# Quantifying Critical Parameter in Disease Transmission

W. C. Kok and J. Labadin
*Department of Computational Science and Mathematics,*
*Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak,*
*94300, Kota Samarahan, Sarawak, Malaysia.*
*woonchee.kok@gmail.com*

***Abstract*—The values of each parameter introduced in a disease model play important role in providing the prediction of a disease transmission. Some parameters values are easily quantified through collected statistical data usually made available from clinical research. However, there may be some parameters that are not easily found. For such case, the parameters values are estimated through many trial-and-error numerical runs. In this paper, it is shown that a statistical modeling approach coupled with the Maximum Likelihood Estimate method can be used to quantify critical model parameters. A Hand-Foot-Mouth disease (HFMD) model was taken as a case study where infected population data provided by the Sarawak State of Health was fitted onto the Susceptible-Infected-Removal (SIR) model. The concerned parameter is the transmission coefficient of HFMD in the year 2012. Using the mentioned method, it was found that the value for the transmission coefficient of HFMD in 2012 is 1.2654 (CI: 1.15-1.43). It can be concluded that the critical parameter with 95% confidence interval in SIR model has been quantified effectively. Due to the possibility of obtaining other sets of infected population data, a web application called the Disease Modeling Parameter Calculator was developed to assist in estimating the transmission coefficient.**

***Index Terms*—Hand, Foot and Mouth Disease; Maximum Likelihood; Parameter Estimation; Susceptible-Infected-Recovered; Statistical Modeling.**

## I. INTRODUCTION

Infectious disease is one of the top ten causes of death in the world [1-3]. Thus, infectious disease modeling plays a key role in basic science and public policy. Disease models summarize what is known about disease epidemiology, prevention and treatment. Disease modeling is beneficial to clinical practitioner, manufacturers, policy makers and researchers to control or eradicate infectious diseases. To date, public health professionals have significantly increased the usage of disease model to assist public health policy decisions and to explore questions in disease control [4]. Disease researcher and modeler formulate a disease model to identify a disease trends, make general forecast, and estimate the uncertainty in the forecast by synthesizing information from different data sources.

In order to understand the human population distribution and the spread of disease, researcher develop mathematical disease model [5, 6]. In different approach and model, there is a flow of process, computing tools and mathematical model to help in disease modeling. Different modeling approach suits different situation, different fields and different problems. One of the key challenges in building a disease model is in quantifying some parameters that are not easily available.

One of the methods to quantify parameter is the Maximum Likelihood Estimation (MLE). "MLE estimates parameter values that make the observed data the most likely to have happened" [7]. The principle of MLE, originally developed by R. A. Fisher in 1920s, states the "most likely" means that one must seek the value of the parameter vector that maximizes the likelihood function [8, 9]. MLE is asymptotically consistent, as the data size gets larger, the estimated parameters gets closer to the true values and converge to the actual values [7]. MLE method is also asymptotically efficient, for large data size it can generate the most precise estimates compared to others. MLE method is scale free or parameterization invariance which the estimated parameters are not affected by the transformation of variables [10]. The values of fit function are independent with the scale of response data [11]. Apart from that, MLE is reported by many researchers as being an unbiased estimation with large data sizes which is more than 30 samples. Sufficiency is one of the most important properties of MLE [7, 12]. Sufficiency indicates the completeness of the information about the parameters that the researcher is interest in. If there is a sufficient statistic for a parameter, the MLE of the parameter is a function of a sufficient statistic. A sufficient statistic is a statistic that uses all of the information in the sample about the parameter of interest [12]. However, MLE can be biased for small samples when the sample size is less than 30. It requires large data sizes in order to overcome the accuracy issue. Normally in infectious disease modeling, likelihood equation can be very complicated, for example in creating likelihood function or other complex model such as negative log-likelihood function.

