Using Local Weather and Geographical Information to Predict Cholera Outbreaks in Hanoi, Vietnam

Nguyen Hai Chau[1], Le Thi Ngoc Anh[2]

[1] Faculty of Information Technology
VNUH University of Engineering and Technology, Hanoi, Vietnam
Email: chaunh@vnu.edu.vn
[2] Information Technology Department
Hanoi Medical University, Hanoi, Vietnam
Email: lengocanh@hmu.edu.vn

Abstract. In 2007, repeated outbreaks of cholera in Hanoi have raised the need to have up-to-date evidence on the impact of factors on cholera epidemic, which is essential for developing an early warning system. We have successfully built models to predict cholera outbreaks in Hanoi from 2001 to 2012 using Random Forests method. We found that geographical factors - the number of cholera cases of a district of interest and its neighbours - are very important to predict accurately cholera cases besides the weather factors. Among weather factors, temperature and relative humidity are the most important. We also found that prediction accuracy of our models, measured in adjusted coefficient of determination, will decrease by 0.0076 if prediction length increases by one day.

Keywords: cholera outbreaks prediction, random forests, geographical information, time series.

1 Introduction

Cholera is a global public health issue despite the decrease of morbidity and mortality in recent years [**ali2012**, **sack2004**]. Cholera is an acute watery diarrhea caused by a multiplication of gram-negative toxigenic bacterium namely Vibrio Cholera (V. Cholera) in human intestine [**nguyen2009**, **who2003**]. It is estimated that approximately 200 serotypes of V.Cholera are available, of which O1 and O139 are the primary cause of cholera epidemics and endemics worldwide [**sack2004**, **who2003**]. Cholera is regularly considered in the relations with unclean water and poor sanitation infrastructures, especially in low and middle-income countries [**who2003**, **kellyhope2008**]. Annually, about 2.9 million cases and 91,000 deaths occur as a consequence of cholera infection.

Along with water and hygiene status, previous studies demonstrated that climate variability partly contributes to the widespread of V. Cholera [**emch2008.1**]. For example, studies in Africa indicated that the increase of temperature and

rainfall results in the rise of cholera cases [**mendelsohn2008**, **reyburn2011**]. Furthermore, studies in Bangladesh showed that temperature and sunshine hours might relate with cholera occurrence [**islam2009**]. In a recent report, the World Health Organization has underlined that climate variables have the central role on the temporal and spatial distribution of infectious diseases, raising the need to develop an early warning system based on meteorological factors [**kovats2003**, **who2004**]. Therefore, establishing climate-based predictive models for cholera epidemic are necessary for prompt prevention and intervention in the longer run.

Vietnam experienced cholera epidemics in the twentieth century, especially in 1960s and 1990s, and most of cases were reported in Southern regions [2]. However, in 2007 and 2008, the cholera outbreaks occurred with the majority of affected provinces in Northern regions, including Hanoi [**nguyen2009**, **gtfoc2008**, **who2008**]. Until April 2008, there were 3,271 cases reported from 18 provinces [**nguyen2009**]. The reasons for this epidemic were argued not to rest only with water or food contamination [**gtfoc2008**]. Therefore, understanding the association between cholera cases and other factors as climate variability is necessary to develop the strategies to control, monitor and prevent the cholera outbreaks.

2 Related works

Ali *et al.* [**ali2013**] studied cholera data of Matlab, Bangladesh from 1988 to 2001 and concluded that the number of cholera cases in the study area is strongly associated with local temperature and sea surface temperature (SST). Time series analysis is the method used in this research.

R. C. Reiner *et al.* [**reiner2012**] successfully build a model that is able to predict the number of cholera cases in Matlab, Bangladesh 11 months in advance. Data sets used in this research are local weather, southern oscillation index (SOI) and flooding condition from 1995 to 2008. As a result from the research, SOI and flooding condition are main positive factors to the number of cholera cases in Matlab. Prediction method used in this research is simulation using multidimensional inhomogeneous Markov chain (MDIMC).

