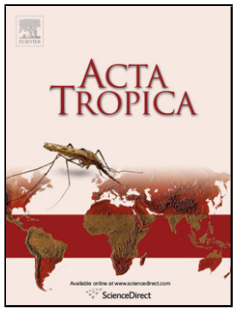


Title: Open Data Mining for Taiwan's Dengue Epidemic

Authors: ChienHsing Wu, Shu-Chen Kao, Chia-Hung Shih, Meng-Hsuan Kan

PII: S0001-706X(17)31294-9
DOI: https://doi.org/10.1016/j.actatropica.2018.03.017
Reference: ACTROP 4618

To appear in: *Acta Tropica*

Received date: 29-10-2017
Revised date: 19-2-2018
Accepted date: 10-3-2018

Please cite this article as: Wu, ChienHsing, Kao, Shu-Chen, Shih, Chia-Hung, Kan, Meng-Hsuan, Open Data Mining for Taiwan's Dengue Epidemic.Acta Tropica https://doi.org/10.1016/j.actatropica.2018.03.017

# Open Data Mining for Taiwan's Dengue Epidemic

ChienHsing Wu, **Correspondence**
Department of Information Management, National University of Kaohsiung,
700, Kaohsiung University Rd., Nanzih District, Kaohsiung 81148, Taiwan, R.O.C.
chwu@nuk.edu.tw

Shu-Chen Kao
Department of Information Management, Kun Shan University,
195, Kunda Rd., YongKang Dist., Tainan, Taiwan, R.O.C.
kaosc@mail.ksu.edu.tw

Chia-Hung Shih
Bachelor Program of Computer and Intelligent Robot, National Pingtung University
51, Minsheng E. Rd., Pingtung City, Pingtung County 90004, Taiwan, R.O.C.
hiroshi.seki@gmail.com

Meng-Hsuan Kan
Department of Information Management, National University of Kaohsiung,
700, Kaohsiung University Rd., Nanzih District, Kaohsiung 81148, Taiwan, R.O.C.
m1053310@mail.nuk.edu.tw

**Abstract**

By using a quantitative approach, this study examines the applicability of data mining technique to discover knowledge from open data related to Taiwan's dengue epidemic. We compare results when Google trend data are included or excluded. Data sources are government open data, climate data, and Google trend data. Research findings from analysis of 70,914 cases are obtained. Location and time (month) in open data show the highest classification power followed by climate variables

(temperature and humidity), whereas gender and age show the lowest values. Both prediction accuracy and simplicity decrease when Google trends are considered (respectively 0.94 and 0.37, compared to 0.96 and 0.46). The article demonstrates the value of open data mining in the context of public health care.

**Keywords:** Open data, data mining, dengue epidemic, Google trend, simplicity

## 1. Background

Dengue fever remains one of the most serious infectious diseases in Taiwan. The Taiwan's Center for Disease Control (TCDC) website (https://nidss.cdc.gov.tw/en/) reported 43,698 confirmed cases of dengue in 2015. Although the number decreased to only 743 in 2016, the Taiwanese government is still cautiously monitoring the epidemic , and implementing innovative strategies such as open data applications, which involve data (or data-driven) science, model development, and domain knowledge (Dhar, 2013; Zuiderwijk & Janssen, 2014; Zeleti et al., 2016; Hsu et al., 2017).

Open data applications to predict dengue epidemic cover a variety of models such as linear and multi regression, moving average, weighted moving average, intra- and inter-seasonal autoregression, time series, networks, correlation, analytical hierarchy process, and their combinations. Models and variables have been tested in various countries including Thailand (Wongkoon et al., 2012), Malaysia (Dom et al., 2013; Dom et al., 2016), Taiwan (Chien and Yu, 2014), Bengal (Banu et al., 2014), Vietnam (Phung et al., 2015), Colombia (Eastin et al., 2014; Delmelle et al., 2016) and Saudi Arabia (Ibrahim Alkhaldy, 2017).

Differently from these studies, we apply data mining to discover prediction themes in dengue open data. Online communities have been increasingly attracting attention to possible insights that could be derived from data mining based on open data sources (Brownstein et al., 2009; Santillana et al., 2014; Yang et al., 2017; Strauss et al., 2017). For example, the use of Google trends to develop data-driven models (classification-oriented models) of dengue epidemic is still under investigation. We expect that the insights derived from this approach to be relevant to the development of government management and control policies related to dengue.

This study has three main purposes. First, we examine the applicability of data mining technique (data-driven classification technique) to predict dengue epidemics.

Second, we investigate the role of Google trend data on prediction accuracy and simplicity of decision trees. Finally, based on the results obtained above we discuss management insights by conducting the mining results.

## 2. Literature Review

### 2.1 Studies of dengue epidemic

Previous studies of dengue epidemics involved various predictors such as population (density and urbanization), environmental attributes (transportation and water sources), temporal attributes (time, season), and climate attributes (temperature, humidity, and precipitations). Few have focused on the living conditions (housing and individual behavior) and geographical attributes (altitude). Recently, web resources such as Google Trends have attracted increasing attention of dengue epidemic researchers. A basic review of literature regarding studies of dengue prediction is presented.

