MODELING DENGUE OUTBREAK DATA USING NONLINEAR MIXED EFFECTS MODEL

Carlos Rafael Sebrango Rodríguez[1]*, Lizet Sánchez Valdés[2]**, Ziv Shkedy***
*University of Sancti Spiritus "José Martí Pérez", Cuba
** Center of Molecular Immunology, Cuba
***Center for Statistics, Hasselt University, Belgium ([3])

ABSTRACT
In recent years, there has been an increased interest in using statistical models for analysis of single dengue outbreaks based on the reported cumulative cases. Sometimes this type of data is collected for all urban areas in a particular region. Modeling in order to estimate epidemiological parameters is usually performed for each area separately, but when the interest lays on estimating the average behavior of a particular area in the population, and variability among and within areas, a nonlinear mixed effects model is recommended. In this research, we describe two approaches that provide estimates of three key epidemiological parameters: the turning point, the final size of outbreak, and the basic reproduction numberR_0, using nonlinear models. The first approach consists of fitting an individual nonlinear model for each area separately. In the second method, we use a nonlinear mixed effects model, which accounts for heterogeneity between areas. In both approaches, the Richards model was used as mean structure. The proposed methods are applied to data of seven Primary-Health Care Areas of Plaza municipality, Havana City, Cuba during 2006 dengue outbreak.

KEYWORDS: Dengue outbreak; parameter estimate; Nonlinear mixed effects model; Richards model; heterogeneity

MSC: 62P10

RESUMEN
Recientemente ha existido un interés creciente en el uso de métodos estadísticos para el análisis de brotes de dengue basado en casos acumulados reportados. Algunas veces este tipo de datos se colecciona para todas las áreas urbanas en una región en particular. La modelación, para estimar parámetros epidemiológicos, se realiza usualmente para cada área por separado, pero cuando el interés consiste en estimar el comportamiento promedio de un área en particular en la población, y la variabilidad entre y dentro de las áreas, se recomienda un modelo no lineal con efectos mixtos. En esta investigación, se describen dos enfoques que proporcionan estimaciones de tres parámetros epidemiológicos primarios: el acmé de la epidemia, el tamaño final y el número reproductivo básico R_0. El primer enfoque consiste en ajustar un modelo no lineal individual para cada área por separado. En el segundo enfoque se utiliza un modelo no lineal con efectos mixtos, el cual tiene en cuenta la heterogeneidad entre las áreas. En ambos enfoque, se utilizó el modelo de Richards como estructura de la media. Los métodos propuestos son aplicados a los datos de siete áreas de salud primarias del municipio Plaza durante el brote de dengue del 2006.

1. INTRODUCTION

Dengue is a mosquito-borne viral infectious disease that causes significant epidemic outbreaks, particularly in tropical and subtropical areas [1]. In recent years, there has been an increased interest in using statistical models for analysis of single dengue outbreaks [2] based on the reported cumulative cases. These models capture the behavior of the outbreak, and also facilitate the estimation of important epidemiological parameters.

Parameters estimation is a key step in modeling epidemiological processes [3] and it provides a useful tool to study the impact of intervention and control measures. Among the most important epidemiologic parameters are the turning point, i.e. the point in time at which the rate of accumulation changes from increasing to decreasing or the infection point of the logistic (S-shaped) curve in a single epidemic outbreak, the final size of epidemic and the basic reproduction number R_0, defined as the number of secondary infections that arise from a typical primary case in a completely susceptible population [4, 5].

Modern techniques for parameter estimation of mechanistic models have gained popularity. A mechanistic model is one where the basic elements of the model have a direct correspondence to the underlying mechanisms in the system being modeled. However, maximum likelihood fitting of phenomenological models remains

[1] sebrango@uniss.edu.cu
[2] lsanchez@cim.sld.cu
[3] ziv.shkedy@uhasselt.be

important due to its simplicity, to the difficulty of using modern methods in the context of limited data, and to the fact that there is not always enough information available to choose an appropriate mechanistic model [6]. In particular, Hsieh et al. [4, 5] proposed to use a nonlinear model, the Richards model, to estimate these three key epidemiological parameters. The Richards model considers only the cumulative infective population size with saturation in growth as the outbreak progresses. The basic premise of the Richards model is that the incidence curve consists of a single peak of high incidence, resulting in an S-shaped epidemic curve and a single turning point of the outbreak [4, 5].