Xu Min *et al.* [**xu2013**] used MaxEnt model, a maximum expectation like model, to analyze China's cholera outbreaks from 2001 to 2008. As their results, precipitation, temperature and the location's altitude are strongly linked to the number of cholera cases. Distance of the location to the sea coast, relative humidity and atmospheric pressure are also linked to the number of cholera cases. The sun hours and river height discharge are independent to the number of cholera cases.

In another research, Xu Min *et al.* [**xu2015**] used satellite and geographical data to find influences of SST, sea surface height (SSH) and ocean chlorophyll concentration (OCC) to the number of cholera cases in China from 1999 to 2008. Results show that changes in SST and SSH are associated immediately to the number of cholera cases while OCC has one month lag effect.

M. Emch *et al.* [**emch2008.2**] studied effects of local weather parameters to the number of cholera cases in Matlab, Bangladesh from 1983 to 2003 and in Nha Trang, Hue (Vietnam) from 1985 to 2003. Results show that high OCC have positive association to the number of cholera cases in Matlab, increasing SST correlates with the number of cholera cases in Hue and increasing of river height correlates with the number of cholera cases in Nha Trang. In the research, the authors show that local weather has two months lag effect. They use univariate and multivariate statistical analysis methods.

The above works show that local weather parameters such as temperature, relative humidity, SOI, SST, SSH have different association to number of cholera cases in different areas.

In Vietnam, several previous studies mentioned the association of local environment with occurrence of cholera cases. A study of Kelly-hope *et al.* in Vietnam suggested the significant link of precipitation and cholera outbreaks at 0-month lag during 1991-2001 [**kellyhope2008**]. Emch *et al.* found that the significant predictors of cholera infected included increasing SST and river height [**emch2008.1**]. However, the outbreak of cholera from 2007 to 2009 in Hanoi has raised the need to have more reliable evidence about the impact of climate factors along with traditional environment indicators. In order to provide more comprehensive and up-to-date evidence, this paper aimed to investigate the relationships of cholera incidence with weather, geographical factors and the SOI climate change indicator in Hanoi during 2001-2012.

The rest of this paper is organized as follows. In section 3, we describe Hanoi local information and data sets used for research. Prediction models building is presented in section 4. We analyze prediction results in section 5 and concludes the paper in section 6.

3 Description of the study area and data sets

To build prediction models of cholera prediction in Hanoi, Vietnam, we use the following data sets: Cholera cases, local weather, geographical data of Hanoi and SOI data set. In the following we describe Hanoi local information and the data sets.

3.1 Study area

Hanoi, located at 21°01′42.5"N, 105°51′15.0"E, is the capital and the second largest city of Vietnam. Its population in 2009 is approximately 2.6 millions in urban districts and 6.5 millions in metropolitan areas. From 2001 to 2012, Hanoi is divided into 11 urban districts, a district-level town and 17 metropolitan districts. The districts and the town are further subdivided into more than five hundreds communes, wards and commune-level town. For simplicity, in this research we refer administrative level 2 as "district", level 3 as "commune" assuming that Hanoi's administrative level is 1.

Hanoi is located in the northern region of Vietnam. It is embraced by the Red River and roughly 100km far from coastal area. Hanoi's climate is warm humid subtropical, identified as *Cwa* in Köppen climate classification system. It has four distinct seasons: spring (February-April), summer (May-August), fall (September-October) and winter (November-January).

3.2 Geographical data sets

The data set contains administrative boundaries of districts and communes, roads, rivers and water areas at 1:50,000 scale. Map of Hanoi and its 29 districts is in Fig. 1.

Fig. 1. Hanoi administrative map. The black circle indicates location of Lang weather station.

