Temporal Disaggregation of Annual Dengue Cases to Monthly Dengue Cases

By Nicole Wayant

INTRODUCTION: Lack of disease incidence data is a common problem within the field of epidemiology (Beale et al. 2008). In order to study the distribution, frequency, patterns, and predictors of disease, epidemiologists need to have a sufficient record of past disease cases. Unfortunately, for many areas of the world, disease incidence data is extremely lacking. At best, data may be available at country level or on an annual basis. From a temporal point of view, this is unacceptable for identifying periods of extreme disease activity, the seasonal patterns of disease, or prediction. Since the global collection of higher fidelity disease data is unlikely, a way to temporally disaggregate annual disease incident data must be identified.

Temporal disaggregation is the process of taking a low frequency time series, such as annual disease cases, and dividing the time series into a higher frequency time series, such as monthly disease cases. This process of down sampling is typically done by 1) dividing the lower frequency time series into equal portions of the higher frequency time series or 2) using one or more related higher frequency time series to model the desired signal (Chamberlin 2010). For example, if disease transmission is dependent on the presence of mosquitoes, then infections are expected to occur more frequently when conditions are good for mosquito survival and reproduction. High frequency indicator time series of mosquito presence, such as temperature and precipitation, are often recorded and readily available.

Within epidemiological literature, disaggregation of disease data has focused on the spatial distribution of disease (Rahman 2012). Random forests, spatial scan statistic, and neural networks (Khan et al. 2005; Kitron et al. 2006; Mendes and Marengo 2009; Rahman 2012) have used disease related variables, such as population density, land cover, and climatic factors, to spatially distribute disease incidents from a large region of interest to several smaller geographic areas. This process provides a finer resolution of the physical location of disease. The process of temporally disaggregating disease incidents, however, has not yet been explored.

Some popular methods for temporal disaggregation include neural networks, splines, and regression (Huth 2002; Kumar et al. 2012; Herath et al. 2016). However, the most commonly used methods are those developed specifically for temporal disaggregation, mainly when dealing with economic data (Chamberlin 2010; Sax and Steiner 2013). The basic framework for these methods, known as Denton, Denton-Cholette, Chow-Lin, Fernandez, and Litterman, can be broken out into three separate steps (Chamberlin 2010; Sax and Steiner 2013). The mathematical notation of these steps and their descriptions are in Table 1.

Table 1. Variable notations and their definitions.	
Notation	Description
y	The unknown time series of interest
y_l	The known low-frequency version of y
$\hat{y}$	The disaggregated time series of y_l
p	A preliminary estimate of y
X	The matrix of related time series
C	The matrix to convert high frequency to low frequency
D	The distribution matrix
Σ	The variance-covariance matrix

- Step 1: Estimate the time-space variance-covariance matrix

The first step in the framework is to estimate the variance-covariance matrix, Σ , in the high-frequency time series space. Within this space, each dimension corresponds to a single point in time in the high-frequency time series. How Σ is calculated differs for each disaggregation method (Sax and Steiner 2013).

- Step 2: Compute a preliminary estimate, p , of the desired signal, y

Denton and Denton-Cholette (Sax and Steiner 2013) simply estimate p as $p = X$, which only works when there is a single indicator series or no indicator series. Chow-Lin, Fernandez, and Litterman (Sax and Steiner 2013) compute p as a Generalized Least Squares (GLS) estimate of the desired signal y . Specifically, $p = X\hat{B}$ where

$$\hat{B} = [X^T C^T (C \Sigma C^T)^{-1} C X]^{-1} X^T C^T (C \Sigma C^T)^{-1} y_l$$

- Step 3: Adjust p using y_l to get a final estimate, $\hat{y}$

When computing p , no restrictions are imposed in relation to the known low-frequency time series, y_l . In particular, for the final estimate $\hat{y}$, disaggregation requires that $C\hat{y} = y_l$. This can be computed using a right-pseudoinverse D of C and the low-frequency error $u_l := y_l - Cp$ in the preliminary estimation of y :

$$\hat{y} = p + Du_l$$

where

$$D = \Sigma C^T (C \Sigma C^T)^{-1}$$

From here, it can be seen that $C\hat{y} = Cp + u_l = y_l$, which is the original signal.

Using the temporal disaggregation framework described above, this technical note (TN) will down sample a time series of annual disease cases to a monthly time series.