Most studies of dengue epidemic in Taiwan have been based on quantitative approaches. For example, Wu et al. (2009) proposed a regression model with urbanization, population characteristics, temperature, and precipitation as predictors and confirmed that most factors were significant except precipitation. The model predicts that when average temperature reaches 18$^{o}$C, dengue epidemic enters the danger period and infection increases by a factor of 1.95 times for every 1$^{o}$C increase in temperature. Hsueh et al. (2012) focused on the geographical distribution in Kaohsiung, Taiwan, investigating regions of high population density, high transportation places, and plentiful water sources where dengue epidemic likely takes place, but they have not included any climate factors (temperature) in their analysis. Chen and Chang (2013) utilized approximate entropy algorithm and pattern recognition to predict dengue outbreaks and confirmed their prediction power for the southern areas of Taiwan (Kaohsiung and Pingtung areas). Chieh and Yu (2014) proposed a nonlinear model based on environment and climate factors, temperature and precipitations concluded that higher temperature leads to higher infection rates; when precipitations is higher than 50mm, the infection rate can last for as long as 15 weeks. Retrospective analysis also showed that temporal patterns are not a likely predictor of dengue outbreaks. This implies that dengue outbreaks are becoming increasingly unpredictable probably due to the rapid climate change (Sanna and Hsieh,

2017).

In regions other than Taiwan, research findings varied. Phung et al. (2015) proposed a prediction model of dengue epidemic in Vietnam based only on climate factors and results showed that mean temperature and mean relative humidity are significant factors, but total precipitations was not, consistent with Wu et al. (2009). Similarly, precipitation was not significant both in a neural network-based prediction model in Singapore (Aburas et al., 2010) and in a simple empirical and graphic model in Mexico (Machado-Machado, 2012). However, areas with plentiful water sources present stronger association with dengue epidemic (Hsueh et al., 2012). This implies that precipitation may result in high humidity and an environment favoring breeding of mosquitoes that transmit dengue. This is supported by Dom et al. (2013) who showed that humidity is a significant predictor of dengue while days of rain are not. Dengue epidemic behavior was also studied through global and local networks (Malik et al., 2017), revealing that some localities were repeatedly affected and implying that location is possibly one factor.

A time series model showed that a dengue epidemic in Thailand was associated with time periods as well as environment and climate (Wongkoon et al., 2012), while temperature and humidity were found to be significant factors associated with the dengue fever epidemic both in Bengal (Banu et al., 2014) and Saudi-Arabia (Alkhaldy, 2017). A weighted regression model to predict dengue fever in Saudi-Arabia (Khormi and Kumar, 2011) showed that housing quality, population size, population density were significant factors, but age was not. In Mexico, population density was not a significant factor (Machado-Machado, 2012). By considering the meteorological data, the intra- and inter-seasonal autoregressive models were proposed to predict dengue outbreaks in Colombia (Eastin et al., 2014). Main research finding indicated that from the perspectives of temporally varying count regression the temperature confined between 18°C and 32°C is very much conducive to dengue outbreaks during warm-dry periods and humidity and rainfall showed various effects on dengue epidemic.

The review above shows that most studies focused on climate factors (temperature, humidity, and precipitations) but have not emphasized environment (city configuration and transportation) and population (population size, population density, and age), while few considered temporal factors (e.g., Eastin et al., 2014;

Yang et al., 2017). Only temperature and humidity are consistently found to be significant factors. In general, the degree of dengue epidemic depends on its antecedents, but previous studies of dengue epidemic have given limited attention to classifying factors according to their predictive power. For example, it is unknown whether temperature is a stronger predictor than humidity or vice-versa. Moreover, little emphasis has been placed on the effect of information sources based on online communities such as Google trends and that of combination with meteorological data on dengue epidemic (Brownstein, et al., 2009; Strauss et al., 2017; Yang et al., 2017).

### 2.2 Trends in information demands by online communities

Over the past decade, thriving online communities not only made information easily sharable but also gave rise to information trends. More and more health-related information on the Internet are available, such as Google trends, which allows Internet users using search engines to find important insights about current disease trends, such as dengue symptoms, treatments, and epidemic (Brownstein, et al., 2009). This signifies that when a dengue fever epidemic is taking place, online community users often search for information and updates, and the level of information demand may correlate with the current level of the epidemic.

To deepen the use of online information demand, Santillana et al. (2014) employed root mean squared errors and Pearson correlations to develop prediction models based on two data sources: Google information trends and cases of H1N1 influenza-like illness from American disease centers, demonstrating that Google information demands were a better predictor. This finding is supported by Strauss et al. (2017) who showed based on Pearson correlations that Google dengue trends were associated with the number of reported cases. Using the multivariate linear regression modeling framework (ARGO, AutoRegressive model with GOogle search queries as exogenous variables), a similar finding was presented that dengue-related Google search trends is associated with dengue activity (Yang et al., 2017). This implies that Google dengue trends have been extensively and successfully used in the predication and surveillance of dengue epidemic. In summary, trends in Google information demands correspond to significant amount of data on a given term such as dengue, enterovirus, or influenza in the context of disease. This study explores the use of Google trends as a predictor of dengue epidemics in Taiwan.