The Susceptible-Infected-Recovered (SIR) model is an epidemiological model that computes the number of susceptible, infected and recovered with an infectious disease in a closed population over time. Disease model is governed by fundamental parameters that include transmission coefficient, recovered rate, birth rate and death rate. Unfortunately, not all parameter values are available therefore the researchers need to estimate the parameters. Researchers need to make an initial estimate of the starting values of some parameters for example transmission coefficient. After making an initial guess of the parameter value, the researcher needs to run the computer simulation and a set of numerical results. If the fitted result is not satisfied and not compromised, researchers need to estimate again and run the simulation again until a minimum discrepancy between the actual data and fitted result is obtained. However, this process is time-consuming because

it is an exhaustive search of the parameter space. And it is less likely to get the best-estimated value that is the nearest to the actual data by using such trial-and-error runs.

Most of the times, researchers use trial-and-error method to fit the data in order to obtain the parameter value. The trial-and-error procedure is complex and computationally expensive. It is time-consuming if the estimation of the parameter values with this conventional approach involved large data set [13-15]. Therefore, researchers need an efficient parameter quantification method to address this problem. Statistical modeling approach promotes cheaper computing power, which allows users to quantify critical parameter quickly and easily compared to the trial-and-error approach. In this study, a set of three parameter values was quantified but our analysis will focus only on the transmission coefficient of an SIR model.

## II. METHODS AND MATERIALS

A Hand-Foot-Mouth disease (HFMD) model was taken as a case study where this HFMD mainly affects young children below 10 years old and occurs in clusters or outbreak as it is a highly viral disease. Typical manifestations of HFMD in children include fever, vesicles in the mouth and skin eruptions on hands and feet. HFMD isolates itself every three years since the large outbreak in year 1997. The prediction suggested a large outbreak in year 2015 and has raised public fear and anxiety due to the outbreak [16]. In most cases for disease models, some parameters are required to be estimated for further analysis. This study investigated the value of the transmission coefficient in the years 2010 until 2014 using the constructed mathematical model. The clinical data is provided from the Sarawak State Health Department, which consists of number of patients versus time from year 2010 until 2014. A deterministic SIR model has been chosen to model the spread of HFMD in Sarawak as shown in equations (1), (2) and (3) [17]. There are different variables and parameters in the SIR model and the description of the variables and parameters are shown in Table 1 and 2.

$$\frac{dS}{dt} = \alpha S(t) - \beta I(t)S(t) - \mu_0 S(t) + \delta R(t) \quad (1)$$

$$\frac{dI}{dt} = \beta I(t)S(t) - \gamma I(t) - (\mu_0 + \mu_1)I(t) \quad (2)$$

$$\frac{dR}{dt} = \gamma I(t) - \delta R(t) - \mu_0 R(t) \quad (3)$$

Table 1
Description of variables for SIR model

| Variable | Description |
|---|---|
| $S(t)$ | Number of susceptible at time *t* |
| $I(t)$ | Number of infective at time *t* |
| $R(t)$ | Number of recovered at time *t* |

Table 2
Description of parameters for SIR model

| Parameter | Description |
|---|---|
| $\alpha$ | Natural birth rate |
| $\beta$ | Transmission coefficient |
| $\gamma$ | Rate at which an infectious individual recovers per unit time |
| $\delta$ | Rate at which a recovered individual loses immunity |
| $\mu_0$ | Natural death rate |
| $\mu_1$ | Rate of death caused by the disease |

This study involves in the analysis of the estimated parameter value of the transmission coefficient only. Transmission coefficient is the rate of susceptible population getting infected with HFMD at the rate of $\beta$. Figure 1 outlines the quantification procedures.

Actual disease data in year 2012 comprises of number of weeks and number of cases has been plotted as shown in Figure 2. If there is any unbalance of the data, pre-process of the data can be performed to trim the data set or sampling the data from the whole data set. The data trend and distribution of the actual data in Figure 2 shows that there is no outlier of the HFMD data. Next, we need to set the state variable, which is dynamically varying characteristic of the model that indicates the storage of volume of time varying quantity of interest within the model. Different state variables, taken together maybe used to define the model "state". The state variables are the susceptible, $S(t)$ and the infected $I(t)$ in this study. To run the simulation, we have prescribed the initial values as $I_0 = 1$ and $S_0 = N - I_0$ where *N* is total population [18]. The returned results of the *ode* function are the number of susceptible and number of infected at a given time *t*. The researcher needs to set the initial values to start the computation. We set the initial number of infected, $I_0$ = 1 and the total population, $N$=6580 according to the actual data obtained for year 2012, and thus the initial number of susceptible, $S_0$=6579.