DATA AND METHODS

Data. From 2005 – 2010, monthly records of dengue for each province in Cambodia were collected from the University of Pittsburgh’s Project Tycho (van Panhuis et al. 2014). Dengue is a vector-borne disease that is spread by the *Aedes aegypti* and *Aedes albopictus* mosquitoes and is typically found within tropical and sub-tropical areas of world. There is currently no vaccine for dengue. While early detection and proper treatment can severely lower fatality rates, dengue has become a leading cause of hospitalization and death among both children and adults in most Asian and Latin American countries (WHO 2019).

Environmental risk to dengue, or any mosquito-borne disease, is related to the parameters of precipitation, temperature, land cover, and vegetation health, all of which affect the habitation of mosquitoes (Chitnis et al 2012; Stuckey et al. 2014; Parham et al. 2015; Siraj et al. 2015). Monthly records of these parameters were obtained from remotely sensed data (Table 2), with Normalized Difference Vegetation Index (NDVI) considered a measure of vegetation health. Once collected, the environmental parameters were sub-setted to each Cambodian province and averaged in order to form a non-spatial time series.

Table 2. Remotely sensed environmental variables and their sources. Variables were spatially averaged for each province for every month of analysis.	
Environmental Variable	Data Source
Normalized Difference Vegetation Index (NDVI)	MODIS (MOD43A4)
Precipitation Sum	CHIRPS
Precipitation Mean	CHIRPS
Precipitation Sum Lagged	CHIRPS
Precipitation Mean Lagged	CHIRPS
Land Surface Temperature (LST) Day Mean	MODIS (MOD11A1)
LST Day Minimum	MODIS (MOD11A1)
LST Day Maximum	MODIS (MOD11A1)
LST Night Mean	MODIS (MOD11A1)
LST Night Minimum	MODIS (MOD11A1)
LST Night Maximum	MODIS (MOD11A1)

Within the dengue time series, there were five provinces that did not have continuous dengue records: Kracheh, Stueng Traeng, Mondol Kiri, and Krong Pailin. In order to only use continuous time series, these provinces were deleted from the dataset. A map of the Cambodian provinces used in the analysis can be seen in Figure 1.

Figure 1. Cambodian provinces used for temporal disaggregation outlined in black.

Method. The first step of the analysis was to determine which environmental variables were temporally correlated to dengue. To do this, the dengue and environmental time series of each Cambodian province were averaged together to create a mean, country level time series of dengue and its predictors. The mean environmental time series were then correlated to the mean dengue time series. Environmental variables that had an R-coefficient of 0.3 or higher, and were significant at $\alpha=0.05$, were kept for disaggregation. Additionally, variables were tested for collinearity. If a pair of variables were found to have a collinearity of at least 0.90, the variable that was least correlated to dengue was discarded.

The second step was to determine if the mean dengue time series was stationary. A stationary time series is one whose mean and variance are constant throughout time. It is relatively easy to predict and most forecasting models assume that a time series is either stationary or can be made stationary. For temporal disaggregation, whether or not a time series is stationary affects which disaggregation method is used. For stationary series, the Chow-Lin disaggregation method is the best choice (Chamberlin 2010; Sax and Steiner 2013). For non-stationary time series, the Fernandez and Litterman models are more optimal (Chamberlin 2010; Sax and Steiner 2013). Please note that the Denton and Denton-Cholette disaggregation methods are only used when zero or one predictor variables are available. Since there is more than one environmental variable related to dengue, these methods were not explored.

The augmented Dickey-Fuller test was run on the mean dengue time series to determine if the series was stationary. The null hypothesis of the augmented Dickey-Fuller test is that the time series is non-stationary. If the augmented Dickey-Fuller statistic is significant at $\alpha=0.05$, then the time series is considered stationary. For this study, the augmented Dickey-Fuller statistic was -3.5237 and had a p-value of 0.046. Therefore, the mean dengue time series is stationary and the appropriate disaggregation method is Chow-Lin.

After stationarity had been determined, the mean dengue time series was aggregated from monthly cases to annual cases. This was done by summing all of the dengue cases within one year. Using Chow-Lin, a model was developed for every possible combination of environmental variables. A root mean square error (RMSE) was calculated for each model and significance levels for each model's coefficients were recorded. The model with the lowest RMSE, most significant coefficients, and most visually pleasing graph was selected as the optimal model for disaggregation. This model was then run on the annually aggregated dengue cases for each individual province of Cambodia and an average RMSE was calculated. All calculations and model development were completed in Mathwork's Matlab.