Sometimes, the reported cumulative cases are collected for all urban areas in a particular region, and modeling is usually performed for each area individually. However when the interest lays on estimating the average behavior of a particular area in the population and variability among and within areas, a nonlinear mixed effects model is recommended [7, 8].

In this research, we describe two approaches that provide estimates of three key epidemiological parameters: the turning point, the final size of outbreak, and the basic reproduction number R_0, using nonlinear models. The first approach consists of fitting an individual nonlinear model for each area separately. In the second method, we use a nonlinear mixed effects model, which accounts for heterogeneity between areas. The proposed methods are applied to data of seven Primary-Health Care Areas of Plaza municipality, Havana City, Cuba during 2006 dengue outbreak.

2. METHODS

Richards model for specific area

Let Y_{ij} a random response, which represents the cumulative number of reported cases in the health area i at timet_j. In this study, we consider the assumption that the cumulative number of reported cases, the response (Y_{ij}), are normally distributed with mean $\mu(t_j,\theta_i)$ and variance σ^2, e.g. $Y_{ij}\sim N(\mu(t_j,\theta_i),\sigma^2)$ where $\mu(t_j,\theta_i)$ is the known nonlinear function; $t_j$is the regressor variable and θ_i the parameter vector, which needs to be estimated.

The mean structure $\mu(t_j,\theta_i)$, which describes the relationship between the cumulative number of reported cases and the time in weeks, is the Richards model [4, 5, 9], and can be expressed as follow:

$$\mu(t_j,\theta_i) = \frac{\alpha_i}{\left[1+k_i.e^{-k_i\gamma_i(t_j-\eta_i)}\right]^{\frac{1}{k_i}}} \quad i=1,\cdots,n \quad j=1,\cdots,J \tag{1}$$

Where $\theta_i = (\alpha_i, k_i, \gamma_i, \eta_i)$ is a health area specific parameter vector to be estimated. The parameter α_i is the final size of the epidemic, γ_i is the growth rate, η_i is the turning point of outbreak and k_i is the exponent of deviation from the standard logistic curve.

Nonlinear Mixed effects Model

The corresponding nonlinear mixed effects model for the cumulative number of reported cases Y_{ij} in the health area i at time t_j is

$$Y_{ij} = \frac{\alpha_i}{\left[1+k_i.e^{-k_i\gamma_i(t_j-\eta_i)}\right]^{\frac{1}{k_i}}} + \varepsilon_{ij} \tag{2}$$

The health area-specific parameter vector is modeled as:

$$\theta_i = X_i\theta + Z_ib_i \tag{3}$$

Here, θ is a fixed parameter vector, b_i is a health area-specific random effects vector, and X_i and Z_i are known design matrices for the fixed effects θ and the random effectsb_i, respectively:

$$X_i = Z_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \theta = \begin{pmatrix} \alpha \\ k \\ \gamma \\ \eta \end{pmatrix} \text{ and } b_i = \begin{pmatrix} b_{i1} \\ b_{i2} \\ b_{i3} \\ b_{i4} \end{pmatrix} \tag{4}$$

It follows from equation (3) that the health area-specific parameter vector can be expressed as

$$\begin{pmatrix} \alpha_i \\ k_i \\ \gamma_i \\ \eta_i \end{pmatrix} = \begin{pmatrix} \alpha + b_{i1} \\ k + b_{i2} \\ \gamma + b_{i3} \\ \eta + b_{i4} \end{pmatrix} \tag{5}$$

The random effect b_i are assumed to be normally distributed as $b_i \sim N(0, \Psi)$ and the within-group error $\varepsilon_{ij} \sim N(0, \sigma^2)$. A general positive-definite matrix is used to represent the random-effects variance–covariance structure Ψ.

