal effect at means*. The basic idea is that the true marginal effect should be evaluated at the means of the dengue incidence, $\bar{d}$, and rainfall

dataset, $\bar{r}$. As in (2), the marginal effect at means is given by $\beta_r^{\text{true}} = \beta_r \Psi' \Theta'|_{\bar{r}} / \Psi' \Theta'|_{\bar{d}}$. The second notion is the *average marginal effect*. As the name tells, it calculates $\Psi' \Theta'$ -terms using d_i and r_i for all i and uses the mean of all calculations $\overline{\Psi' \Theta'|_d} := (1/IJ) \sum_{i=1}^{IJ} \Psi' \Theta'|_{d_i}$ and $\overline{\Psi' \Theta'|_r} := (1/IJ) \sum_{i=1}^{IJ} \Psi' \Theta'|_{r_i}$ to replace the numerator and denominator in (2), respectively. Due to the absence of such individuals $\bar{d}$ and $\bar{r}$, the average marginal effect provides better results than the marginal effect at means in terms of interpretability. Therefore, this study uses the average marginal effects $\beta_r^{\text{true}} = \beta_r \overline{\Psi' \Theta'|_r} / \overline{\Psi' \Theta'|_d}$, $\beta_t^{\text{true}} = \beta_t / \overline{\Psi' \Theta'|_d}$, and $\beta_h^{\text{true}} = \beta_h / \overline{\Psi' \Theta'|_d}$.

Unless stated otherwise, in what follows we will only use the transformed data in order not to envisage possible confusion as well as keep I, J to denote the number of time points and that of districts, respectively.

2.6 Clustering

The aims of clustering in our study are not only to increase the number of explanatory variables but also to drop meteorological data that correspond to the zero incidence levels. The clustering idea is inspired by our previous work on seasonal models in (Fakhrudin et al., 2019). Let $D = (d_{ij})$, $R = (r_{ij})$, $T = (t_{ij})$, $H = (h_{ij}) \in \mathbb{R}^{I \times J}$ (where $i = 1, \dots, I$, $j = 1, \dots, J$) represent the panel incidence, rainfall, temperature, and humidity dataset, respectively. The data clustering in this study centers on the dengue incidence dataset. We restrict our investigation to the use of three clusters with a

lower b_l and an upper barrier b_u . An incidence case level d_{ij} is classified as low if $d_{ij} < b_l$, medium if $b_l \leq d_{ij} < b_u$, or high if $d_{ij} \geq b_u$. For brevity, we collect the barriers into a vector $b := (b_l, b_u)$. We can set an auxiliary function $\delta_l(d_{ij}, b)$, which is 1 in case d_{ij} belongs to the low cluster, i.e., $d_{ij} < b_l$, or otherwise 0. Constructions of $\delta_m(d_{ij}, b)$ and $\delta_u(d_{ij}, b)$ can be done similarly. Now the matrices

$$E^l := \begin{pmatrix} \delta_l(d_{11}, b) & \cdots & \delta_l(d_{1J}, b) \\ \vdots & \ddots & \vdots \\ \delta_l(d_{I1}, b) & \cdots & \delta_l(d_{IJ}, b) \end{pmatrix} \quad (3)$$

and E^m, E^u constructed as for E^l by replacing δ_l with δ_m, δ_u , respectively, denote the *clustering-specific effects*. These are binary matrices serving as dummy variables. Observe that $E^l + E^m + E^u = E$, the one-matrix. Let $P \odot Q = (p_{ij}q_{ij})$ denote the Hadamard product between two matrices P, Q of equal size, which is the entrywise product. Then, the matrices

$$R^l := R \odot E^l, \quad R^m := R \odot E^m, \quad R^u := R \odot E^u \quad (4)$$

denote the clustered rainfall datasets manifested from the clustering of the incidence dataset, where it holds $R^l + R^m + R^u = R$. The clustering in the temperature and humidity dataset can likewise be done as in the rainfall dataset.

2.7 Cross-correlation

We analyze the time lags between dengue incidence and meteorological factors before they are put into the modeling. The utilized tool is Spearman's correlation coefficient. We use Spearman's in favor of Pearson's correlation coefficient for its robustness against monotonic transformations over the datasets. Based on 455 data points from the first week of January 2009 until the third week of September 2017, we fixed the last 403 dengue incidence data and moved the 403-week windows of meteorological parameters back in time, for all districts. This way, we tested the correlations for the full one year's shift. During every window's fixation, we calculate the correlation coefficients and later find the maximum. The time lags corresponding to the maximal correlation coefficients can be seen in Tab. 2.

We found that all the incidence-rainfall, incidence-temperature and incidence-humidity correlations change in sign, i.e., indicating wavering sine-like curves. Especially for the incidence-temperature and incidence-humidity correlations, the maximum absolute value of the negative part is relatively slightly smaller than that of the positive part. The incidence-temperature correlations mostly start with negative coefficients, overtaken by positive coefficients as the lag increases. Meanwhile, the incidence-rainfall and incidence-humidity correlations behave the other way around. Such indicators of instability may be related to the fact that only

	Rainfall	Temperature	Humidity
North	11w, 0.2616	44w, 0.2288	9w, 0.3446
East	11w, 0.4308	22w, 0.2684	9w, 0.3149
South	7w, 0.3867	34w, 0.4089	6w, 0.4524
West	6w, 0.4037	21w, 0.2520	10w, 0.5086
Central	8w, 0.4565	29w, 0.2780	7w, 0.5083

Table 2 Time lags and maximal correlation coefficients found after Spearman's cross-correlation between dengue incidence and meteorological factors in the five districts in Jakarta. The unit w stands for week.

proper combinations of temperature and humidity can sustain *Aedes* mosquitoes' lives, whereas the rest can kill evermore (Juliano et al., 2002; de Almeida Costa et al., 2010). Now, let $\tau_j^r, \tau_j^t, \tau_j^h$ denote the time lags of rainfall, temperature, and humidity with respect to dengue incidence on the j th district, whose values can be taken from Tab. 2. As the maximum of all the time lags is 44w, we use the last 412 data points for dengue incidence, meanwhile the meteorological parameters are shifted back according to the time lags. Eventually, we change to $I = 412$ while $J = 5$ remains.

3 Random-effect model

Let us recall the one-matrix E and define new matrices based on the cross-correlations

$$R := (r_{i-\tau_j^r, j}), \quad T := (t_{i-\tau_j^t, j}), \quad H := (h_{i-\tau_j^h, j}).$$

Alluded from the panel regression is the use of individual-specific effects $\beta_1, \dots, \beta_{J-1}$, which only differ by districts. Leaving out an individual-specific effect for one district is intentionally avoiding linear dependence with the pre-determined constant β_0 . In order to fit into a matrix equation, we use the following operation for the individual-specific effects

$$\beta^d \otimes \mathbb{1} := (\beta_1 \cdots \beta_{J-1}) \otimes \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \Bigg\} I.$$

The notation $P \otimes Q = (pq_{ij})$ denotes the usual Kronecker product between two matrices P and Q . A simple form of linear regression model only reveals the direct interrelationship among the datasets

$$D = \beta_0 E + \beta^d \otimes \mathbb{1} + \beta_R R + \beta_T T + \beta_H H + \varepsilon. \quad (5)$$

The variable ε in the model (5) denotes the *idiosyncratic error*.

The eligibility of the linear model (5) relies on several assumptions, according to which certain diagnostics have to be done. First, the model must be linear with respect to all the coefficients β 's, which is evident, and $\mathbb{E}(\varepsilon_{ij}) = 0$ and

$\mathbb{E}(\beta_j) = 0$ for all i, j . The latest assumption is what we instantly impose in this study without any test, even though the standard t-test advocates all the individual-specific effects being significantly different from zero. Second, all the involved variables are assumed to be *i.i.d.* which we endeavor to see in Sec. 2.4. Additionally, we assume that all the individual-specific effects also serve as normal random variables without requiring any test. Third, the idiosyncratic error cannot be correlated with the explanatory variables as well as individual-specific effects, i.e., exogeneity. In connection with the first assumption, this means that $\mathbb{E}(\varepsilon_{ij}) = 0$ independently from the other predictor variables. Fourth, the individual errors $\varepsilon_j := (\varepsilon_{1j}, \dots, \varepsilon_{Ij})^\top$ are uncorrelated, i.e., no spatial correlation in the individual errors, and there is no serial correlation on ε_{ij} , i.e., $\text{Cov}(\varepsilon_{ij}, \varepsilon_{kj}) = 0$ independently from all the predictor variables when $i \neq k$. Violation to the latter especially leads to smaller standard errors of the coefficients and a higher coefficient of determination. We will use the Wooldridge test for serial correlation. A p-Value smaller than α rejects the null hypothesis that no serial correlation is found from the error.