### 2.4 Knowledge discovery

Knowledge discovery (KD) using data mining (DM) has many advantages (Ristoski ety al., 2015; Chemchem and Drisas, 2015; Lausch et al., 2015) such as being entirely a data-driven approach, as well as its learnability, high classification accuracy, and reliance on multi-context datasets. Although not a new KD mechanism, DM is mainly based on the five modes of association, clustering, regression, summarization, and classification (Chen et al., 1996). Particularly, when data have the characteristics of dimensionality, multi-collinearity, and non-homogeneity, traditional statistical approaches based on causal effects (regression models) or difference analysis may be not suitable, while data-driven and top-down classification techniques (ID3 and its extensions) are appropriate (Quinlan, 1986; Ture et al., 2009; Ramezankhani et al., 2014).

ID3 was introduced by Quinlan (1986) and has been extensively used to derive information entropy for data sets of under multiple classes (Wu et al., 2013; Pombo et al., 2014; Kargarfard et al., 2015). Based on information theory principles, ID3 adopts a top-down induction method to return the degree of separation between variables. The higher entropy of a variable, the higher its power of classification. The outcomes of the top-down induction mechanism are the classification of variables according to their power and a decision tree that contains a set of decision rules. Using decision trees, Ramezankhani et al. (2014) have proposed decision rules and a prediction model for diabetes with 90.5% accuracy. However, tree size (simplicity), quality of the decision rule (accuracy and support) and feedback from domain practitioners and specialists remain important issues that KD applications need to address in the context of dengue epidemic management.

The data sources included in our study are dengue open data (personal and location data), climate factors, and Google trends. These data exhibit the features of dimensionality, multi-collinearity, and non-homogeneity (location, age, number of cases, temperature, and number of searches by online communities). After converting continuous data into discrete variables, we applied the ID3 mining algorithm to obtain decision trees and decision rules, as it is an unsupervised method with comparatively higher efficiency.

## 3. Method

A three-phase method (preparation, implementation, and validation) was used to derive research results. Preparation refers to data preprocessing, implementation applies the mining mechanism to data, and validation criteria are based on accuracy (correctness of decision rules) and simplicity (precision of the decision tree). The design feature is presented in Table 1.

**< Table 1 >**

### 3.1 Data sources

We relied on three data sources. The dengue open dataset (ODD) having 70914 tuples was collected from Taiwan's Centers for Disease Control (http://data.gov.tw/node/9912). The climate open dataset having 4149 tuples (ODC) was obtained from Taiwan's Center Weather Bureau (http://www.cwb.gov.tw/V7/climate/monthlyData/mD.htm). Google trends open data was extracted from local Taiwan's Google online community (ODG) (https://trends.google.com/trends/explore?date=all&geo=TW&q=dengue) by using the key term 'Dengue'. The original ODD was a weekly-based dataset with six attributes (year, county or city, gender, age, week, residence) and one classification attribute (confirmed case). The ODC is a monthly-based dataset with four attributes (year, month, county or city, mean temperature, total precipitations, and mean humidity). The ODG is a monthly-based dataset with two attributes (month and number of inquiry).

**< Table 2 >**

### 3.2 Data preprocessing

Data preprocessing involves two tasks: to convert weekly-based dengue dataset (ODD) into monthly-based data, and to granulate continuous data into discretized ones for the mining mechanism to perform the classification task. The technique used to conduct data discretization is unsupervised equal width interval (Wu et al., 2013). The number of discrete levels is determined by the minimum k for $2^k \geq n$, where n is sample size (Wu, 2002) and k is the number of granule (or levels). The conversion

function (CF) is presented in equation (1) as below.

$$CF_{UEWI}(x_i) = \begin{cases} Attr_j, & if \ \ x_{min} + (j\text{-}1)d \le x_i < x_{min} + jd \\ Attr_m, & if \ \ x_i = x_{max} \\ null & otherwise \end{cases} \quad (1)$$

Where, $x_i$ : value of $i^{th}$ data, i=1, 2, …, n, n is the size of data set
$Attr_m$ ：$x_i$ is converted into the $j^{th}$ level, $j$= 1, …k, k is the number of levels
$x_{min}$：the minimum value
$x_{max}$：the maximum value
$d$ ：interval

There were seven tasks in data discretization and data combination. They are: (1) discretization of age, (2) conversion of weekly-based to monthly-based, (3) discretization of confirmed cases (classification attribute), (4) combination of ODD and ODC into ODDC (open dataset for dengue and climate), (5) discretization of mean temperature, total precipitation, and mean humidity in ODDC, (6) discretization of Google trend data, and (7) combination of ODDC and ODG into ODDCG (open dataset for dengue, climate, and Google trends).

Particularly, to transform the ODD dataset to be consistent with other data sources that have different scales, we performed three data pre-processing operations. The first operation is to convert weekly-based data tuples into monthly-based data tuples. The converted dataset therefore has 16458 tuples. The second is to granulate data tuples of the attribute Age. The granulated dataset has 16458 data tuples. The final is to combine data tuples that have the same values of Year, Month, County, Gender, and level of Age. It is noticed that doing so is to have the dataset applicable for data mining mechanism to perform classification. The total confirmed cases remain the same and the size of the combined dataset is therefore reduced to be 4149 tuples. Details of dataset are presented in Table 2.

Although it is necessary to discretize continuous data prior to application of the mining algorithm, discretized data may contain inconsistent data leading to different conclusions (Wu et al., 2013). Theoretically, same conditions in a dataset should result in the same conclusion (or decision rules). However, deleting all inconsistent data may lead to loss of important information. We therefore adopted a rule. If the decision rule derived from a dataset with a number of cases more than or equal to half the size

of an inconsistent sub-dataset, keep all cases of the rule and eliminate the remaining inconsistent cases. Otherwise, eliminate the whole inconsistent sub-dataset.