RESULTS: The R-coefficients of the correlations between the mean dengue time series and the mean environmental predictors are listed in Table 3. The variables with an absolute R-coefficient greater than 0.3 are as follows: NDVI, precipitation sum, precipitation mean, and Land Surface Temperature (LST) night maximum. Of these variables, precipitation sum and precipitation mean were highly correlated to each other. Because of this, precipitation mean was not kept as an environmental predictor since precipitation sum was more correlated to dengue.

Table 3. R-coefficient between the mean dengue time series and the mean environmental variables time series.	
Environmental Variable	R-Coefficient
Normalized Difference Vegetation Index (NDVI)	-0.35
Precipitation Sum	0.48
Precipitation Mean	0.47
Precipitation Sum Lagged	0.25
Precipitation Mean Lagged	0.25
Land Surface Temperature (LST) Day Mean	0
LST Day Minimum	0.13
LST Day Maximum	-0.09
LST Night Mean	-0.29
LST Night Minimum	-0.16
LST Night Maximum	-0.36

A total of seven models, plus a null model that evenly distributed dengue cases throughout the year, were created from every possible combination of dengue correlated environmental variables. Of these models, only models 1, 4, and 6 had significant environmental coefficients (Table 4). This significance occurred in the models that only used NDVI, precipitation sum, and LST night maximum as variables. Of these three models with significant coefficients, model 4, the model with the precipitation sum predictor, had the lowest RMSE.

Table 4. Model name, variables, the p-value of each model's coefficients and the model's RMSE. * denotes variables that have significant coefficients at $\alpha=0.05$.			
Model Name	Model Variables	Environmental Variable Significance	RMSE
Null	Null	NaN	12.210
1	NDVI*	0.0236	13.833
2	NDVI Precipitation Sum	0.980 0.896	10.71990
3	NDVI Precipitation Sum LST Night Maximum	0.679 0.978 0.553	18.88614
4	Precipitation Sum*	0.0233	10.760
5	Precipitation Sum LST Night Maximum	0.314 0.431	24.16393
6	LST Night Maximum*	0.0312	12.80885
7	LST Night Maximum NDVI	0.444 0.336	20.54409

The graphs of the models provide a better picture of how well the mean dengue time series has been disaggregated (Figure 2). Of the three models with significant coefficients, model 4 does the best job of capturing the general seasonality of dengue. Model 6, in contrast, closely mirrors the null model and does not capture the peaks and valleys of the dengue time series. Both the null model and model 6 were only able to capture the step increases and decreases in dengue cases from year to year. Lastly, model 1, which only uses NDVI as a predictor variable, predicts a large number of dengue cases when a lower number occurred and predicts periods of low disease activity when extreme dengue activity happened. The precipitation model (model 4), however, was able to capture every step change, peak and valley in the dengue time series, and only missed the extreme amplitude of cases in 2007.

Figure 2. Graph of the mean Cambodian dengue time series, the null model, model 1, model 4, and model 6.

Based on its low RMSE score, significant predictor variable, and its ability to capture the general seasonality of the mean Cambodian time series, the Chow-Lin disaggregation method with the monthly time series of an area's total precipitation was applied to each individual province in Cambodia. The mean RMSE among the disaggregation of all these time series was 12.925. The disaggregation of four Cambodian provinces' dengue time series are in Figure 3.

Figure 3. Graphs of the dengue time series and annual dengue disaggregated using model 4 for the Cambodian provinces of a) Battambang (RMSE = 3.67), b) Banteay Meanchey (RMSE = 18.53367), c) Kandal (RMSE = 31.1175), d) Siem Reap (RMSE = 20.53035).

DISCUSSION AND SUMMARY: The variables that were the most significantly correlated to Cambodian dengue were NDVI, precipitation, and maximum night LST. The temporal disaggregation models created with these predictors did not all achieve RMSEs lower than the null model and many produced variable coefficients that were not significant. Of all the models tested, model 4, which only used precipitation data as a predictor, performed the best. Model 4 had a lower RMSE than the null model, the coefficient of the precipitation variable was significant at $\alpha=0.05$, and visually disaggregated the dengue time series better than any other model. Thus, temporal disaggregation of dengue cases can be achieved using the Chow-Lin disaggregation method with the total sum of an area's precipitation as a predictor variable. More research is

needed in this area to see if the same technique could be used for other types of vector and non-vector-borne diseases.