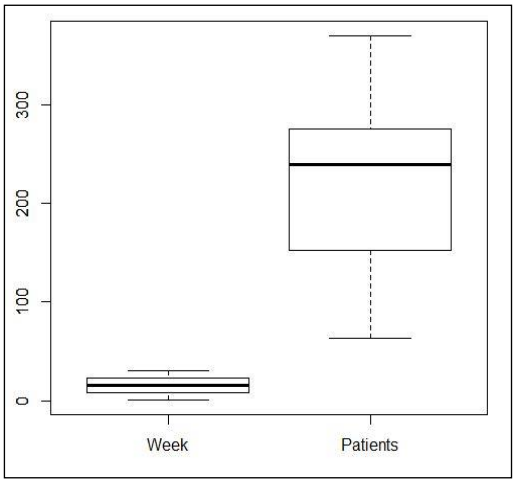


Figure 2: Boxplot of HFMD data in year 2012

Next, researchers need to specify the span of time to run the model. Researchers need to call a *seq* function, which is a built-in function in *R* package. This function can create time sequences of the model. The data contains of 30 weeks and we set the interval between these 30 weeks as 200 intervals. Researchers can decide the number of interval themselves. However, the smaller steps of the time sequence make the result more accurate [20-22]. In other words, small time step sizes are desirable for better accuracy. A system of differential equations can be solved by using *deSolve* function which is the default integration routine in *R* package. Having (1), (2) and (3) solved, the built-in *ode* function in *R* packages takes input of the initial values of the parameters where $\beta = 1$, $\gamma = 0.5$ and $\delta = 1$. $\alpha$, $\mu_0$ and $\mu_1$ in Table 2 are conveniently available from some published literature. In this study, we adopted the parameter values from [17]: $\alpha = 0.02923$, $\mu_0 = 0.01077$ and $\mu_1 = 0.001731$.

Function *ode* returns an object of class *deSolve* with a matrix that contains the values of the state variables at the requested output times. The model function returns outputs number of susceptible ($S$) and number of infected ($I$) at time $t$.