The appearance of individual-specific effects β^d automatically designates the model to be either a fixed-effect or a random-effect model. Durbin–Wu–Hausman test (Davidson and MacKinnon, 1993) is performed to see if a random-effect model is chosen in favor of a fixed-effect model, i.e., the null hypothesis. The test is in line with the aforementioned third assumption on the eligibility of the linear model that the errors are uncorrelated with all the predictor variables. Additional to the eligibility of a random-effect model is the absence of multicollinearity. The higher the multicollinearity, the higher the standard deviations of the coefficients, and therefore the wider the confidence intervals and the smaller the t-statistics. An overt sign of multicollinearity is that the suspected explanatory variables have the same correlation to the response variable and standard deviation, yet their significance levels are radically different (Farrar and Glauber, 1967; Willis and Perlack, 1978). To see multicollinearity, we calculate the *Inverse of Variance Inflation Factor* (1/VIF). Any value of it measures one minus the coefficient of determination resulting from a linear regression among all explanatory variables where the one under measurement serves as the response variable. Consequently, a 1/VIF smaller than, say 0.1, indicates a multicollinearity problem (Mansfield and Helms, 1982). After all, multicollinearity only scopes out redundancy or possible overfitting and does not influence the validity of the significance tests.

A purposive measure to see the validity of the significance tests is error normality. One can say that the standard deviations of the coefficients cannot be reliable in case the error is not normal. This means that the confidence intervals should have been too wide or too narrow. Nonetheless, even

if the distribution is grossly non-normal, the significance tests will still provide good approximations for the standard deviations. We used Shapiro–Wilk W-test for normality (null hypothesis) of any random variable (Chen, 1971; Hanusz et al., 2016). Note that error normality is independent of whether the computed optimal coefficients are biased or unbiased. Finally, we check for heteroscedasticity, a specific feature showing that the variance of the error is dependent of the explanatory variables, otherwise homoscedasticity. One obvious sign of heteroscedasticity is, that when a certain explanatory variable tends to step up larger values, the error simultaneously exhibits larger jumps. When the pooled estimator is preferred over the random-effect estimator, we can use Breusch–Pagan test for heteroscedasticity (Gujarati and Porter, 2008). The p-Value is gained using a chi-squared statistic, of which smaller than α rejects the null hypothesis that the error is homoscedastic.

From the datasets, we are given the number of time points $I = 412$ and of districts or individuals $J = 5$. The panel data with such a characteristic are categorized as of a small sample size. This study compares the usual pooled estimator and random-effect estimator. Despite the evidence that the random-effect estimator is only consistent and asymptotically normally distributed under a large sample size (Schmidheiny, 2019), we perform such comparison using Breusch–Pagan Lagrange Multiplier test, where the null hypothesis states that the variance across districts is negligible (no panel effect). The pooled estimator, on the other hand, is unbiased in small sample sizes given clearance on the eligibility of the linear model and random-effect model. The only caveat against pooled estimator for panel regression is its inefficiency, i.e., the variances of the coefficients are not minimal and incorrect due to error normality (Schmidheiny, 2019). Keeping us warned on this fact, the pooled estimator is retained against the random-effect estimator for the sake of keeping further computations the simplest. At last, the usual variable significance t-test is done to determine whether every single coefficient involved in a regression model is significantly different from zero. Along with the single tests, the F-test's result is also presented to see if the overall coefficients of the predictor variables are significantly different from zero, i.e., that the current model provides a better fit to the data than a model that contains no explanatory variables, except the constant. In some cases, p-Values returning from t-tests can be large, whereas the accompanying p-Value from F-test is sufficiently small. This is another indicator of multicollinearity (Johnston, 1972).

We fixed $\alpha = 0.01$ overall the statistical tests in our study. According to Tab. 3, Durbin–Wu–Hausman test strongly recommends the random-effect in favor of the fixed-effect model. Moreover, Breusch–Pagan Lagrange Multiplier test gives p-Value 1, suggesting no difference in the pooled and random-effect estimator. All the ex-

	Val	AME	StD	p-Value	F p-Value	1/VIF	R ²	Adj R ²	W	B-P	D-W-H	Wo
β_0	-1.3564		.1181	0								
β_1	-.1211		.0097	0		.6218						
β_2	-.0813		.0106	0		.5254						
β_3	-.0327		.0105	.002		.5349						
β_4	-.0025		.0106	.813	0	.5292	.3002	.2978	0	0	1	0.0510
β_R	.1092	7.2914×10^{-9}	.0143	0		.7491						
β_T	.0052	2.8191×10^{-7}	.0005	0		.7451						
β_H	.0094	4.5642×10^{-7}	.0007	0		.6613						

Table 3 Results from model (5). AME, StD, Adj R², W, B-P, D-W-H, Wo stand for average marginal effect, standard deviation, Adjusted R², p-Value of Shapiro-Wilk W-test for normality, p-Value of Breusch-Pagan test for heteroscedasticity, p-Value of Durbin-Wu-Hausman test for a random-effect vs. a fixed-effect model, and Wooldridge test for the serial correlation, respectively.

planatory variables are significant, except for the fourth individual-specific effect. The F-test, however, indicates that all the variables are significant, which also exhibit no multicollinearity according to 1/VIF values. The model produces a non-normal, heteroscedastic error evincing no serial correlation. The average marginal effects reveal that the increases in rainfall, temperature, humidity by 10% at the means of the datasets $(\bar{D}, \bar{R}, \bar{T}, \bar{H}) = (7.4689 \times 10^{-6}, 45.7152, 195.97, 77.8258)$ results in that in the incidence by 0.45%, 73.97%, 47.56%, respectively.

4 Clustering-integrated model

4.1 General expression

So far, we have collected new explanatory variables based on the clustering, namely the clustering-specific effects E^l, E^m, E^u and clustered meteorological factors $R^l, R^m, R^u, T^l, T^m, T^u, H^l, H^m, H^u$. One trait of infectious diseases is the impact of past incidences on the number of current incidences. In this regard, autoregressive terms are also added to represent the sequential connection between past and current incidences. Our analysis reveals that the autocorrelation coefficients conceal relatively decreasing amplitudes. Focusing on positivity, we thus utilize for the sake of less computation effort, but not limited to, incidence cases in the past four weeks. In the matrix form, these cases can be collected into $D_{-1} := (d_{i-1,j}), \dots, D_{-4} := (d_{i-4,j})$. Let $k \in \{l, m, u\}$ and $s \in \{1, \dots, 4\}$. Then, the general form of the model integrating clustering is presented as follows

$$D = \beta_0 E + \sum_k \beta_E^k E^k + \beta^d \otimes \mathbb{1} + \sum_k \beta_R^k R^k + \sum_k \beta_T^k T^k + \sum_k \beta_H^k H^k + \sum_s \beta_{-s} D_{-s} + \varepsilon. \quad (6)$$

Note that some of the explanatory variables in (6) are still subject to dropping due to possible multicollinearity (overfitting). We also note that all the clustering-specific effects and clustered meteorological factors – therefore the

mean squared error – are subject to changes as we perturb the clustering barriers b_l, b_u . In the sequel, we will look for the optimal barriers b_l, b_u that give the minimum of the mean squared error.

4.2 Optimal clustering barriers

The model (6) can be arranged into a seasonal model taking the form

$$Y = X(b)\theta + \varepsilon, \quad b = (b_l, b_u),$$

with the pooled estimator $\theta = [X(b)^\top X(b)]^{-1} X(b)^\top Y$. Let $y_{\min} := \min Y$ and $y_{\max} := \max Y$. The dependency $X = X(b)$ returns from the fact that the clustered meteorological factors are dependent on the clustering barriers, see (3) and (4). Finding such optimal barriers requires to solve the following optimization problem

$$\begin{aligned} \min_b \quad & f(b) := \frac{1}{2} \left\| X(b) [X(b)^\top X(b)]^{-1} X(b)^\top Y - Y \right\|^2 \\ \text{s.t.} \quad & y_{\min} \leq b_l \leq b_u \leq y_{\max}. \end{aligned} \quad (7)$$

The fact that Heaviside functions $\delta_l(d_{ij}, b) = 1 - \mathcal{H}(d_{ij} - b_l)$, $\delta_m(d_{ij}, b) = \mathcal{H}(d_{ij} - b_l) \cdot (1 - \mathcal{H}(d_{ij} - b_u))$, and $\delta_u(d_{ij}, b) = \mathcal{H}(d_{ij} - b_u)$ from (3) take part in the objective function f in (7) gives us a non-smooth constrained optimization problem.