Basic reproduction number $\mathbf{R_0}$

The initial exponential growth rate of an epidemic is an important measure of disease spread, and is commonly used to infer the basic reproduction numberR_0. R_0 is determined by γ and the distribution of the generation interval, i.e. the time between a case and the secondary cases resulting from it.

For an infection where all secondary infections are exactly equal to the mean generation interval T, the distribution conforms to a so-called delta distribution. It has been shown mathematically that, given the growth rate γ, the expression $R_0 = e^{\gamma T}$ provides the upper bound of the basic reproduction number regardless of the distribution of the generation interval used [10].

To take into account the extrinsic and intrinsic incubation periods as well as the duration of viraemia, we use an estimated generation time of $T = 19$ days or $T = \frac{19}{7}$ weeks with a range of 16–34 days [4]. The ESTIMATE statement from NLMIXED procedure in SAS [11] was used to compute approximate standard errors for the R_0 estimates using the delta method. The approximate standard error for R_0 estimate using the delta method is given by the expression $SE(R_0) \approx Te^{\gamma T} SE(\gamma)$.

3. APPLICATION TO THE DATA

Dengue outbreak data

The research was conducted in all Primary-Health Care Areas of the municipality of Plaza, a part of Havana City, where cases were reported during 2006 dengue outbreak. The seven areas are numbered as follows: 1 ("15 y 18"), 2 ("19 de Abril"), 3 ("Corynthia"), 4 ("Moncada"), 5 ("Puentes Grandes"), 6 ("Plaza") and 7 ("Rampa"). The data used in this study is the weekly distribution of confirmed Dengue cases per Health Area by date of onset of symptoms. The weekly data were converted into cumulative case curves.

Model for specific health area

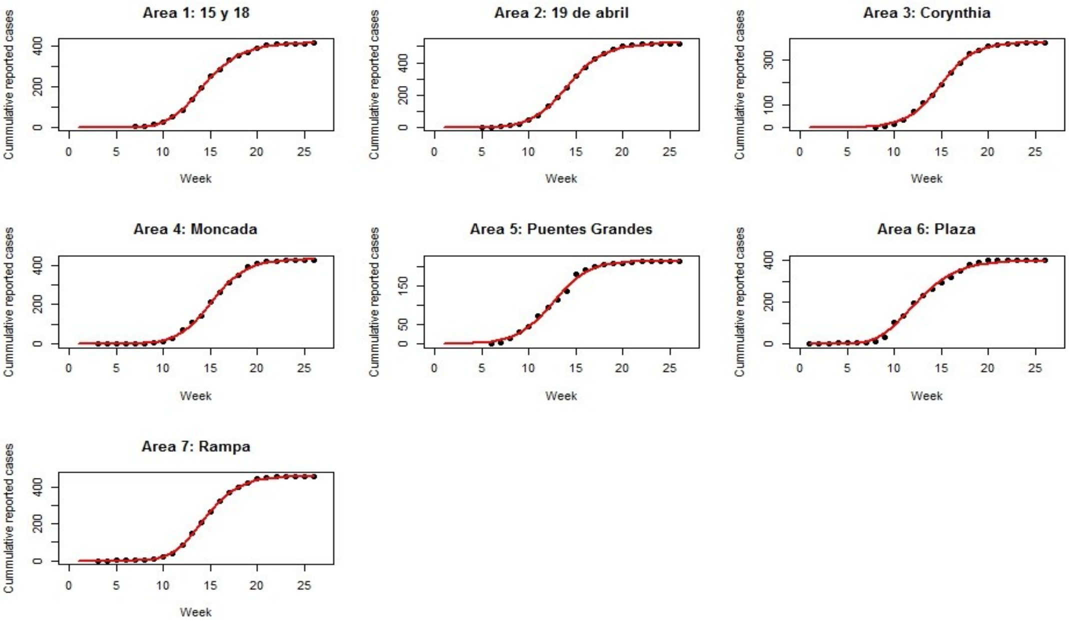
The nonlinear Richards model in equation (1) was fitted for each health area separately by using NLMIXED procedures in SAS [11] and nlsList function from nlme package in R software [12]. Figure 1 shows the individual fitted models of the seven Primary-Health Care Areas. Table 1 shows the health area-specific ML parameter estimates (and 95% CI). A remarkable variability is observed in the individual parameter estimates.