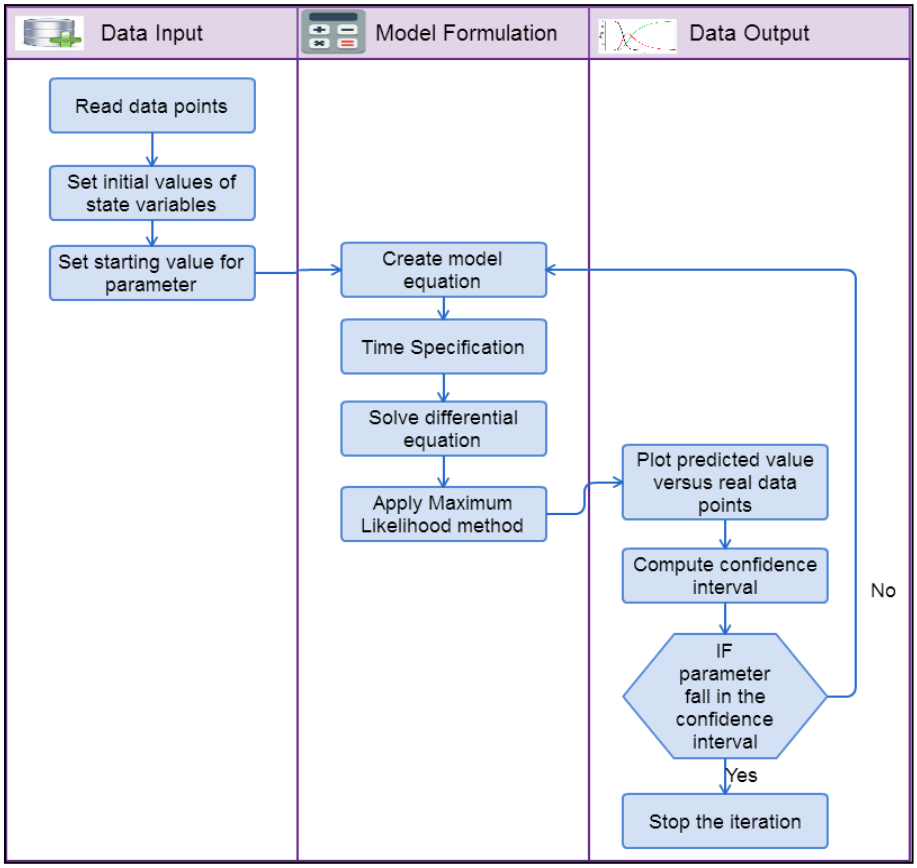


Figure 1: Step to quantify disease model parameter

The *model* function, which is a user-defined function, will return the rate of change and the parameters vector, which are $\beta$, $\gamma$ and $\delta$. Function *ode* returns an object of class *deSolve* with a matrix that contains the values of the state variables at the requested output times. The model function returns outputs number of susceptible ($S$) and number of infected ($I$) at time $t$.

The most important step is the researcher writes a function to return a negative log-likelihood of the data. This is due to the fact that the value of the transmission coefficient (Beta/ $\beta$) must always be a positive value. After a simple transformation of these two parameters, $\beta = e^b$ and $= e^g$ . These $b$ and $g$ values are defined from negative infinity to positive infinity. This helps the numerical algorithms to perform better and provide better result. Theoretically, the idea of finding the maximum or the minimum of a function by taking its derivative is based on the extreme value theorem. This means if a function $f(x)$ is continuous on a closed interval [a,b], then $f(x)$ has a maximum and minimum value on the interval [a,b]. For multiple unknown parameters, researchers need to determine simultaneous solution set for $n$ equations, where $n$ is the number of unknown parameters. Particularly, for the negative log likelihood function *neg log* $\mathcal{L}$ and $n = 2$, the system is shown in equations (4) and (5):

$$\frac{\partial neg\ \log \mathcal{L}(\beta,\gamma)}{\partial \beta} = 0, \quad (4)$$

$$\frac{\partial neg\ \log \mathcal{L}(\beta,\gamma)}{\partial \gamma} = 0. \quad (5)$$

We solve equations (4) and (5) by applying *mle2* built-in function in *bbmle R* package. This package provides the routine for maximum likelihood estimation. The package is an optimiser from the *stats* package that is based on *Nelder-Mead* algorithm. The *Nelder-Mead* algorithm is the default optimiser in the function *optim* in *R* Packages and can approximate covariance matrix for the parameters by inverting the Hessian matrix at the optimum, which can be later used to derive confidence intervals. The *mle2* function returns the estimated parameter results as shown in Figure 3.

```
Call:
mle2(minuslogl = mLL, start = start, method = "Nelder-Mead",
    control = list(trace = 1))

Coefficients:
      beta      gamma      sigma
 1.2654298  0.8443535 110.4219993

Log-likelihood: -184.16
```

Figure 3: Return parameter results from the *mle2* function

We plot the predicted values versus actual data points to visualize the effectiveness of applying the Maximum Likelihood method by using the algorithm as shown in Figure 4 of which $x$-axis is the time unit and $y$-axis represents the number of cases. Here, the red dots shows the actual cases while the red line indicates the predicted cases. The plot of predicted values is based on the estimated parameters: $\beta = 1.2654, \gamma = 0.8443$ and $\delta = 110.4$. If there is any outlier or not resemble curve, researcher may need to pre-process the data set by sampling or trimming the data size into smaller

size and quantify the parameters again.