The first treatment on the problem (7) is to transform the constrained into an unconstrained type. The basic idea is to augment the constraint functions into the original objective function using a penalty function. The new objective function becomes sufficiently large as the argument verges on the boundary of the feasible domain. We use a penalty function that not only gives such a feature but also saturates the objective function to a certain large value in case the argument is actually outside the feasible domain. The message here is that staying outside the feasible domain is allowed, but one can never find an optimal solution there. This strategy would also hinder the objective function from obtaining

arbitrary large values around the boundary, possibly infinity, like in the usual penalty log function. Let $g \geq 0$ be a constraint function. The penalty function with the preceding specifications is modeled as

$$\mathcal{P}(g; \gamma, v) := \frac{1}{v} \left(\frac{\pi}{2} + \tan^{-1} \left(\frac{\gamma - g}{v} \right) \right).$$

The penalty function $\mathcal{P}$ resembles a Z-shaped curve bounded below by zero, where a downward jump happens around the mid-point γ . The parameter v , on the one hand, stretches the curve to obtain large values when $g < \gamma$. This means that, by passing a small value $\gamma \simeq 0$, $\mathcal{P}$ becomes dominantly large as $g \simeq 0$ and saturates to a large constant as $g < 0$. On the other hand, v also controls the stiffness of the jump. When v is sufficiently small, the curve of $\mathcal{P}$ stretches and exhibits tremendous stiffness where the slope at γ tends to $-\infty$. Including the penalty function, the objective function transforms into

$$f(b) \mapsto f(b) + \sum_{i=1}^3 \mathcal{P}(g_i(b); \gamma, v) \quad (8)$$

where $g_1 = b_l - y_{\min}$, $g_2 = b_u - b_l$, $g_3 = y_{\max} - b_u$.

This study undertook tests for the aforementioned parameters in which the optimization results are compared with the approximate optima using brute-force computations. We obtain a similar result as v and γ are varied within $[10^{-8}, 10^{-1}]$ and $[10^{-3}, 10]$, respectively.

4.3 PSO-solver

Let us rewrite $f(b) = \frac{1}{2} \langle V(b), V(b) \rangle^2$ from (7). The derivative is given by $f'(b) = \langle V(b), V'(b) \rangle$ where

$$\begin{aligned} V &= X[X^\top X]^{-1}X^\top Y - Y, \\ V' &= \left\{ X'[X^\top X]^{-1}X^\top + X[X^\top X]^{-1}X'^\top \right. \\ &\quad \left. - X[X^\top X]^{-1}[X'^\top X + X^\top X'] [X^\top X]^{-1}X^\top \right\} Y. \end{aligned}$$

Computing V' requires to relax the Heaviside functions in $X(b)$, e.g. $\mathcal{H}(d_{ij} - b_l)$, using their smooth approximation

$$\frac{1}{1 + e^{-2\zeta(d_{ij} - b_l)}} \xrightarrow{\zeta \rightarrow \infty} \mathcal{H}(d_{ij} - b_l).$$

With this approximation, the problem (7) becomes a smooth problem, to be handled by any gradient-based method. The advantage of using gradient-based methods is that only one “player” is used on every iteration. A disadvantage appears as the relaxation parameter ζ is set to be extremely large, in which case the derivative term V' discloses infinities. Also, imminent from such methods is the calculation of V' itself, which requires a number of matrix multiplications and inverse. Metaheuristics, such as Particle Swarm Optimizer

(PSO), do not require any information regarding the gradient of the objective function. PSO uses several “players” that create clusters shrinking to an optimal solution at times. The stochastic nature pronounced by random numbers in the algorithms helps increase the probability of finding the global optima, at least from the empirical point of view (Trelea, 2003). Chances will be higher with a higher number of players. Even though there has been immense development of PSO algorithms recently (Rana et al., 2011; Wang et al., 2018), this study uses the following variant for its simplicity

$$\begin{aligned} b_{t,i} &= b_{t-1,i} + c_0 \cdot v_{t,i}, \\ v_{t+1,i} &= \omega \cdot \text{rand} \cdot v_{t,i} \\ &\quad + c_1 \cdot \text{rand} \cdot (p_{t,i} - b_{t,i}) + c_2 \cdot \text{rand} \cdot (q_t - b_{t,i}). \end{aligned}$$

Here, rand denotes a random number in $[0, 1]$.

The algorithm features the update of players’ position b (in this case, the barriers) and velocity v . There, t, i denote the pointers for the iteration and player, respectively. Let us in addition introduce a measure for the individual best optimum $y_{t,i}$, which together with $p_{t,i}$, are given initial conditions. Let f_g denote the minimum of the objective function evaluations over all players and the entire history. As $f(b_{t,i}) < f_g$, we update $p_{t,i} = b_{t,i}$ and $y_{t,i} = f(b_{t,i})$. In the algorithm, the entity $q_t := p_{t,i^*}$ where $i^* = \arg \min_i y_{t,i}$ denotes the best player among all players on the current iteration. The *self confidence* c_1 determines how confident every player in performing local search as if the first equation signifies $p_{t,i}$ when c_1 is dominant among other parameters. The *player’s confidence* c_2 measures every player’s confidence against the global best position found on every iteration, making all players attracted to a single best position when $c_1 \ll c_2$. The case $c_1 \gg c_2$ means that players are trapped in their local best positions. The *inertia* ω emphasizes the influence of the historical velocities in determining the current velocities, while the *constriction factor* c_0 determines how far players can jump around between iterations. This study engineers 100 players, which indicate a relatively large swarm. For smooth objective functions, empirical studies suggest to use between 5 to 30 players (van den Bergh and Engelbrecht, 2001; Brits et al., 2002). Previous studies (van den Bergh, 2002; Brits et al., 2002; van den Bergh and Engelbrecht, 2006) recommend to have a tradeoff between disproportionate wandering ($c_1 \gg c_2$) and premature finding of optima ($c_1 \ll c_2$), therefore we attested $c_1 = c_2 = 1$. Additionally, we use $\omega = 1$ and $c_0 = 0.3$.

In connection with the previous section, we show in Fig. 2 under excessive brute-force computations how the relaxation of the objective function $f(b)$ progresses towards its true evaluation as the relaxation parameter ζ increases. Along with this result, the PSO iterations are shown to not only approach the approximated global minimum from the brute-force computation but also locates the closest point to the true one.

Fig. 2 Excessive brute-force computation of the objective function using (a) $\zeta = 2$, (b) $\zeta = 10$, (c) $\zeta = 30$, (d) $\zeta = 70$, (e) $\zeta = 100$ where the small blue circle encodes the approximated global minimum. Realization of PSO algorithm on (f) initial condition, (g) 5th-iteration, (h) 10th-iteration and (i) 20th-iteration.

4.4 Diagnostic results

We test the model (6) giving the trade-off between a large coefficient of determination, multicollinearity-free, and significance. We thus drop a variable when it is insignificant in order to preserve the coefficient of determination, then check the multicollinearity. Under this process, the following variables are dropped in the order of reading: H^l , β_3 , β_4 ,

E^m , E^l , E^u , R^u , T^u , T^l , H^m . The final model is given as

$$D = \beta_0 E + (\beta_1, \beta_2) \otimes \mathbb{1} + \beta_R^l R^l + \beta_R^m R^m + \beta_T^m T^m + \beta_H^u H^u + \sum_s \beta_{-s} D_{-s} + \varepsilon, \quad (9)$$

with the corresponding diagnostics presented in Tab. 4. This last format suggests that only rainfall corresponding to the low and medium, temperature to the medium, humidity to the high dengue incidence levels as well as the lag autoregressive terms determine the entire incidence level.