Table 1: Parameter estimates of Richards model and R_0 estimate for specific health area.

Áreas	α	η	γ	k	R_0
15 y 18	420.05 (415.36,424.75)	13.76 (13.53,13.99)	1.07 (0.65,1.48)	0.40 (0.21,0.58)	18.22 (0*,38.94)
19 de Abril	524.14 (520.35,527.93)	13.87 (13.70,14.03)	0.73 (0.60,0.85)	0.64 (0.49,0.79)	7.17 (4.75,9.59)
Corynthia	381.09 (374.98,387.19)	14.79 (14.42,15.17)	0.59 (0.42,0.77)	0.92 (0.50,1.34)	5.00 (2.60,7.40)
Moncada	431.13 (425.10,437.16)	14.92 (14.63,15.22)	0.69 (0.49,0.89)	0.74 (0.44,1.04)	6.48 (3.02,9.94)
P. Grandes	215.32 (211.11,219.52)	12.58 (11.97,13.18)	0.51 (0.31,0.72)	1.08 (0.38,1.79)	4.03 (1.77,6.30)
Plaza	399.14 (387.35,410.94)	11.73 (11.39,12.07)	1.12 (0.56,1.68)	0.36 (0.14,0.57)	20.67 (0*,52.05)
Rampa	460.88 (456.82,464.93)	14.01 (13.81,14.21)	0.96 (0.68,1.25)	0.50 (0.31,0.68)	13.66 (3.15,24.17)

* Max (0, lower bound)

Although the models for specific health areas fit the data well, from our point of view, this approach has limitations. It uses 28 coefficients to represent the individual cumulative cases profiles and does not take into account the obvious similarity among the individual curves, indicated in Figure 1. This approach (individual nonlinear model for each area separately) is useful when one is interested in modeling the behavior of a particular, fixed set of areas, but it is not adequate when the areas are regarded as sample from a (perhaps hypothetical) population and inference should focus on this population.


Area 1: 15 y 18

Area 2: 19 de abril

Area 3: Corynthia

Area 4: Moncada

Area 5: Puentes Grandes

Area 6: Plaza

Area 7: Rampa

Figure 1: Individual fitted (red line) and observed (dot) cumulative number of cases of the seven Primary-Health Care Areas.

Nonlinear Mixed effect model

As in this case, the interest lays on estimating the average behavior of a health area in the population and the variability among and within health areas, a mixed-effects model is developed. A crucial step in the model-building of mixed-effects models is deciding which of the coefficients in the model need random effects to account for their between-areas variation and which can be treated as purely fixed effects.

Our modeling strategy was to fit different models taking into account in each model different parameters with random effects. Within the strategy, no parameters with random effects were considered at first. Then, all combinations of parameters with random effects were used. In all cases, the α parameter, the final size of the epidemic, was considered as random effect for its great variability. Table 2 shows the Akaike information criteria (AIC) and Bayesian information criteria (BIC) for nonlinear mixed models with all combinations of parameter with random effects, and parameter estimates for the best fitted nonlinear mixed effects model by using NLMIXED procedures in SAS [11] and nlme() function from nlme package in R software [12].

Table 2: AIC and BIC for nonlinear mixed models with different random effects (a), parameter estimates for nonlinear mixed model with lowest AIC and BIC (b) and the correlation matrix of random effects (c).

a				b		c		
Random effects	**AIC**	**BIC**		**Parameters**	**Estimate (se)**			
-	1730.0	1745.3		**Fixed effects**				
α	1434.7	1453.0		α	406.76 (33.31)			
α, η	1135.6	1160.0		k	0.5176 (0.062)			
α, γ	1410.1	1434.5		γ	0.8691 (0.080)			
α, k	1438.7	1463.1		η	13.5559 (0.39)			
$\boldsymbol{\alpha, \eta, k}$	**1124.2**	**1157.7**		**Residual variance**				
α, η, γ	1127.8	1161.3		σ^2	38.3170 (4.65)			
α, k, γ	1307.4	1340.9		**Random effects**		**Correlation Matrix**		
α, η, γ, k	1132.8	1177.93		$StdDev(b_{i1})$	88.0630 (23.55)		α	k
				$StdDev(b_{i2})$	0.03025 (0.012)	k	0.262	
				$StdDev(b_{i4})$	1.0139 (0.27)	η	0.559	0.934