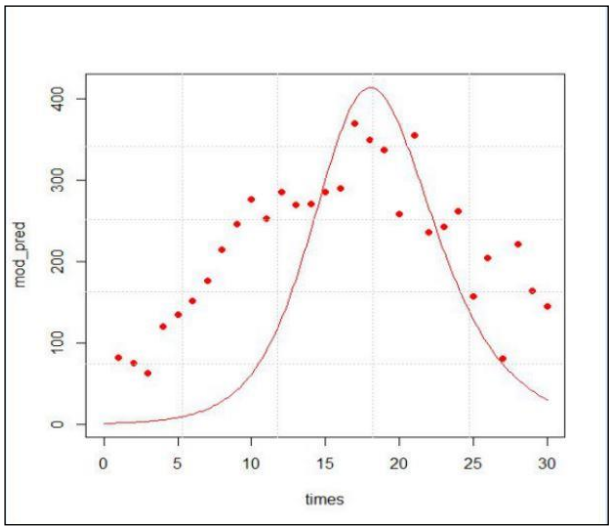


Figure 4: Predicted cases versus actual cases

Quantifying the parameters of a disease model can be defined as finding the parameters that make a disease model fit the actual data best or as close as possible. Researchers' goodness-of-fit metrics is based on the likelihood which the probability or chance of seeing the collected data given a particular model. To validate the parameters that are quantified through the Maximum Likelihood method, we compute the confidence interval and superimpose the plot of the predicted cases versus actual cases as shown in Figure 5. The standard is 95% confidence level, the blue dotted line is the confidence interval for this case, from the plot, researcher can observe whether the actual data points fall between the confidence interval or not. The results validated the MLE algorithms as majority of the actual case (green dots) fall into the intervals. If the actual data points fall in confidence interval, the estimated parameters can be considered to be effective and are of good quality. If they do not fall in the confidence interval, researchers may choose to re-run the test again.

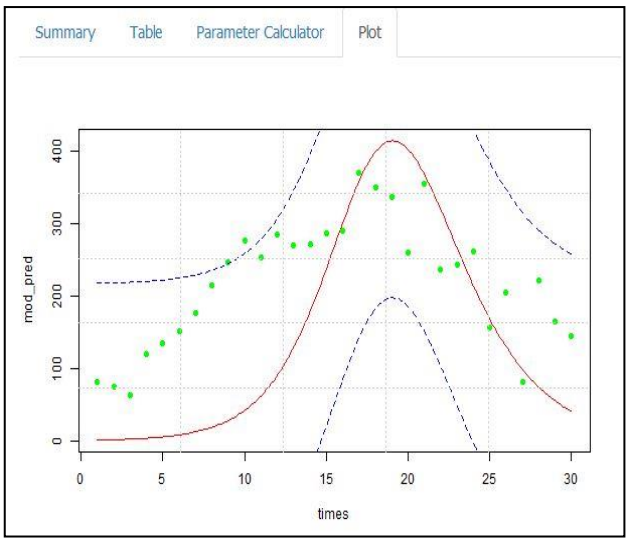


Figure 5: Predicted cases versus actual cases with 95% confidence interval

We also computed the likelihood profile for the fitted model. By constructing the likelihood profile, researchers can plot and look for confidence interval at several different α values, so it is more efficient. By default, the plot method for likelihood profiles displays the square root of the deviance difference (twice the difference in negative log-likelihood from the best fit), so it will be a V-shaped for cases where the approximation works well. The likelihood profile of the transmission coefficient in year 2012 for $\beta = 1.2654$ is shown in Figure 6. Figure 6 shows that the $\beta$ value falls into 95% confidence interval which is between 1.15-1.43 and the $\beta$ value with 1.2654 has the lowest z-score. Z-score shows how many standard deviations in an element is from the mean. A z-score equal to 0 represents the element equal to the mean.