The least average marginal effect for β_R^l also suggests to drop the corresponding variable, from which case the

Fig. 3 Panel (a) depicts the data and clustering with the barriers found from the PSO algorithm. The low, medium, and high incidence levels are colored green, blue, and red, respectively. Panels (b)–(d) show the comparison between the model (9) using the coefficients' values in Tab. 4 (in blue) with the real data (in black).

clustering strategy successfully clears up the zero-inflated problem from the incidence data. Durbin–Wu–Hausman test strongly suggests using the random-effect model, whereas, Breusch–Pagan Lagrange Multiplier test gives p-Value 1, suggesting to opt for the pooled estimator. All the predictor variables are found to be significant, with a note that the second individual-specific effect is only significant at $\alpha = 0.1$ level. The F-test indicates the significance of the overall predictor variables. No multicollinearity is found. The produced coefficient of determination is relatively high, with the warning of overestimation due to the presence of serial correlation in the error. Moreover, the error is also found to be heteroscedastic with improved normality as compared to the simple model (5). Fig. 3 depicts the computation results for the barriers and for the optimized model (9).

5 Prediction strategy

Here we pay our attention to a prediction algorithm. Making a prediction for the model involving clustered variables is not as instant as that for the simple model. Suppose that, with the help of a meteorological agency, the meteorological factors are predicted for several weeks in the future. To determine the predicted incidences, we need to cluster those factors. However, clustering cannot be defined unless we also know the incidences in the future. We are just encountering an indefiniteness.

Let Δ be the number of incidences on the first week in the prediction horizon. Using one row from the matrix equation (9), we see that the clustering-specific effects, as well as the clustered meteorological parameters, are dependent not only on the optimal barriers but also on Δ . In other words,

	Val	AME	StD	p-Value	F p-Value	1/VIF	R ²	Adj R ²	W	B-P	D-W-H	Wo
β_0	-.0463		.0097	0								
β_1	-.0235		.0039	0		.8318						
β_2	-.0068		.0038	.076		.8861						
β_R^l	-.0909	2.5585×10^{-10}	.0299	.002		.5263						
β_R^m	.0349	7.9554×10^{-9}	.0066	0		.6135						
β_T^m	.0012	1.6785×10^{-8}	.0000	0	0	.1589	.8496	.8489	0	0	1	.0031
β_H^u	.0058	1.6216×10^{-7}	.0001	0		.1496						
β_{-1}	.2383	.2724	.0149	0		.3278						
β_{-2}	.1245	.1962	.0158	0		.2944						
β_{-3}	.0518	.0147	.0156	.001		.2982						
β_{-4}	.0726	.0831	.0145	0		.3447						

Table 4 Results from model (9). AME, StD, Adj R², W, B-P, D-W-H, Wo stand for average marginal effect, standard deviation, Adjusted R², p-Value of Shapiro–Wilk W-test for normality, p-Value of Breusch–Pagan test for heteroscedasticity, p-Value of Durbin–Wu–Hausman test for a random-effect vs. a fixed-effect model, and Wooldridge test for the serial correlation, respectively.

the “correct” Δ should satisfy one row of the matrix equation given the optimal coefficients. For short, this row equation reads as

$$\Delta = \mathcal{S}(\Delta). \quad (10)$$

For every fixed district index j , eq. (10) unfolds the scalar equation

$$\Delta_j = \beta_0 + \beta_j + \beta_R^l \delta_l(\Delta_j, b) r_j + \beta_R^m \delta_m(\Delta_j, b) r_j + \beta_T^m \delta_m(\Delta_j, b) t_j + \beta_H^u \delta_u(\Delta_j, b) h_j + \sum_s \beta_{-s} \Delta_{-s,j} \quad (11)$$

where $\Delta_{-s,j}$ represent all the incidences in the past 4 weeks and $b = (b_l, b_u)$ denote the optimal barriers. For $j \geq 3$, we have $\beta_j = 0$.

Algorithm 1 Fixed point iteration for prediction.

Require: error tolerance χ , $\Delta_{\text{pred}} := \emptyset$, and Δ_{-1}

```

1: for  $i = 1, \dots, n$  do
2:   set  $\Delta^{(0)} := \Delta_{-1}$ 
3:   set  $\delta := \chi + 1$ 
4:   set  $k := 1$ 
5:   while  $\delta \geq \chi$  do
6:      $\Delta^{(k)} := \mathcal{S}(\Delta^{(k-1)})$ 
7:     set  $\delta := \|\Delta^{(k)} - \Delta^{(k-1)}\|$ 
8:     set  $k := k + 1$ 
9:   end while
10:  set  $\Delta_{\text{pred}} := \Delta_{\text{pred}} \cup \{\Delta^{(k-1)}\}$ 
11:  set  $\Delta_{-1} := \Delta^{(k-1)}$ 
12: end for
Ensure:  $\Delta_{\text{pred}}$ 

```

Our key method to solve (10) is using the standard fixed point iteration. Algorithm 1 walks us through the basic idea. The procedure for a prediction test entails the division of the time window emanated from the original data into a testing and a prediction window. The testing window consists of 380 time points and the prediction window of 32 time points. Basically, we impose $n = 32$ in the algorithm. The optimal

coefficients of the model (9) are calculated using the barriers produced by PSO only for the corresponding data in the testing domain. The resulting coefficients are then used to evaluate the prediction in step 6 in Algorithm 1. We note that the definite values of the Heaviside functions in $\delta_l, \delta_m, \delta_u$ help designate the right-hand side of (11) having zero derivatives almost everywhere except at the barriers. The piecewise-affinity and continuous differentiability implying Lipschitz continuity suggest that the iterative algorithm converges as long as Δ_j never traces one of the barriers during iterations. One can appoint a sufficiently small Lipschitz constant for showing the contractivity of the right-hand side function. Violating the condition on not passing through a barrier might still lead to convergence, but is not guaranteed theoretically. The entire prediction result is then presented in Fig. 4. The t-test has been done to show that the error is normal with the mean zero at $\alpha = 0.1$ level.

6 Next generation operator and risk mapping

This section is devoted to the estimation of a number determining risk, relying solely on the incidence data. Therefore, the forthcoming discussions use the original density data without any transformation. The underlying idea is taken from the next generation operator in population dynamics. Accompanied by the significance of the autoregressive terms (cf. Tab. 4), we depart from the assumption that the incidence cases in a certain district in the current week are attributed to those in the other districts in the past week. Of course, one can also consider the past two-week or three-week cases for more sophistication. Short incubation-to-illness period (Ogden and Lindsay, 2016), short adult mosquito lifespan and mosquito traveling distance (Christophers, 1960) are among several reasons why we focus on such “freshly” infected cases. Therefore, all

Fig. 4 Prediction result (in blue) based on Algorithm 1 in comparison with the real data (in black) on the 32-week prediction window.

cases happening at a certain region in the past week introduce marginal effects to the cases at the other regions in the current week. Recall our incidence panel data $D = (d_{ij})$ and let $D_i = (d_{i1} \cdots d_{iJ})^\top$ denote the incidence cases collected from all districts at time t_i . The basic model is then governed by the linear difference equation

$$D_{i+1} = \mu + GD_i, \quad \mu \geq 0, \quad (12)$$

where every element of G , g_{ij} , denotes the marginal effect of the incidence cases at district j to those at district i . The vector μ denotes a certain baseline determining to which level the D -sequence converges, when it does. In the present context, the matrix G serves as the *next generation operator*, since it determines the “next generation” of incidence cases by knowing the present cases. If d_{ij} were the number of individuals of an age class j in the population demography, then the matrix G is called *Leslie matrix* (Caswell, 2000).