The lowest AIC and BIC was obtained for the model with the parameters α, k and η as random effects, i.e. considering only the γ as a fixed parameter. As the key epidemiological parameter R_0 is derived from γ, then R_0 parameter estimate, and therefore the transmission, would be the same for each health area if this model is considered, i.e. $R_0 = 10.58$ (4.22,16.94) . Although the best fit is obtained for this model, it might make no sense from an epidemiological point of view, in which case it is better for inference to use the model when the parameters with random effects are α,η and γ, having an AIC = 1127.8 and BIC = 1161.3 very similar to the above.

The near-zero estimate for standard deviation of the k random effect suggests that this term could be dropped from the model. The remaining estimated standard deviations suggest that the other random effects should be kept in the model. In order to test if k random effect can be removed from the model, a likelihood ratio test was performed. The results are shown in table 3. The inclusion of k random effects causes a significant improvement in the log-likelihood.

Table 3: Likelihood ratio test for the model with α,η as random effects versus the model with α,η,k as random effects.

Model	**AIC**	**BIC**	**logLik**	**Test**	**L. Ratio**	***p*-value**
Model 1 (α,η)	1135.607	1060.006	-559.80			
Model 2 (α,η,k)	1124.182	1157.731	-551.09	1 vs 2	17.42	6e-04

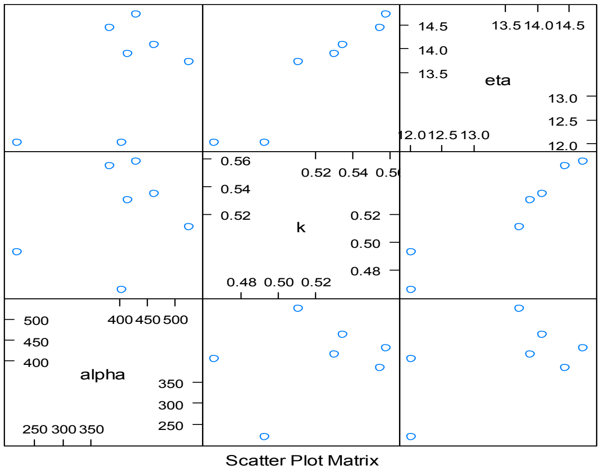


Figure 2: Pairs plot for the random-effects estimate corresponding to the model with α,η,k as random effects.

The estimated correlation of 0.934 between $\eta$and k suggests that the estimated variance-covariance matrix is ill-conditioned and that the random-effects structure may be over-parameterized. The scatter-plot matrix of the estimated random effects provides a useful diagnostic plot for assessing over-parameterization problems. The nearly perfect alignment between η and k random effects (Figure 2) further indicates that the model is over-parameterized. The large correlation between k and η random effects and the small correlation between these random effects and the α random effect suggest that a block-diagonal Ψ could be used to represent the variance–covariance structure of the random effects. In order to test if a block-diagonal Ψ could be used to represent the variance–covariance structure of the random effects, a likelihood ratio test was performed (Table 3). The large *p*-value for the likelihood ratio test and the smaller values for BIC corroborated the block-diagonal variance–covariance structure.

Table 4: Likelihood ratio test for the model with variance-covariance matrix Ψ positive-definite versus the model with variance-covariance matrix Ψ block diagonal.