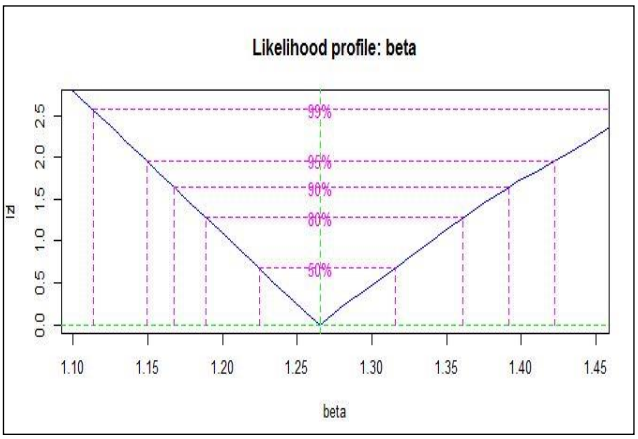


Figure 6: Likelihood profile for transmission coefficient

## III. Results and Discussion

To make the quantification process easier for the researchers who have no programming knowledge, a web-application named Disease Modeling Parameter Calculator was developed to assist in quantifying the disease model parameter(s) for instance the transmission coefficient. This calculator was developed using *R* and Shiny package (a web application framework available in *R* that can turn the statistical analysis into interactive web applications). This tool is integrated in one disease monitoring system named Online Communicable Disease Monitoring System (OCDMS) [23]. This OCDMS aimed to provide occurrences prediction to inform the public health authorities about the seriousness of infected disease if no control measures are taken. This OCDMS driven by the SIR model and the calculator provide the critical parameters to this mathematical model in the OCDMS.

This tool provides a simple graphical user interface (GUI) that can estimate the parameter values with no additional programming skill. The GUI provides wizards which guide end users through the process as portrayed in Figure 7. Users can select a suitable model either an SIR model or a Susceptible-Exposed-Infected-Recovered (SEIR) model. Next, users can import the epidemiological dataset from Comma-Separated Values (CSV) files or text (txt) files by one click. Then, users need to input an initial value for the parameter for instance the transmission coefficient and finally users need to specify the time sequence. After finishing all parameterization in the first tab of the calculator, the calculator will automatically prompt the boxplot about the dataset for users to observe whether there is any outlier. Besides that, the transmission coefficient value will be

prompted in third tab and finally the final tab displays the visualization of the actual data and the estimated data according to the estimated transmission coefficient value from the calculator. The users can navigate and use the transmission coefficient value in their disease model for further analysis.

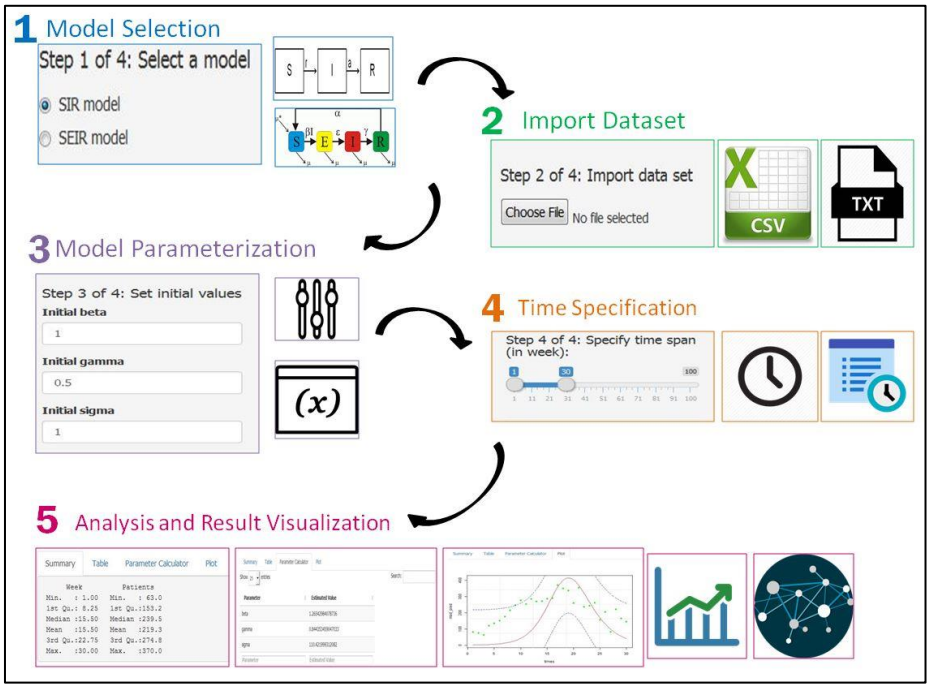


Figure 7: Graphical Interface of the Disease Modeling Parameter Calculator

As mentioned earlier, we would like to investigate the transmission coefficient of HFMD cases in Sarawak from year 2010 until 2014. The epidemiological data from year 2010 until 2014 are imported into the calculator and the results of transmission coefficient are tabulated in Table 3.