Due to stochastic nature of the data, the deterministic model defined above contains some seasonal error. Technically speaking, the error returns from the exclusion of unobservable entities, which may also be meteorological factors. Moreover, districts that experience large fluctuations in the incidences would need to be distinguished from others. In the present aim, we would like to model the uncertainty taking into account its state-dependence. Only if a proper noise in the model is found based on data can the risk mapping

be more realistic. Our model including uncertainty is then governed by

$$D_{i+1} = \mu + GD_i + H(D_i)\xi_{i+1}, \quad (13)$$

where (ξ_i) are multivariate $\mathcal{N}(0,1)$ -distributed random variables. Our next step is determining the state-dependent covariance function $H(D_i)$ via the representation of the stochastic difference equation (13) in a stochastic differential equation. We first reform the equation (13) into

$$D_{i+1} = D_i + \mu + [G - \text{id}]D_i + H(D_i)\xi_{i+1}.$$

Observe that this equation is the Euler–Maruyama approximation (Kloeden and Platen, 1992) for the stochastic differential equation

$$dD = (\mu + [G - \text{id}]D) dt + H(D) dW \quad (14)$$

using the step size of 1 week. All elements μ_i , $(g_{ij} - \delta_{ij})d_j$ (with $i, j = 1, \dots, J$ and δ_{ij} denoting the Kronecker delta) represent “events”, indexed by k where $k = 1, \dots, J(J+1)$. The frequencies of these events are given by

$$\lambda_k := \begin{pmatrix} 0 \\ \vdots \\ \text{sign}(\mu_i) \\ \vdots \\ 0 \end{pmatrix} \longleftarrow i$$

for $k = 1, \dots, J$ and

$$\lambda_k := \begin{pmatrix} 0 \\ \vdots \\ \text{sign}(g_{ij} - \delta_{ij}) \\ \vdots \\ 0 \end{pmatrix} \longleftarrow i$$

for $k = J+1, \dots, J(J+1)$. The probabilities of these events happening after a certain time step Δt are given by $P_k = |\mu_i| \Delta t$ for $k = 1, \dots, J$ and $P_k = |g_{ij} - \delta_{ij}| d_j \Delta t$ for $k = J+1, \dots, J(J+1)$. Using the idea in (Allen and Burgin, 2000; McCormack and Allen, 2006), the incidence increase ΔD is assumed to follow the Gaussian process

$$\Delta D = \mathbb{E}(\Delta D) + \sqrt{\text{Var}(\Delta D)} \frac{\Delta W}{\sqrt{\Delta t}} \quad (15)$$

where W denotes the multidimensional Wiener process such that $\Delta W / \sqrt{\Delta t} \sim \mathcal{N}(0, 1)$. We have $\mathbb{E}(\Delta D) = \sum_{k=1}^{J(J+1)} \lambda_k P_k = (\mu + [G - \text{id}]D)\Delta t$, which is our original *drift* in (14). Meanwhile, $\text{Var}(\Delta D) = \mathbb{E}(\Delta D \Delta D^\top) - \mathbb{E}^2(\Delta D) \approx \mathbb{E}(\Delta D \Delta D^\top) = \sum_{k=1}^{J(J+1)} P_k \lambda_k \lambda_k^\top$ for a sufficiently small Δt . Since this formula is linear in Δt , then the factor Δt in the *noise* of (15) entirely disappears. We can now take $\Delta t \rightarrow 0$ in (15) to get (14), where then

$$H^2(D) = \sum_{k=1}^{J(J+1)} \begin{pmatrix} 0 & \dots & \dots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & \dots & P_k & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & \dots & \dots & 0 \end{pmatrix}.$$

Apparently, H^2 is just a linear function and has the format of a diagonal matrix. Surely is taking the square root of H^2 straightforward. Now with the function H being formulated, we are still worrying if this noise would lead the solution of the original model (13) jumps over larger values than the error emanated from the optimization. For this reason, we introduce a *damping factor* σ to control the volatility emanated from the state-dependent noise $H(D)$ and correct the model (13) into

$$D_{i+1} = \mu + G D_i + \sigma H(D_i) \xi_{i+1}. \quad (16)$$

The value of σ is to be estimated by how far the error from the deterministic model delineates that from the datasets. Under the model (16), we know that the endemic equilibrium $D^* := \mathbb{E}(D_i) = [\text{id} - G]^{-1} \mu$ is stochastically stable (i.e. stable in probability) providing that the maximal real part of the eigenvalues of the Jacobian $[G - \text{id}]$ is negative. The latter happens if all the eigenvalues of G lie in the open unit disc in $\mathbb{C}$, which is the case if the spectral radius $\rho(G) < 1$. Moreover, the variance $\text{Var}(D_i) =$

$[\text{id} - G^2]^{-1} \sigma^2 H^2(D_{i-1}) \Sigma$, where $\Sigma = \text{Var}(\xi_i)$, appears to be state-dependent.

Suppose that G_+, G_- are nonnegative matrices containing absolute entries of G that are nonnegative and negative, respectively. Otherwise, their elements are zero. We have $G_+ - G_- = G$. If $\|G_-\|$ is sufficiently small, then the endemic equilibrium enjoys the approximation $D^* = [\text{id} - G_+ + G_-]^{-1} \mu \approx [\text{id} - G_+]^{-1} \mu - [\text{id} - G_+]^{-1} G_- [\text{id} - G_+]^{-1} \mu$. When non-zero entries of G_- are solely off-diagonal, then a contradiction can be done to show that $\rho(G) < 1$ implies $\rho(G_+) < 1$. Therefore, we obtain $D^* \approx [\text{id} - G_+]^{-1} \mu$ in the first-order approximation, which is always nonnegative due to the property of M-matrix $[\text{id} - G_+]$, see (Berman and Plemmons, 1994). On optimality, D^* approximates the means of the datasets since otherwise the solution of (16) deviates away from the data. This implies that when $\rho(G_+)$ is close enough to 1 such that $[\text{id} - G_+]^{-1}$ is ready to blow up, μ adjusts to be small enough. Moreover, $D^* = (d_1^*, \dots, d_j^*) \rightarrow \mu$ when $\rho(G) \rightarrow 0$.

Let d_{sep} denote a barrier separating the *low-risk regime* ($d_j^* < d_{\text{sep}}$) and *high-risk regime* ($d_j^* > d_{\text{sep}}$). In the low-risk regime, one would say that the disease is such insignificant that the surveillance and treatments can be kept in the current magnitudes. Based on the biological existence of the endemic equilibrium, we estimate a *risk measure* as $\mathcal{R}^j = A_0 + A_1 (d_j^* - d_{\text{sep}})$. Two conditions are assigned in order to determine the unobservable parameters A_0, A_1 . Here, we assume that $\mathcal{R}^j = 1$ when $d_j^* = d_{\text{sep}}$, implying $A_0 = 1$. We also assume that under the low-risk regime, $\rho(G)$ dominates the value of $\mathcal{R}^j$ and eventually $\mathcal{R}^j = \rho(G)$ when $d_j^* = 0$. We acquire $A_1 = (1 - \rho(G)) / d_{\text{sep}}$ and finally

$$\mathcal{R}^j := \rho(G) + (1 - \rho(G)) \frac{d_j^*}{d_{\text{sep}}}. \quad (17)$$

The formula in (17) assures the nonnegativity of $\mathcal{R}^j$, which has a certain biological implication. Observe that under the low risk regime ($d_j^* < d_{\text{sep}}$), it always holds $\mathcal{R}^j < 1$ independent of $\rho(G)$. The value $\mathcal{R}^j$ gets even smaller as $\rho(G) \ll 1$. Only when d_j^* is small enough does $\rho(G)$ determine $\mathcal{R}^j$. Under the high risk regime ($d_j^* > d_{\text{sep}}$), $\mathcal{R}^j > 1$ but $\mathcal{R}^j$ gets larger as $\rho(G) \ll 1$. Note that when $\rho(G) \ll 1$, the noise emanated from (16) exhibits small jumps, meaning that the smaller $\rho(G)$, the more attractive the endemic equilibrium is or the higher the sureness that the solution trajectory converges to the endemic equilibrium. As the regional *risk measure* gives individual persistence of the disease, we define the global *risk measure* as their mean

$$\mathcal{R} := \sqrt[j]{\mathcal{R}^1 \dots \mathcal{R}^J}. \quad (18)$$

The geometric mean is used in the formula in order to avoid the severe effect of outliers. The facts that the datasets have positive means, and apparently, $\rho(G) \gg 0$ avoid any single

$\mathcal{R}^j$ being close to zero. In this study, the **risk measure** (18) clearly relies on the definition of the threshold parameter d_{sep} , under and above which every district or region can be categorized as of low and high risk. This allows one to share the responsibility in defining the transmission severity from a modeler to the decision-maker.

For a test case, we define $d_{\text{sep}} := (\max_j d_j^*)/2 \approx 5.49 \times 10^{-6}$, roughly approximating half of the maximal mean of the incidence datasets or 55 per 100,000 inhabitants. This yields the **regional risk measures** .9798, .9914, 1.0556, 1.0660, 1.0986 and **global risk measure** $\mathcal{R} \approx 1.0373$. With the chosen d_{sep} , North and East Jakarta are essentially of low risk, whereas South, West, and Central Jakarta are essentially of high risk. Additionally, attributed to $\rho(G) \approx .9063$ is the weak stochastic stability of all the equilibrium states. This advocates the situation that all the districts remain to be of considerable non-vanishing risk.