Model	**AIC**	**BIC**	**logLik**	**Test**	**L. Ratio**	***p*-value**
Model 1 (Ψ positive-definite)	1124.182	1157.731	-551.09			
Model 2 (Ψ Block diagonal)	1124.365	1151.814	-553.18	1 vs 2	4.18	0.1235

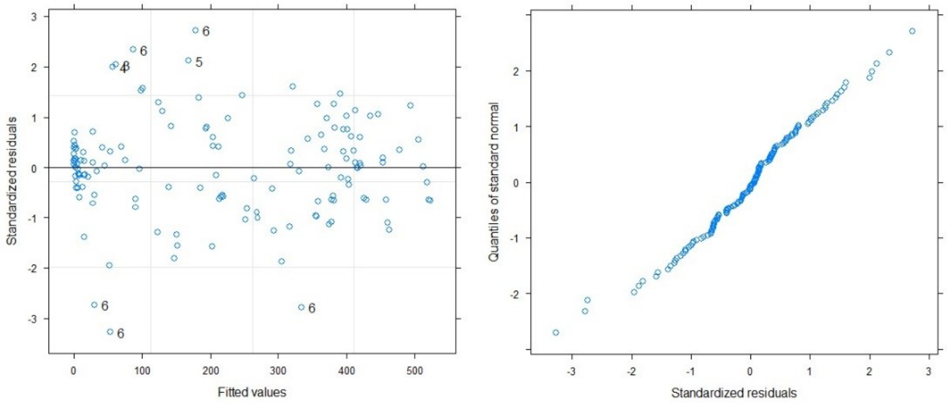


Figure 3: Scatter plot of standardized residual versus fitted values (Left panel) and normal plot of standardized residual (Right panel) for the model with variance-covariance matrix Ψ Block diagonal.

The plot of standardized residuals versus the fitted values corresponding to the model with α, η, k as random effects and with variance-covariance matrix Ψ block diagonal presented in the left panel of Figure 3, shows that the residuals are distributed symmetrically around zero with an approximately constant variance. It does not indicate any departures from the nonlinear mixed effects model assumptions, except for some possible outlying observations for the Health area 6 (Plaza). The normal probability plot of the standardized residual, shown in the right panel in Figure 3 does not indicate any violations of the normality assumption for the within-group errors.

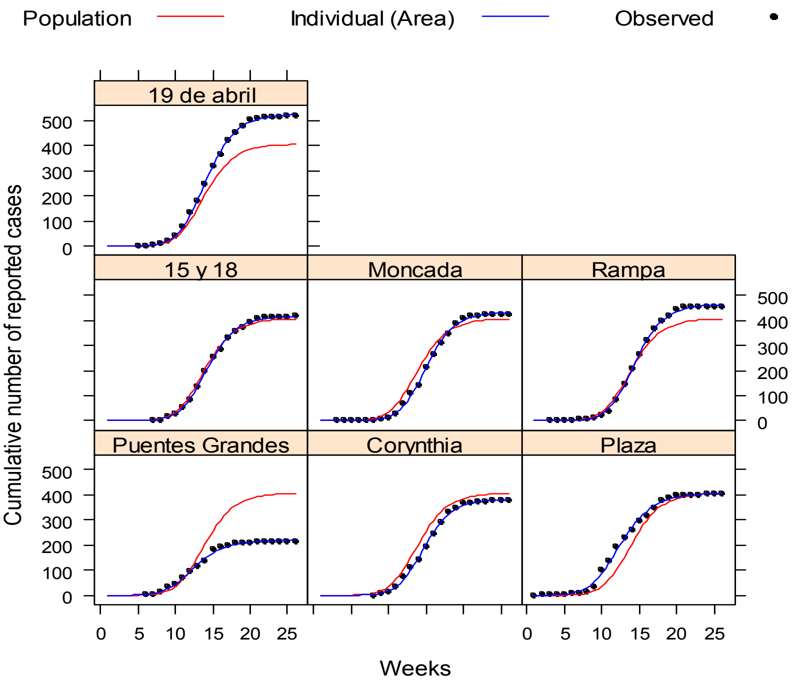


Figure 4: Population predictions (population), within-group predictions (area), and observed cumulative number of cases (dot) versus time in week, for the best-fitted mixed model.

The plot of the augmented predictions in Figure 4 gives a final assessment of the adequacy of this model. For comparison and to show how individual effects are accounted for in the nonlinear mixed effects model, both the population predictions (corresponding to random effects equal to zero) and the within-group predictions (obtained using the estimated random effects) are displayed. Note that the within-group predictions are in close agreement with the observed cumulative cases, illustrating that the nonlinear mixed effects model can accommodate individual effects.