Table 3
Estimated Parameter Results

| Year | Number of Cases | Transmission Coefficient Values |
|---|---|---|
| 2010 | 3904 | 37.438 |
| 2011 | 979 | 38.592 |
| 2012 | 10077 | 1.26 |
| 2013 | 5877 | 36.178 |
| 2014 | 6580 | 29.941 |

Infectious disease is transmitted from some source to the susceptible individual. Transmission coefficient represents the infectiousness of a disease [21]. Transmission coefficient of infectious disease varies in time. For instance, the influenza varies seasonally due to the change of seasonal humidity [24]. A disease transmission from infected individuals to susceptible individuals defines the dynamic of an infectious disease. Transmission coefficient can be defined as the product of the total contact rate and transmission probability [5, 25]. Effective contact can be defined as any kind of contact between two individuals when one is infectious and another is susceptible. Effective contact rate is effective contacts per unit time while this can be expressed as total contact rate, which is the total number of contacts either effective or not, per unit time. The transmission probability, on the other hand, is the risk of infection given the contact between an infectious and a susceptible individual.

The result shows the relationship where transmission rate in Table III is inversely proportional to number of cases. As the transmission coefficient increases, the numbers of HFMD cases will decreases. This happens because the number of population affect. We observed this relationship with the concept of parameter units in Table 4. The transmission coefficient is inversely proportional to the people or patients or can be considered as cases in Table 4.

Table 4
SIR Model Parameter Units [24]

| Parameter | Description | Unit |
|---|---|---|
| $\beta$ | Transmission coefficient | $\frac{1}{people\ x\ day}$ |
| $\gamma$ | Recovery coefficient | $\frac{1}{day}$ |
| $S(t,x)$ | Number of susceptible people at time t and space x | $people$ |
| $I(t,x)$ | Number of infected people at time t and space x | $people$ |
| $R(t,x)$ | Number of recovered at time t and space x | $people$ |

## IV. CONCLUSION

It can be concluded that we implemented the algorithm to quantify disease parameter effectively. By using the statistical modelling approach coupled with the Maximum Likelihood Estimation method, the parameter quantification process can be done in lesser time. Nevertheless, the transmission coefficient has been quantified with accuracy of 95% confidence interval. Furthermore, an automated prototype, Disease Modeling Parameter Calculator has been developed to assist end-user to estimate the parameter in shorter time and less hassle. This automated tool is used to quantify parameter for Susceptible-Infected-Recovered (SIR) routine, the tool is ready to include different model routines

for example Susceptible-Exposed-Infected-Recovered (SEIR) model to quantify other parameters.

## ACKNOWLEDGEMENT

Thanks to the Faculty of Computer Science and Information Technology and Universiti Malaysia Sarawak for providing the opportunity and facilities to carry out the research. We would like to deliver our greatest gratitude to UNIMAS Zamalah Vice Chancellor Research Award. Special thanks to the UNIMAS lecturers and staff members for their kind assistance who made this project possible. Our heartfelt thanks go to the Sarawak State Health Department for providing the data generously.

## REFERENCES

[1] Van Lerberghe, W. (2008). *The world health report 2008: primary health care: now more than ever*. World Health Organization.

[2] Murray, C. J., & Lopez, A. D. (1997). Mortality by cause for eight regions of the world: Global Burden of Disease Study. *The lancet*, *349*(9061), 1269-1276.

[3] Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., & Murray, C. J. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *The Lancet*, *367*(9524), 1747-1757.

[4] Reich, N. G., Lauer, S. A., Sakrejda, K., Iamsirithaworn, S., Hinjoy, S., Suangtho, P., ... & Lessler, J. (2016). Challenges in Real-Time Prediction of Infectious Disease: A Case Study of Dengue in Thailand. *PLoS Negl Trop Dis*, *10*(6), e0004761.

[5] Keeling, M. J., & Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton University Press.