7 Spatial autocorrelation

Spatial autocorrelation, in the present context, performs a comparison of the incidence levels of a certain district with those of its neighbors. The underlying concept thus relies on the definition of *close neighbors*, i.e., a set of neighboring districts in which commuting between neighbors is possible. In the presence of spatial autocorrelation, the incidence levels are either positively or negatively linked to each other. A positive linkage indicates similar incidence levels appearing among neighbors from a certain timestamp, which is called *spatial mutualism*. When dissimilar incidence levels are instead observed, a negative linkage or *spatial competition* happens.

7.1 Global autocorrelation coefficient

The widely-used idea for computing the spatial autocorrelation coefficient is due to Moran (1950). In our case, we consider $J = 5$ close neighbours. Let d_{ij} ($i = 1, \dots, I$ and $j = 1, \dots, J$) be the incidence cases of the close neighbours during the discrete time points $t_1, \dots, t_I$. By writing

$$\bar{d} = \frac{\sum_{i,j} d_{ij}}{IJ} \quad \text{and} \quad z_{ij} = d_{ij} - \bar{d} \quad (\text{i.e. } \sum_{i,j} z_{ij} = 0),$$

Moran's autocorrelation coefficient is defined as

$$r := \left(\frac{IJ}{2IJ - I - J} \right) \frac{\sum_i \sum_{j=1}^{J-1} z_{ij} z_{i,j+1} + \sum_{i=1}^{I-1} \sum_j z_{ij} z_{i+1,j}}{\sum_{i,j} z_{ij}^2}.$$

The mean and standard deviation of this statistic are presented in the original paper (Moran, 1950), allowing to compute p-Value after normalization thus perform a statistical

inference under sufficiently large datasets. The coefficient is outlined by three numbers: -1 (perfectly negative autocorrelation), 0 (no autocorrelation) and 1 (perfectly positive autocorrelation). Since the coefficient is parameter-free, a straightforward computation returns $r = 0.7530$. This suggests that the close neighbors co-vary positively on the overall observation time (p-Value = 0.623). Practically, we are suggested not to only worry on some but all districts during epidemics.

7.2 Time-dependent autocorrelation coefficient

To break down the autocorrelation measurement on the weekly level, a time-dependent coefficient is used instead. Let us freeze time and consider J close neighbours with incidence levels d_j ($j = 1, \dots, J$). Denote by w_{jk} ($j, k = 1, \dots, J$) the *spatial connectivity measure* between j th and k th district. Write $J\bar{d} = \sum_j d_j$ and $z_j = d_j - \bar{d}$. The so-called Moran's time-dependent autocorrelation coefficient (Cliff and Ord, 1981; Anselin, 1995) is presented as

$$r_M := \left(\frac{J}{\sum_{j,k} w_{jk}} \right) \frac{\sum_j \sum_k w_{jk} z_j z_k}{\sum_j z_j^2}.$$

A drawback of Moran's autocorrelation coefficient is its instability for general problems, as seen shortly. Datasets representing the same title can have different values in case there are several parties who do the same measurement/estimation (e.g. local and international bodies), or can lose certainty during data collection and transfer. We can thus generate noise over the incidence data Δd , defined uniformly for all d_j , i.e. $d_j \mapsto d_j + \Delta d$. An example could be $\Delta d = \sigma_d Z$ for some σ_d where $Z \sim \mathcal{N}(0, 1)$. Due to such noise, Moran's correlation coefficient suffers from the following change

$$\begin{aligned} r_M &\mapsto \frac{\sum_j z_j^2}{\sum_j (z_j + \Delta d)^2} r_M \\ &+ \frac{\sum_j z_j^2}{\sum_j (z_j + \Delta d)^2} \left(\frac{J}{\sum_{j,k} w_{jk}} \right) \frac{\sum_j \sum_k w_{jk} (z_j + z_k)}{\sum_j z_j^2} \Delta d \\ &+ \frac{\sum_j z_j^2}{\sum_j (z_j + \Delta d)^2} \left(\frac{J}{\sum_{j,k} w_{jk}} \right) \frac{\sum_j \sum_k w_{jk}}{\sum_j z_j^2} \Delta d^2. \end{aligned}$$

It is clear that the perturbed coefficient depends on not only the incidences but also weighting measures, cf. (Upton and Fingleton, 1985).

The more stable coefficient was introduced by Geary (Geary, 1954; Anselin, 1995), which is called Geary's autocorrelation coefficient

$$r_G := \left(\frac{J-1}{2 \sum_{j,k} w_{jk}} \right) \frac{\sum_j \sum_k w_{jk} (d_j - d_k)^2}{\sum_j z_j^2}.$$

The less sensitivity is told by the behaviour of the coefficient with respect to perturbation over the data. Introducing noise as before, Geary's autocorrelation coefficient takes the following change

$$r_G \mapsto \frac{\sum_j z_j^2}{\sum_j (z_j + \Delta d)^2} r_G.$$

One can see that the perturbed Geary's r_G only takes the first term of Moran's r_M , which solely depends on the incidences. According to (Cliff and Ord, 1981), $\mathbb{E}(r_G) = 1$ represents no spatial autocorrelation, while $r_G < 1$ and $r_G > 1$ point to a positive and a negative spatial autocorrelation among the close neighbours. A high similarity is indicated by the situation where $d_j \simeq d_k$ for any j, k and therefore, r_G close to 0.

To compute the time-dependent coefficient, we require to define the connectivity measures w_{jk} . A first insight was gained from the 2014 commuter data in Jakarta, representing the average daily percentages of commuters from all the regions (Irawan and Pragesari, 2014). However, the same percentages from two different districts of discrepant populations inevitably hide different proportions. To avoid bias, we instead use the proportions, each of which represents the percentage multiplied by the total population of the district where the commuters depart. The inability to access similar data from different years leads us to constant percentages but time-dependent w_{jk} due to varying populations, see Tab. 1. The computation result of timely r_G can be seen in Fig. 5(a). The mean value $\bar{r}_G \approx .0934$ reveals that all the close neighbors co-vary significantly positively from time to time, amplifying the global autocorrelation.

8 Outbreak propagation

This section asks how far the trajectory for the highest peak of incidences can be learned. In case the highest peak shows spatial periodicity, it would help decision-maker see which districts the highest peaks locate on a certain prediction horizon. As a result, the targeting for the intervention measures can be more accurate. Technically, we aim to locate the maximum of the incidence level for every week that correspond to which district giving the maximum. Then, data containing district indices related to such highest epidemics can be presented as a time series $\text{idx} = (\text{idx}_i)$. To see the periodicity, we calculate the Fourier spectrum of the created seasonal data using Fast Fourier Transform. The result is given in Fig. 5(b). One can see that the leading low frequencies are not clearly distinguished from the high frequencies, meaning that the index data are weakly periodic. For the sake of bearing a rule of thumb or guidance for control interventions, revealing a certain spatial periodicity for the peak of

outbreaks can still be important. Based on the Fourier analysis, the computed leading frequency corresponds to the period of 52 weeks. The period suggests that the incidence peaks resemble annual cases. Now we can generate a periodic function of period 52 weeks that is as close as possible to the index data. The corresponding optimization problem reads as

$$\begin{aligned} \min_{x \in \{1, \dots, J\}^I} \quad & \sum_{i \in I} 1 - \delta(x_i, \text{idx}_i) \\ \text{s.t.} \quad & x_i = x_{i+52}, i = 1, \dots, I - 52, \end{aligned}$$

where $\delta(x, y)$ denotes the Kronecker delta function. This is a constrained integer optimization problem with the cardinality of the solution space $J^I = 5^{412}$. A first step to treat the problem is redesigning it to integrate the constraint into the search space, which returns

$$\begin{aligned} \min_{x \in \{1, \dots, J\}^{52}} \quad & \sum_{j=0}^{\lfloor I/52 \rfloor - 1} \sum_{i=1}^{52} 1 - \delta(x_i, \text{idx}_{52j+i}) \\ & + \sum_{i=1}^{I - \lfloor I/52 \rfloor 52} 1 - \delta(x_i, \text{idx}_i). \end{aligned} \quad (19)$$

Now the cardinality reduces to $J^{52} = 5^{52}$. This study employs a genetic algorithm to solve (19) engineering 20 players (individuals) in every generation. The best objective function value and the mean of all the 20 values for all generations can be seen in Fig. (5)(c). On the 97th-iteration onwards, the best objective function value turns to be constant at 265. This means that only 18 out of 52 data points per 52 weeks can be fitted by the optimal periodic solution x as depicted in Fig. (5)(d). This figure is in line with the finding on the risk measure, that North and East Jakarta are considered of low risk as the peaks only occur once and twice per year, respectively.