### 3.3 Data mining and validation

The final discretized dataset (ODDCG) contained a sample of 4149 individuals. It was then randomly divided into a subset with two-thirds of cases (2766) used for training the mining algorithm and a subset with the remaining one-third (1383) cases used for validation. The ID3 was used to mine the larger subset. Accuracy estimation is based on application of the rules mined from the larger subset to the smaller subset. Simplicity is based on the consideration that a rule supported by a larger number of cases is likely to be stronger, while a rule based on fewer conditions is likely to be simpler (Wu, 2003). In other words, the higher the simplicity, the better the mined decision tree. Information entropy, expected information of an attribute (denoted by Attr), and final gained information of an attribute (denoted by Gain (Attr)) are determined by formulae (1), (2), and (3). Simplicity of a decision tree is determined by the equation (4).

$$I(n_{c_1}, n_{c_2}, .....n_{c_n}) = (-\frac{n_{c_1}}{T}\log_2\frac{n_{c_1}}{T}) + .... + (-\frac{n_{c_n}}{T}\log_2\frac{n_{c_n}}{T}) \quad (1)$$

$n_{C_i}$ : The number of records that return to class $C_i$, i=1,2,....,n.
$T$ : The total number of records.

$$Exp(Attr) = \sum_{i=1}^{t}[(\frac{n_{vi}}{T})I(a_{viC_1}, a_{viC_2}, ...., a_{viC_m})] \quad (2)$$

t：The number of different values that attribute Attr can take on.
$n_{vi}$：The total number of records that attribute Attr takes on value $V_i$ , i=1,2,...,t.
$a_{viCj}$ ：the total number of records that attribute Attr takes on value $V_i$ and returns to class $C_j$, i=1,2,...,t, j=1,2,...,m.
$T$：The total number of records.

$$\text{Gain}(\text{Attr}) = \text{I}(n_{c_1}, n_{c_2}, .....n_{c_n}) - \text{E}(\text{Attr}) \quad (3)$$

$$E(S) = \sum_{i=1}^{r}\frac{1}{C_i} \times P_{Ri} \quad (4)$$

$C_i$ : the number of conditions of $i^{th}$ decision rule,
i=1, 2, 3,…, r, r is the total number of rules
$P_{Ri}$ : the support of the $i^{th}$ rule

## 4. Results and Discussions

Data reveal a positive association between Google trends and confirmed cases (Figure 1), suggesting that Google trends are likely to be a predictor of dengue epidemic. Figure 2 shows that location (County) plays a role in the number of confirmed cases, with most occurrences of dengue epidemic centered in three areas (Kaohsiung city, Tainan city, and Pintung county) accounting for more than 96% confirmed cases (68196 cases). Figure 3 depicts data of Month against confirmed cases, and reveals that the number of confirmed cases is higher between August and December in almost every sampled year.

Furthermore, entropy computed for attributes of the discretized dataset is presented in Table 3. When Google trends are excluded, County holds the highest expected entropy followed by the Month. Therefore, location and time are the main predictors of dengue epidemic in Taiwan. As for climate, Mean Temperature has the highest entropy, followed by Total Precipitation and Mean Humidity. Individual characteristics (Age and Gender) exhibit less predictive power. When Google trends are included in the analysis, they exhibit higher expected entropy than any other variable (County, Month, Temperature, Precipitation, Humidity, Age, and Gender).

**< Table 3 > < Figure 1, 2, 3 >**

Table 4 presents data on prediction accuracy and simplicity of the returned decision rules. When Google trends are not considered, 104 rules are produced with an accuracy of 0.96 and simplicity of 0.46. In contrast, when Google trends are included 107 rules were mined from the data, with prediction accuracy down to 0.94, and simplicity down to 0.37. Discussions are addressed below.

First, findings indicate that location and time (month) are the main predicators of dengue fever epidemic in Taiwan. This is consistent with the findings by Malik et al. (2017) in Malaysia and those by Hsu et al. (2017), but uncovered by most other reports. This implies that using climate variables alone may be insufficient to predict and control dengue fever epidemics. From north to south, Taiwan is 394 km long, but temperature range in a year varies approximately from10$^{o}$C to 35$^{o}$C. Temperatures are higher in southern Taiwan, which explains the presence of dengue mosquitoes

(*Aedesa Egypti* and *Aedes albopictus*) and the higher incidence of dengue mostly in this region.

Second, as discussed earlier, simplicity represents the width and depth of the mined decision tree and it is a measure of the mining performance of ID3. Results showed that simplicity of the mined outcomes was higher when Google trends were not considered. Google trends presented the highest classification power among variables, but leads to less simplicity. This implies that Google trends results in a comparatively complex decision tree and less supported decision rules. A similar situation was observed in the case of prediction accuracy. From a practical perspective, this suggests that Google trends may be useful to reveal the situations of dengue epidemic. From a theoretical perspective, Google trends at this particular point of time were not demonstrated to be a better decision generator in the context of dengue fever epidemic. Further studies are required to explore the issue in more detail.

Finally, unlike most previous reports, our study adopts multiple data resources (open data, climate data, and online information demand) as predictors of dengue fever epidemic. The predication accuracy of ID3 was acceptable, suggesting that this data mining technique has enough classification power to predict dengue epidemic. To deepen the knowledge and insights, the qualitative study using empirical survey on dengue practitioners was carried out.