[6] Dorjee, S., Poljak, Z., Revie, C. W., Bridgland, J., McNab, B., Leger, E., & Sanchez, J. (2013). A review of simulation modelling approaches used for the spread of zoonotic influenza viruses in animal and human populations. *Zoonoses and public health*, *60*(6), 383-411.

[7] Frances, L. C. (2012). *Comparison of Maximum Likelihood, Bayesian, Partial Least Squares, and Generalized Structured Component Analysis Methods for Estimation of Structural Equation Models with Small Samples: An Exploratory Study*. Master's thesis. University of Nebraska, Lincoln, United States. Retrieved from University of Nebraska Digital Commons.

[8] Jae, M. I. (2002). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90-100.

[9] Latha, S. and Lilly, F. (2012). A Comparison of Parameter Best Estimation Method for Software Reliability Models. *International Journal of Software Engineering and Applications*, *3*(5).

[10] Joseph, S. (2007). *Statistical Estimation in HLM Models.* Retrieved fromhttp://pages.uoregon.edu/stevensj/HLM/data/Estimation_HLM_models.pdf.

[11] Mahesh, S. K. (2008). *Parameter Estimation of Copula Using Maximum Likelihood Estimation Method*. ProQuest.

[12] John, A. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912 – 1922. *Statistical Science*, *12*(2), 162-176.

[13] Dong, L., Wu, K., & Tang, G. (2016). A Data-Centric Approach to Quality Estimation of Role Mining Results. *IEEE Transactions on Information Forensics and Security*, *11*(12), 2678-2692.

[14] Lechner, J., & Günthner, W. A. (2016, September). A tool for a fast evaluation of UHF RFID installations. In *RFID Technology and Applications (RFID-TA), 2016 IEEE International Conference on* (pp. 117-122). IEEE.

[15] Kaveh, A., & Farhoudi, N. (2011). A unified approach to parameter selection in meta-heuristic algorithms for layout optimization. *Journal of Constructional Steel Research*, *67*(10), 1453-1462.

[16] Ooi, M. H., Wong, S. C., Podin, Y., Akin, W., Del Sel, S., Mohan, A., ... Solomon, T. (2007). Human enterovirus 71 disease in Sarawak, Malaysia: a prospective clinical, virological, and molecular epidemiological study. *Clinical Infectious Diseases*, *44*(5), 646-656.

[17] Chuo, F., Tiing, S., & Labadin, J. (2008). A simple deterministic model for the spread of hand, foot and mouth disease (HFMD) in Sarawak. In Proceedings of the *Second Asia International Conference on Modeling & Simulation, 2008. AIMS 08.* (pp. 947-952). IEEE.

[18] Barnes, B., & Fulford, G. R. (2011). *Mathematical modelling with case studies: a differential equations approach using Maple and MATLAB*. CRC Press.

[19] Bolker, B. M. (2008). *Ecological Models and Data in R.* New Jersey, US*:* Princeton University Press.

[20] Gustafsson, L., & Sternad, M. (2007). Bringing consistency to simulation of population models–Poisson Simulation as a bridge between micro and macro simulation. *Mathematical biosciences*, *209*(2), 361-385.

[21] Draganescu, A., Knottenbelt, W., & Heinis, T. (2015). Uncertainty Quantification of Epidemic Phenomena and the Parallel Simulator Tool.

[22] Gholampour, A. A., Ghassemieh, M., & Razavi, H. (2011). A time stepping method in analysis of nonlinear structural dynamics. *Applied and Computational Mechanics*, *5*(2).

[23] Kok, W. C., Labadin, J., Mohammad, A., Wong, K. S., & Chang, Y. L. (2016, May). Android-based Disease Monitoring. In *Information and Communication Technology (ICICTM), International Conference on* (pp. 97-103). IEEE.

[24] Lofgren, E., Fefferman, N. H., Naumov, Y. N., Gorski, J., & Naumova, E. N. (2007). Influenza seasonality: underlying causes and modeling theories. *Journal of virology*, *81*(11), 5429-5436.

[25] Alaa, E. (2013). *Transmission Rate in Partial Differential Equation in Epidemic Models.* Degree Thesis. Marshall University. Retrieved from Marshall University Digital Theses.