9 Concluding remarks

We derive a panel random-effect model describing a direct relationship between lagged meteorological factors (rainfall, temperature, humidity) and dengue incidence level. We show that rainfall has the time lag ranging between 6w to 11w, temperature between 21w to 44w, humidity between 6w to 10w in the past leading to the current magnitudes of incidence levels for all the five districts in Jakarta. Based on the model, all the meteorological factors are shown to influence the incidence level significantly positively. The fact that a small coefficient of determination leads to biased error in the prediction window (not zero-mean normal) and that the incidence data suffer from the zero-inflated problem, we initiate a clustering to the incidence data. The timely meteorological factors that correspond to a certain cluster are classified as to cause the incidence level stays in the cluster. Under

Fig. 5 Panel (a) presents the computation result of Geary's autocorrelation coefficient. Panel (b) displays the Fourier amplitudes related to the frequencies of the index data, with the leading frequencies (in red) correspond to the period of 52 weeks, 227.5 weeks, and 32.5 weeks, respectively. Panel (c) depicts the computed best (in black) and average objective function value from (19) using a genetic algorithm. Panel (d) shows the optimal periodic solution from (19) where index 1 to 5 represent North, East, South, West, and Central Jakarta, respectively.

three clusters, this strategy triples the number of explanatory variables, poised to achieve better fittings. The optimal barriers minimizing the mean squared error are sought with the aid of a Particle Swarm Optimizer applied to the corresponding constrained, non-smooth optimization problem. The new clustering-integrated model with the optimal barriers returns a relatively high coefficient of determination. As the classification on the meteorological factors is based on the incidence level, making a prediction with predicted meteorological factors is not straightforward. A fixed point iteration is used in this study to deal with such indefiniteness. The prediction result is shown to underlie unbiased error for the forthcoming 32 weeks at $\alpha = 0.1$ level. We then propose a certain formula for estimating the (regional) risk measure relying only on the incidence data. The handling equation is a vector autoregression model with state-dependent noise. The noise is determined via connecting the model with the stochastic differential equation counterpart. The endemic equilibrium is shown to be stochastically stable, with the regional risk measure either smaller than 1 (low-risk) or greater than 1 (high-risk). Under a test case for the threshold incidence level, we show that North and East Jakarta are relatively of low risk while South, West, and Central Jakarta are relatively of high risk. To see the spatial connection among the districts, we calculate the global and local spatial auto-

correlation coefficient. Our computation shows that all the districts co-vary highly positively in terms of epidemics occurrence, both generally and timely. However, the local government might still need to know which district requires the most attention per week. We then compute the highest peak of the outbreaks from all the districts and study its behavior over time. Using Fourier analysis, we see that the highest peak resembles annual periodicity in the propagation among all the districts. The recurrence of the peak combined with the risk measures suggest that the following districts require the most to the least attention during epidemics and inter-epidemics: Central, South, West, East, and North Jakarta.

We point out several limitations of this study. First, the population data in Tab. 1 show an increasing pattern in time due to imports, with an anomaly in the year 2011 as the numbers suddenly jump high, especially in Central Jakarta. We failed to find the reason why. **Second, our study was not compared with the usual treatment of dropping time points corresponding to zero incidence levels. We based our reasoning for doing this on having limited data points on the testing window, which will return weak prediction power.** Third, since the final model only incorporates meteorological factors and autoregressive terms, the prediction accuracy highly depends on those entities. Moreover, we observe that the prediction outcome only affords unbiased error, and

no essential information regarding absolute error is required since in our case, the explanatory variables are covariance-stationary, i.e., the mean and variance of each time series do not change over time. To maximize the prediction power, another prediction of incidence levels from medical experts combined with predicted meteorological factors can be included as a certain regulator for the updated epidemic alerts. Fourth, both the simple and clustering-integrated model return heteroscedastic error. Consequently, the error is steered by some of the explanatory variables, and the prediction can be harmed by possible irregularities in the predicted values of those variables. It is the nature of the disease to interplay with various confounding factors other than climate, for example, the socio-economic elements inspired by several studies mentioned in the introduction part. A future study using such additional instrumental variables is thus needed to clear up the heteroscedasticity issue. Fifth, the stochastic stability of the endemic equilibrium from the vector autoregression model is independent of the regional risk measures, but the latter depend on the threshold incidence level of d_{sep} . Such a threshold d_{sep} should carefully be estimated to minimize false alarms that would inflict the epidemic response capabilities of certain districts.

Acknowledgements DA has financially been supported by University of Indonesia through QQ Research Grant scheme, 2019 [No. NKB-0268/UN2.R3.1/HKP.05.00/2019] and MF by the Indonesia Ministry of Research and Technology through PMDSU Program [No. 1511/E4.4/2015].

Conflict of interest

The authors declare no conflict of interest existing in the submission of the manuscript.

References

- Allen LJS, Burgin AM (2000) Comparison of deterministic and stochastic SIS and SIRS models in discrete time. *Mathematical Biosciences* 163:1–33
- de Almeida Costa EAP, de Mendonca Santos EM, Correia JC, de Albuquerque CMR (2010) Impact of small variations in temperature and humidity on the reproductive activity and survival of *Aedes aegypti* (Diptera, Culicidae). *Revista Brasileira de Entomologia* 54(3):488–493
- Anselin L (1995) Local Indicators of Spatial Association – LISA. *Geographical Analysis* 27(2):93–115
- Arcari P, Tapper N, Pfueller S (2007) Regional variability in relationships between climate and dengue/DHF in Indonesia. *Singapore Journal of Tropical Geography* 28:251–272
- Astuti EP, Dhewantara PW, Prasetyowati H, Ipa M, Herawati C, Hendrayana K (2019) Paediatric dengue infection in Cirebon, Indonesia: a temporal and spatial analysis of notified dengue incidence to inform surveillance. *Parasites & Vectors* 12(1):186
- van den Bergh F (2002) An Analysis of Particle Swarm Optimizers. PhD thesis, Department of Computer Science, University of Pretoria, Pretoria, South Africa
- van den Bergh F, Engelbrecht A (2006) A study of particle swarm optimization particle trajectories. *Information Sciences* 176:937–971
- van den Bergh F, Engelbrecht AP (2001) Effects of swarm size on cooperative particle swarm optimisers. In: Spector L, Goodman ED, Wu A (eds) *GECCO'01: Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, Morgan Kaufmann Publishers Inc., San Francisco, CA, United States, pp 892–899
- Berman A, Plemmons RJ (1994) *Nonnegative Matrices in the Mathematical Sciences*. Society for Industrial and Applied Mathematics, Philadelphia, USA
- Bouzzid M, Colón-González FJ, Lung T, Lake IR, Hunter PR (2014) Climate change and the emergence of vector-borne diseases in Europe: case study of dengue fever. *BMC Public Health* 14:781–12
- Brits R, Engelbrecht A, van den Bergh F (2002) A niching particle swarm optimizer. In: Wang L (ed) *Proceedings of the Fourth Asia-Pacific Conference on Simulated Evolution and Learning*, Nanyang Technological University, School of Electrical & Electronic Engineering, Singapore, pp 692–696
- Byttebier B, Majo MSD, Fischer S (2014) Hatching response of *Aedes aegypti* (Diptera: Culicidae) eggs at low temperatures: Effects of hatching media and storage conditions. *Journal of Medical Entomology* 51(1):97–103
- Cameron AC, Trivedi PK (2010) *Microeconometrics Using Stata*, Revised Edition. Stata Press, College Station, Texas, USA
- Caswell H (2000) *Matrix Population Models: Construction, Analysis, and Interpretation*, 2nd edn. Sinauer Associates Inc., Sunderland, Massachusetts, USA
- Chen EH (1971) The power of the Shapiro-Wilk W test for normality in samples from contaminated normal distributions. *Journal of the American Statistical Association* 66(336):760–762
- Chen SS, Gopinath RA (2000) Gaussianization. In: Leen TK, Dietterich T, Tresp V (eds) *NIPS'00: Proceedings of the 13th International Conference on Neural Information Processing Systems*, MIT Press, Cambridge, MA, United States, pp 402–408
- Chien LC, Yu HL (2014) Impact of meteorological factors on the spatiotemporal patterns of dengue fever incidence. *Environment International* 73:46–56
- Christophers SR (1960) *Aedes aegypti – The Yellow Fever Mosquito: Its Life, Bionomics, and Structure*. Cambridge University Press, London, UK