Although previous studies indicated that man-made factors such as transportation conditions, population density, and living environment have relevant effects on dengue epidemic(Hsueh et al., 2012; Khormi and Kumar, 2011; Dom et al., 2016; Delmelle et al., 2016), location has natural characteristics affecting dengue epidemics. We suggested that governments or agencies should consider location as a priority variable, and develop location-based management policies, regulations, and strategies to cope with dengue. A possible solution would be a location-based awareness system in highly epidemic areas such as Kaohsiung, Pingtung and Tainan.

In the case of the Season variable, it is believed that the peak epidemic period lasts from the middle of summer (May) until the November (see Figure 3). To reduce epidemic breakouts, the preventive action should therefore occur from May or June, especially after the confirmation of the first case in the main counties and cities. The traditional Season variable no longer obviously reflects dengue epidemic because of increasing climate uncertainty, implying that Season may be gradually losing its

classification power, and as long as the temperature is appropriate (25-30$^{o}$C), mosquitoes are likely to be active.

In relation to climate, temperature has been seen as the main reason determining activity of disease mosquitoes. Our research finding is consistent with most previous literature (Wu et al., 2009; Dom et al., 2013; Phung et al., 2015; Ibrahim Alkhaldy, 2017). Although supporting this view, removing the breeding sources (water) of disease mosquito would be an effective way to reduce the epidemic in any season, not only in summer. Whenever possible, efforts should therefore be made to eliminate or mitigate any natural conditions favoring the reproduction of disease mosquitoes.

As for the variable Precipitation, it is not very much emphasized in management and control of dengue epidemic in Taiwan. Regarding possible reasons, however, although rainfall is not directly related to dengue, it contributes to the formation of small water ponds that favor breeding of mosquitoes (Chien and Yu, 2014). This factor is especially relevant in the case of rapid and strong rains whose frequency seem to be increasing as a result of extreme climate.

As for the variable Google trends, the information source has not yet been incorporated into management practices of dengue epidemic in Taiwan. Google trends are useful according to Santillana et al. (2014) and Strauss et al. (2017). However, trends have increased neither prediction accuracy nor simplicity of decision trees in comparison to the variables County and Season. The application of Google trends Taiwan's dengue epidemic needs further investigation. For example, a bi-direct causal model testing whether Google trends predict dengue epidemic or vice-versa has not yet been proposed. In general, the possible role of online social network in management and control of dengue epidemic should be addressed by future studies.

## 5. Concluding Remarks

The article highlights the value of open data mining in the context of public health care. It has three main purposes: the utilization of data mining technique (ID3) to predict dengue epidemic, the integration of multiple open data sources, and the derivation of practical insights from an empirical qualitative survey of domain specialists and experts. Results show that the classification-oriented data mining technique can be successfully applied to Taiwan's dengue open data. From the viewpoint of health care management, variables used to describe Taiwan's dengue

epidemic are either congenital (location, season, climate, and individual) and hence mostly uncontrollable, or acquired (Google trends) and created through relationships. We reported and ordered the predicted power of congenital variables through quantitative analyses. Moreover, there has been a gap between academic and practical perceptions. To reduce the gap, the use of discovered knowledge via an empirical survey may be advantageous to the development of dengue control strategy and policy. This requires the development of an in-depth understanding of dengue epidemic in particular and of the data modeling in general. Furthermore, climate is changing rapidly, which may greatly influence living environments. This means that the effects of location, time, and climate variables on dengue epidemics are very likely to change. Future studies should therefore keep tracking the possible impact of those on the dengue fever epidemic.

## References

Aburas, H.M., Cetiner, B.G., and Sari, M. 2010, Dengue confirmed-cases prediction: A neural network model, Expert Systems with Applications, 37(6), 4256-4260.

Alkhaldy, I. 2017, Modelling the association of dengue fever cases with temperature and relative humidity in Jeddah, Saudi Arabia - A generalized linear model with break-point analysis, Acta Tropica, 168, 9-15.

Brownstein, J.S., Freifeld, C.C. and Madoff, L.C. 2009, Digital disease detection - harnessing the web for public health surveillance, New England Journal of Medicine, 2009. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2917042/

Banu, S., Hu, W., Guo, Y., Hurst, C. and Tong, S. 2014, Projecting the impact of climate change on dengue transmission in Dhaka, Bangladesh, Environment International, 63, 137-142.

Chemchem, A. and Drisas, H. 2015, From data mining to knowledge mining: Application to intelligent agents, Expert Systems with Applications 42, 1436–1445.

Chen, C.C. and Chang, H.C. 2013, Predicting dengue outbreaks using approximate entropy algorithm and pattern recognition, Journal of Infection, 67(1), 65-71.

Chen, M.S., Han, J. and Yu, P.S. 1996, Data mining: an overview from a database perspective, IEEE Transaction on Knowledge and Data Engineering, 8, 866–883.

Chien, L.C. and Yu, H.L. 2014, Impact of meteorological factors on the spatiotemporal patterns of dengue fever incidence, Environment International, 73, 46-56.

Delmelle, E., Hagenlocher, M., Kienberger, S. and Casas, I. 2016, A spatial model of socioeconomic and environmental determinants of dengue fever in Cali, Colombia, Acta Tropica, 164, 169-176.