REFERENCES

- Beale, L., J. J. Abellan, S. Hodgson, and L. Jarup. 2008. Methodologic issues and approaches to spatial epidemiology. *Environmental Health Perspect*, 116: 1105-1110.
- Chamberlin, G. 2010. Temporal disaggregation. *Economic and Labour Market Review*, Nov. 2010: 106-121.
- Chitnis, N., D. Hardy, and T. Smith. 2012. A periodically-forced mathematical model for the seasonal dynamics of malaria in mosquitoes. *Bull Math Biol*, 74: 1098-1124.
- Herath, S. M., P. R. Sarukkalige, and V. T. V. Nguyen. 2016. A spatial temporal downscaling approach to development of IDF relations for Perth airport region in the context of climate change. *Hydrological Sciences Journal*, 61(11): 2061-2070.
- Huth, R. 2001. Statistical downscaling of daily temperature in Central Europe. *Institute of Atmospheric Physics, Prague, Czech Republic*.
- Khan, M. S., P. Coulibaly, and Y. Dibike, Y. 2005. Uncertainty analysis of statistical downscaling methods. *Journal of Hydrology*, 319: 357-382.
- Kitron, U., J. A. Clennon, M. C. Cecere, R. E. Gurtler, C. H. King, and G. Vazquez-Prokopec. Upscale or downscale: Applications of fine scale remotely sensed data to Chagas disease in Argentina and schistosomiasis in Kenya. *Geospatial Health*, 1(1): 49-58.
- Kumar, J., B. G. Brooks, P. E. Thornton, and M. C. Dietze. 2012. Sub-daily statistical downscaling of meteorological variables using neural networks. *Procedia Computer Science*, 9: 887-896.
- Mendes, D., and J. A. Marengo. 2009. Temporal downscaling: A comparison between artificial neural network and autocorrelation techniques over the Amazon Basin in present and future climate change scenarios. *Theoretical and Applied Climatology*, 100(3-4): 413-421.
- Parham, P. E., J. Waldoock, G. K. Christophides, D. Hemming, F. Agosto, K. J. Evans, N. Fefferman, H. Gaff, Ab. Gumel, S. LaDeau, S. Lenhart, R. M. Mickens, E. N. Naumova, R. S. Ostfeld, P. D. Ready, M. B. Thomas, J. Velasco-Hernandez, and E. Michale. 2015. Climate, environmental and socio-economic change: Weighing up the balance in vector-borne disease transmission. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 370: 1665.
- Rahman, A. 2017. Estimating small area health-related characteristics of populations: A methodological review. *Geospatial Health*, 12(495): 3-14.
- Sax, C., and P. Steiner. 2013. Temporal disaggregation of time series. *The R Journal*, 5(2): 80-87.
- Siraj, A. S., M. J. Bouma, M. Santos-Vega, A. K. Yeshiwondim, D. S. Rothman, D. Yadeta, P. C. Sutton, and M. Pascual. 2015. Temperature and population density determine reservoir regions of seasonal persistence in highland malaria. *Proceedings Royal Society B*, 282: 20151383.
- Stuckey, E. M., T. Smith, and N. Chitnis. 2014. Seasonally dependent relationships between indicators of malaria transmission and disease provided by mathematical model simulations. *PLoS Comput Biol*, 19(9): e1003812. *Trends in Parasitology*, 29(10): 477-482.
- van Panhuis, W. G., J. Grefenstette, S. Y. Jung, Ni. S. Chok, A. Cross, H. Eng, B. Y. Lee, V. Zadorozhny, S. Brown, D. Cummings, and D. S. Burke. 2014. Contagious diseases in the United States from 1888 to the present. *New England Journal of Medicine*, 369(22): 2152-2158.
- World Health Organization (WHO). 2019. Dengue and severe dengue. <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue> 15 April 2019.

NOTE: The contents of this technical note are not to be used for advertising, publication, or promotional purposes. Citation of trade names does not constitute an official endorsement or approval of the use of such products.