- Cliff AD, Ord JK (1981) *Spatial Processes*. Pion, London, UK
- Davidson R, MacKinnon JG (1993) *Estimation and Inference in Econometrics*. Oxford University Press, UK
- Deng SQ, Yang X, Wei Y, Chen JT, Wang XJ, Peng HJ (2020) A review on dengue vaccine development. *Vaccines* 8(1):63–13
- Fakhrudin M, Putra PS, Wijaya KP, Sopaheluwakan A, Satyaningsih R, Komalasari KE, Mamenun, Sumiati, Indratno SW, Nuraini N, Götz T, Soewono E (2019) Assessing the interplay between dengue incidence and weather in Jakarta via a clustering integrated multiple regression model. *Ecological Complexity* 39:100768–8
- Farrar DE, Glauber RR (1967) Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics* 49(1):92–107
- Geary RC (1954) The contiguity ratio and statistical mapping. *The Incorporated Statistician* 5(3):115–146
- Gujarati DN, Porter DC (2008) *Basic Econometrics*, 5th edn. McGraw Hill Book Co., New York, USA
- Hanusz Z, Tarasinska J, Zielinski W (2016) Shapiro-Wilk test with known mean. *REVSTAT-Statistical Journal* 14(1):89–100
- Harrington LC, Fleisher A, Ruiz-Moreno D, Vermeylen F, Wa CV, Poulson RL, Edman JD, Clark JM, Jones JW, Kitthawee S, et al. (2014) Heterogeneous feeding patterns of the dengue vector, *Aedes aegypti*, on individual human hosts in rural Thailand. *PLoS Neglected Tropical Diseases* 8(8):e3048–16
- Haryanto B (2016) Health adaptation scenario and dengue fever vulnerability assessment in Indonesia. In: Akhtar R (ed) *Climate Change and Human Health Scenario in South and Southeast Asia*, *Advances in Asian Human-Environmental Research*, Springer, Cham, Switzerland, pp 221–236
- Irawan N, Pragesari NN (2014) *Statistik Komuter Jabodetabek: Hasil Survei Komuter Jabodetabek 2014*. Statistics Indonesia, Jakarta, Indonesia
- Jansen CC, Beebe NW (2010) The dengue vector *Aedes aegypti*: what comes next. *Microbes and Infection* 12(4):272–279
- Johnston J (1972) *Econometric Methods*, 2nd edn. McGraw Hill Higher Education
- Juliano SA, O'Meara GF, Morrill JR, Cutwa MM (2002) Desiccation and thermal tolerance of eggs and the coexistence of competing mosquitoes. *Oecologia* 130:458–469
- Kakarla SG, Caminade C, Mutheneni SR, Morse AP, Upadhyayula SM, Kadiri MR, Kumaraswamy S (2019) Lag effect of climatic variables on dengue burden in India. *Epidemiology and Infection* 147:e170(1)–10
- Kloeden PE, Platen E (1992) *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, Berlin–Heidelberg, Germany
- Kusriastuti R, Sutomo S (2005) Evolution of dengue prevention and control programme in Indonesia. *Dengue Bulletin* 29:1–7
- Lowe R, Bailey TC, Stephenson DB, Graham RJ, Coelho CA, Carvalho MS, Barcellos C (2011) Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. *Computers & Geosciences* 37:371–381
- Mansfield ER, Helms BP (1982) Detecting multicollinearity. *The American Statistician* 36(3a):158–160
- McCormack RK, Allen LJS (2006) Multi-patch deterministic and stochastic models for wildlife diseases. *Journal of Biological Dynamics* 1:63–85
- Moran PA (1950) Notes on continuous stochastic phenomena. *Biometrika* 37(1/2):17–23
- Morin CW, Comrie AC, Ernst K (2013) Climate and dengue transmission: evidence and implications. *Environmental Health Perspectives* 121(11–12):1264–1272
- Ogden NH, Lindsay LR (2016) Effects of climate and climate change on vectors and vector-borne diseases: ticks are different. *Trends in Parasitology* 32(8):646–656
- Ramachandran VG, Roy P, Das S, Mogha NS, Bansal AK (2016) Empirical model for estimating dengue incidence using temperature, rainfall, and relative humidity: a 19-year retrospective analysis in East Delhi. *Epidemiology and Health* 38:e2016052–8
- Ramadona AL, Lazuardi L, Hii YL, Holmner A, Kusnanto H, Rocklöv J (2015) Prediction of dengue outbreaks based on disease surveillance and meteorological data. *PloS ONE* 11(3):e0152688–18
- Rana S, Jasola S, Kumar R (2011) A review on particle swarm optimization algorithms and their applications to data clustering. *Artificial Intelligence Review* 35:211–222
- Schmidheiny K (2019) *Panel Data: Fixed and Random Effects*. Universität Basel, Basel, Switzerland
- Shi Y, Liu X, Kok SY, Rajarethinam J, Liang S, Yap G, Chong CS, Lee KS, Tan SS, Chin CKY, Lo A, Kong W, Ng LC, Cook AR (2016) Three-month real-time dengue forecast models: An early warning system for outbreak alerts and policy decision support in Singapore. *Environmental Health Perspectives* 124(9):1369–1375
- Spiegel J, Bennett S, Hattersley L, Hayden MH, Kitayapong P, Nalim S, Wang DNC, Zielinski-Gutiérrez E, Gubler D (2005) Barriers and bridges to prevention and control of dengue: the need for a social-ecological approach. *EcoHealth* 2(4):273–290
- Suwandono A, Kosasih H, Kusriastuti R, Harun S, Ma'roef C, Wuryadi S, Herianto B, Yuwono D, Porter KR, Beckett CG, et al. (2006) Four dengue virus serotypes found circulating during an outbreak of dengue fever and dengue haemorrhagic fever in Jakarta, Indonesia, during 2004. *Transactions of the Royal Society of Tropical Medicine*

- and Hygiene 100(9):855–862
- Trelea IC (2003) The particle swarm optimization algorithm: convergence analysis and parameter selection. *Information Processing Letters* 85:317–325
- Upton GJG, Fingleton B (1985) *Spatial Data Analysis by Example, Volume 1: Point Pattern and Quantitative Data*, 1st edn. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section (Book 182, Wiley
- Wang D, Tan D, Liu L (2018) Particle swarm optimization algorithm: an overview. *Soft Computing* 22(2):387–408
- Wang X, Tang S, Wu J, Xiao Y, Cheke RA (2019) A combination of climatic conditions determines major within season dengue outbreaks in Guangdong Province, China. *Parasites & Vectors* 12(45):1–10
- Wijaya KP, Aldila D, Schäfer LE (2019) Learning the seasonality of disease incidences from empirical data. *Ecological Complexity* 38:83–97
- Wilder-Smith A (2020) Dengue vaccine development: status and future. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz* 63(1):40–44
- Willis CE, Perlack RD (1978) Multicollinearity: effects, symptoms, and remedies. *Journal of the Northeastern Agricultural Economics Council* 7(1):55–61
- Withanage GP, Viswakula SD, Gunawardena YINS, Hapugoda MD (2018) A forecasting model for dengue incidence in the District of Gampaha, Sri Lanka. *Parasites & Vectors* 11:262–10
- Xu L, Stige LC, Chan KS, Zhoud J, Yang J, Sang S, Wang M, Yang Z, Yan Z, Jiang T, Lu L, Yue Y, Liu X, Lin H, Xu J, Liu Q, Stenseth NC (2017) Climate variation drives dengue dynamics. *Proceedings of the National Academy of Sciences of the United States of America* 114(1):113–118
- Yu HL, Yang SJ, Yen HJ, Christakos G (2011) A spatio-temporal climate-based model of early dengue fever warning in southern Taiwan. *Stochastic Environmental Research and Risk Assessment* 25:485–494