Dhar, V. 2013, Data science and prediction, Communications of the ACM, 56(12),

64-73

Dom, N.C., Hassan, A.A., Latif, Z.A. and Ismail, R. 2013, Generating temporal model using climate variables for the prediction of dengue cases in Subang Jaya, Malaysia, Asian Pacific Journal of Tropical Disease, 3(5), 352-361.

Dom, N.C., Hassan Ahmad, A., Abd Latif, Z. and Ismail, R. 2016, Application of geographical information system-based analytical hierarchy process as a tool for dengue risk assessment, Asian Pacific Journal of Tropical Disease, 6(12), 928-935.

Eastin, M.D., Delmelle, E., Casas, I., Wexler, J. and Self, C. 2014, Intra-and interseasonal autoregressive prediction of dengue outbreaks using local weather and regional climate for a tropical environment in Colombia, The American journal of tropical medicine and hygiene, 91(3), 598-610.

Hsu, J.C., Hsieh, C.L. and Lu, C.Y. 2017, Trend and geographic analysis of the prevalence of dengue in Taiwan, 2010–2015, International Journal of Infectious Diseases, 54, 43-49.

Hsueh, Y.H., Lee, J. and Beltz, L. 2012, Spatio-temporal patterns of dengue fever cases in Kaoshiung City, Taiwan, 2003–2008, Applied Geography, 34, 587-594.

Ibrahim Alkhaldy, I. 2017, Modelling the association of dengue fever cases with temperature and relative humidity in Jeddah, Saudi Arabia—A generalized linear model with break-point analysis, Acta Tropica, 168, 9-15.

Kargarfard, F., Sami, A. and Ebrahimie, E. 2015, Knowledge discovery and sequence-based prediction of pandemic influenza using an integrated classification and association rule mining (CBA) algorithm, Journal of Biomedical Informatics 57, 181–188.

Khormi, H.M. and Kumar, L. 2011, Modeling dengue fever risk based on socioeconomic parameters, nationality and age groups: GIS and remote sensing based case study, Science of The Total Environment, 409(22), 4713-4719.

Lausch, A., Schmidt, A. and Tischendorf, L. 2015, Data mining and linked open data – New perspectives for data analysis in environmental research, Ecological Modelling 295, 5–17.

Machado-Machado, E.A., 2012, Empirical mapping of suitability to dengue fever in Mexico using species distribution modeling, Applied Geography, 33, 82-93.

Malik, H.A.M., Waheed Mahesar, A., Abid, F., Waqas, A. and Ridza Wahiddin, M. 2017, Two-mode network modeling and analysis of dengue epidemic behavior in Gombak, Malaysia, Applied Mathematical Modelling, 43, 207-220.

Phung, D., Huang, C., Rutherford, S., Chu, C., Wang, X., Nguyen, M., Nguyen, N.H. and Manh, C.D., 2015, Identification of the prediction model for dengue incidence in Can Tho city, a Mekong Delta area in Vietnam, Acta Tropica, 141, Part A, 88-96.

Pombo, N., Araújo, P. and Viana, J. 2014, Knowledge discovery in clinical decision support systems for pain management: A systematic review, Artificial Intelligence in Medicine, 60(1), 1-11.

Quinlan, J.R., (1986), Induction of decision tree, Machine Learning, 1, 81-106.

Ramezankhani, A., Pournik, O., Shahrabi, J., Khalili, D., Azizi, F., and Hadaegh, F. 2014, Applying decision tree for identification of a low risk population for type 2

diabetes. Tehran Lipid and Glucose Study, Diabetes Research and Clinical Practice, 105(3), 391-398.

Ristoski, P., Bizer, C. and Paulheim, H. 2015, Mining the web of linked data with RapidMiner, Web Semantics: Science, Services and Agents on the World Wide Web, 35, 142–151.

Sanna, M. and Ying-Hen Hsieh, Y.H. 2017, Temporal patterns of dengue epidemics: The case of recent outbreaks in Kaohsiung, Asian Pacific Journal of Tropical Medicine, 10(3), 292-298.

Santillana, M., Zhang, D.W., Althouse, B.M. and Ayers, J.W. 2014, What can digital disease detection learn from (an external revision to) Google flu trends? American Journal of Preventive Medicine, 47(3), 341-347.

Strauss, R.A., Castro, J.S, Reintjes, R. and Torres, J.R. 2017, Google dengue trends: An indicator of epidemic behavior. The Venezuelan Case, International Journal of Medical Informatics, 104, 26-30.

TCDC, Taiwan's Center Disease Control 2016, Access date: 2016, January, http://www.cdc.gov.tw/english/index.aspx

Ture, M., Tokatli, F. and Kurt, I. 2009, Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients, Expert Systems with Applications, 36(2), 2017-2026.

Wongkoon, S., Jaroensutasinee, M. and Jaroensutasinee, K. 2012, Development of temporal modeling for prediction of dengue infection in Northeastern Thailand, Asian Pacific Journal of Tropical Medicine, 5(3) 249-252.

Wu, C.H. 2003, On the granulation simplicity for the decision rule discovery in databases: EWI Vs. EFI, International Journal of Science and Technology, 14, 28-36.

Wu, C.H., Kao, S.C. and Okuhara, K. 2013, Examination and comparison of conflicting data in granulated datasets: Equal width interval vs. equal frequency interval, Information Sciences, 239 (1), 154-164.