4. CONCLUSIONS

Modeling dengue outbreak data collected for all urban areas in a particular region is usually performed for each area separately. In this study, we described two approaches to estimate three key epidemiological parameters: turning point, final size and the basic reproduction number. The first approach consisted in fitting an individual nonlinear model for each area separately and the second approach proposed to use a nonlinear mixed effects model. Both approaches were applied to data of seven Primary-Health Care areas of Plaza Municipality, Havana, Cuba, during 2006 dengue outbreak.

For this particular setting, the second approach is highly recommended because the areas are regarded as sample from a population and it does not ignore variability among and within areas. However, the first approach constitutes a powerful tool for the model building of the second approach because the individual estimates can suggest the type of random effects structure to use and also provide starting values for the parameters.

The best-fitted nonlinear mixed effects models were obtained for the models with the parameters α, k, η and α, η, γ as random effects, respectively. Although the AIC and BIC of these two models are very similar, they have different interpretations. In the first model, the growth rate parameter γ is regarded as fixed parameter indicating that the transmission is the same in all areas. This might make no sense from an epidemiological point of view because the transmission depends on mosquito's population, which should not be the same for all areas. The second model considers that the three key epidemiological parameters vary among areas, which makes sense from epidemiological point of view. For that reason, it is better for inference to use the second model.

These modeling approaches could be a valuable tool to public health policymakers for responding to future dengue outbreaks, because they give estimates of key epidemiological parameters like turning point, which could provide vital information pertaining to the changing trends of the epidemic and possibly indicating changes in intervention and control.

RECEIVED: MAY, 2016
REVISED: NOVEMBER, 2016

REFERENCES

[1] CHOWELL, G., FUENTES, R., OLEA, A., AGUILERA, X., NESSE, H. and HYMAN, J. M. (2013): The Basic Reproduction Number R_0 and Effectiveness of Reactive Interventions during Dengue Epidemics: The 2002 Dengue Outbreak in Easter Island, Chile. **Math Biosci Eng.**; 10(0): 1455–1474.

[2] NISHIURA, H. (2006) Mathematical and statistical analyses of the spread of dengue. **Dengue Bulletin** 30: 51–67.

[3] KLECZKOWSKI, A. and GILLIGAN, C. A. (2007): Parameter estimation and prediction for the course of a single epidemic outbreak of a plant disease, **J. R. Soc. Interface** 4, 865-877.

[4] HSIEH, Y.H. and MA, S. (2009): Intervention measures, turning point, and reproduction number for dengue, Singapore, 2005, **American Journal of Tropical Medicine and Hygiene** 80, 66-71.

[5] HSIEH, Y.H., ARAZOZA, H. and LOUNES, R. (2013): Temporal trends and regional variability of 2001-2002 multiwave denv-3 epidemic in Havana City: did hurricane Michelle contribute to its severity? **Tropical Medicine and International Health** 18 (7), 830-838.

[6] MA, J., DUSHOFF, J., BOLKER, B. M. and EARN, D.J.D. (2014): Estimating initial epidemic growth rate, **Bull Math Biol** 76:245-260. DOI 10.1007/s11538-013-9918-2.

[7] BATES, D., PINHEIRO, J. (2000): **Mixed Effects Models in S and Splus**, New York: Springer.

[8] DEMIDENKO, E. (2013): **Mixed models: theory and applications with R**, Second edition, Wiley series in probability and statistics.

[9] RICHARDS, F.J. (1959): A flexible growth function for empirical use. **Journal of Experimental Botany** 10, 290–300.

[10] WALLINGA, J. and LIPSITCH, M. (2007): How generation intervals shape the relationship between growth rates and reproductive numbers. **Proceedings. Biological Sciences** 274, 599–604.

[11] SAS/STAT SOFTWARE, VERSION 9.3 (2011) by **SAS Institute Inc**.

[12] R DEVELOPMENT CORE TEAM (2015). R: a language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. http://www.R-project.org/.