Wu, P.C., Lay, J.G., Guo, H.R., Lin, C.Y., Lung, S.C. and Su, H.J. 2009, Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical Taiwan, Science of The Total Environment, 407(7), 2224-2233.

Yang, S, Kou, S.C., Lu, F., Brownstein, J.S., Brooke, N., and Santillana, M. 2017, Advances in using Internet searches to track dengue, PLoS Computation Biology, 13(7), e1005607. https://doi.org/10.1371/journal.pcbi.1005607

Zeleti, F.A., Ojo, A., and Curry, E. 2016, Exploring the economic value of open government data, Government Information Quarterly, 33(3), 535-551.

Zuiderwijk, A. & Janssen, M. 2014, Open data policies, their implementation and impact: A framework for comparison, Government Information Quarterly, 31(1), 17-29.

Table 1: Design features

| Features | Description |
|---|---|
| Objectives | (1) Applicability of data mining in open data of dengue epidemic<br>(2) Comparison of Google trend considered with not considered in the mined results<br>(3) Management insights of dengue epidemic |
| Datasets | (1) Dengue open data,<br>(2) Climate,<br>(3) Google trends |
| Data preprocessing | (1) Combination of dengue open data, climate data, and Google trend<br>(2) There are 4 attributes and 1 classification attribute (i.e., confirmed case) in the Dengue open data, 3 attributes in the Climate (mean temperature, precipitations, and mean humidity), and 1 attribute in Google trend<br>(3) Data discretization: age, mean temperature, precipitations, and mean humidity, confirmed cases, and Google trend |
| Discretization | Unsupervised equal width interval (i.e., the smallest k for $2^k \geqq n$, n is the dataset size, k is the number of granule) |
| Mining mechanism | ID3 |
| Mining outputs | (1) The order of classification power of variables<br>(2) Decision rules |
| Validation and comparison | (1) Accuracy (2/3 for training and 1/3 for validation)<br>(2) Simplicity of decision tree generated |

Table 2：Datasets and their attributes

| Dataset and attributes | | Value | Being preprocessed |
|---|---|---|---|
| Dengue | | | |
| | Data period | 2006/January~2016/March | |
| | Size | 70914 | |
| | Week | 1-53 | Weekly-based is converted into monthly-based (12 months) |
| | County | - | 21 counties (21 categories) |
| | Gender | Male/Female | 2 categories |
| | Source | From domestic/oversea | 2 categories |
| | Age | Every 4-year (e.g., 0~4, 5~8) | Converted into 8 levels |
| | Confirmed cases | Weekly-based | Converted into monthly-based (12 categories) |
| Climate | | | |

| | | | |
|---|---|---|---|
| | Data period | 2006/January~2016/March | |
| | Size | 4149 | Fitted to the converted dengue dataset |
| | County | - | 21 counties (21 categories) |
| | Mean temperature | - | Data is discretized (13 levels) |
| | Total precipitations | - | Data is discretized (13 levels) |
| | Mean humidity | - | Data is discretized (13 levels) |
| Google trend | Information inquiry | 2006/January~2016/March | Data is discretized (13 levels) |

Table 3: The gained entropy for attributes

| Categories | Google trend not considered | | Google trend considered | | |
|---|---|---|---|---|---|
| | Attributes | Exp. entropy | Attributes | Exp. entropy | Order |
| Online demand | - | - | Google trend | 0.07175793 | 1 |
| Location | County/City | 0.06028398 | County | 0.06028398 | 2 |
| Time | Month | 0.05253638 | Month | 0.05253638 | 3 |
| Climate | Mean temperature | 0.04748686 | Mean temperature | 0.04748686 | 4 |
| | Total precipitations | 0.03947917 | Total precipitations | 0.03947917 | 5 |
| | Mean humidity | 0.03780933 | Mean humidity | 0.03780933 | 6 |
| Individual | Age | 0.02103388 | Age | 0.02103388 | 7 |
| | Gender | 0.00249399 | Gender | 0.00249399 | 8 |

Table 4: The mined results

| Criteria | Results | |
|---|---|---|
| | Google trend not considered | Google trend considered |
| The number of mined rules | 104 | 107 |
| Prediction accuracy | 0.96 | 0.94 |
| Simplicity | 0.46 | 0.37 |

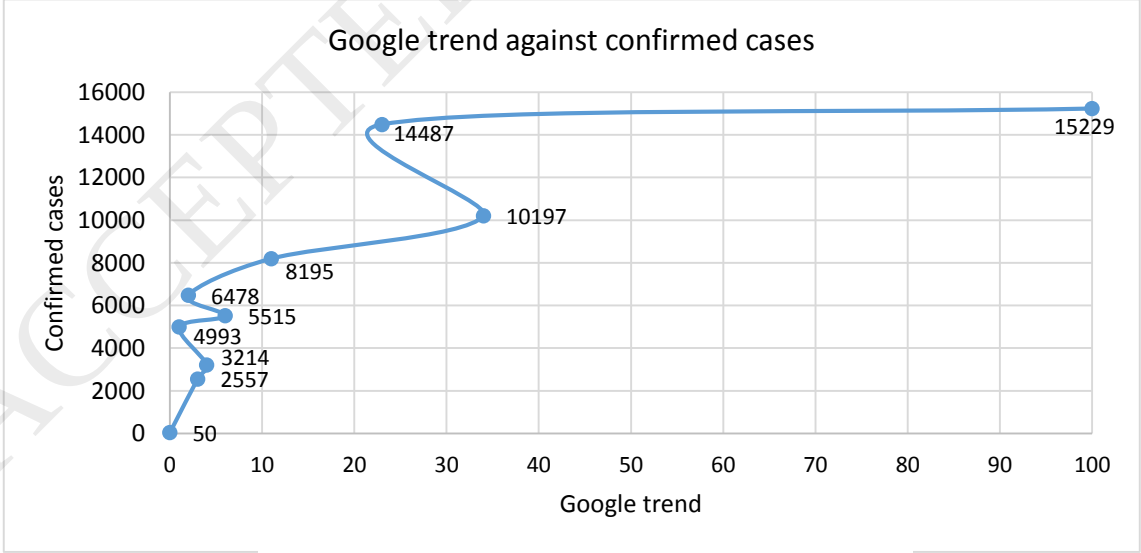


Figure 1: Google trend against confirmed cases

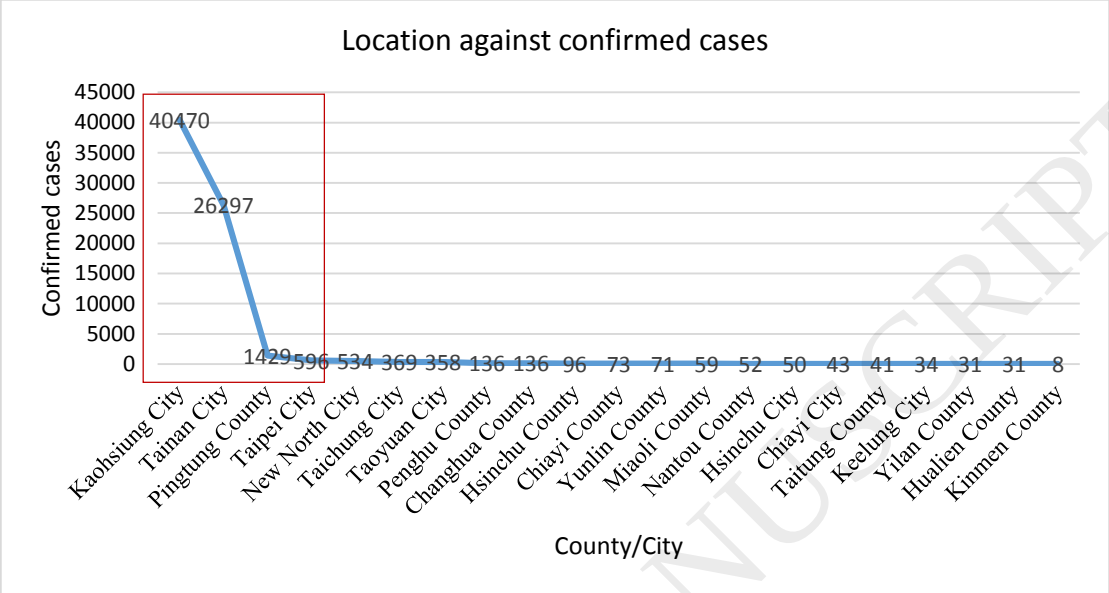


Figure 2: Location against confirmed cases

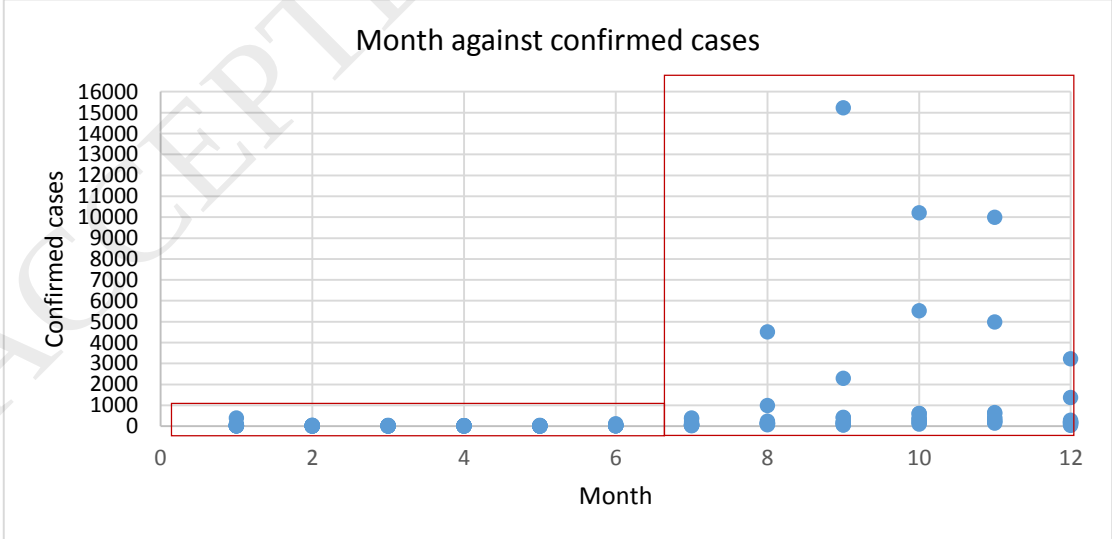


Figure 3: Month against confirmed cases