3.3 Cholera data set

The data set contains all observed cholera cases of Hanoi from Jan 01, 2001 to Dec 31, 2012. Each record of the data set contains patient's name, age, sex, date of infection and his/her home address at least to commune administrative level. We aggregate the data set to calculate the number of daily cholera case per district. Daily aggregated cholera cases of Hanoi and daily aggregated cholera

cases per district are in Fig. 2. The data set shows that there were five cholera outbreaks in Hanoi in 2004, 2007, 2008, 2009 and 2010. The other years are cholera-free.

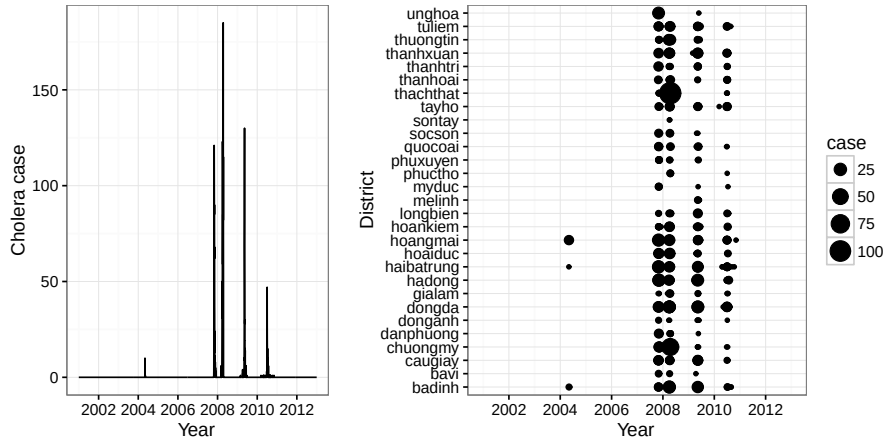

Fig. 2. Cholera cases of Hanoi from 2001 to 2012. Left: Daily cholera cases of Hanoi. Right: daily cholera cases by district; size of black circles is proportional to the number of cholera cases.

3.4 Local weather data set

This data set contains daily relative humidity (min, max and mean), daily temperature (min, max and mean), daily sun hours, daily wind speed and daily precipitation measured by Lang weather station in Hanoi. The Lang weather station locates in Cau Giay district of Hanoi as shown in Fig. 1 and its measurements are representative for Hanoi weather. Fig. 3. illustrates the local weather data set.

3.5 SOI data set

We use SOI data collected from a website of Queensland government, Australia [**soi**]. The data set contains daily SOI measurement from 1991 to the day of getting data. Daily SOI data from 2001 to 2012 is plotted in Fig. 4.

4 Features selection and models building

4.1 Features selection

In previous works of cholera prediction, researchers often use monthly aggregated data as their main data sets. We use daily aggregated data. The reasons include

sparsity and the number of data points in our monthly aggregated data set. If we aggregate the cholera and weather data sets by month, we have 144 observations among which 19 are with cholera. In addition, the maximum length of each cholera outbreak of Hanoi in month is only three, making "positive" data series very short and difficult to predict.

We aggregate data sets described in section 3, except the geographical data set, by day and mix them into one. We consequently have a final data set, denoted as FS, with 35 variables and 4383 observations. Six of the variables are weather ones including temperature, relative humidity, precipitation, sun hour, wind speed and SOI. The others are daily cholera cases of 29 districts. Fig. 5 shows a correlogram of the FS data set. It is obvious that the weather variables have very close to zero correlation with the cholera case variables. However the cholera case variables are correlated. Cholera case variables of geographically closed districts are generally more correlated. This fact suggests us to use subsets of cholera case variables as additional predictors in combination with weather variables for building prediction models. However, the cholera case variables of districts are also outcomes. Therefore, we can only use the past values of cholera variables for prediction. In the following, we describe the building of our prediction models.

4.2 Models building

While other studies on cholera outbreak prediction mainly consider study areas as atomic, we use districts of Hanoi as geographical units. For each district of Hanoi, we build three predictive models namely complete, weather-independent and geographical-independent, abbreviated as CP, WI and GI, respectively. Our purpose is twofold. Firstly, we want to choose the best model for all districts. Secondly, we want to evaluate effects of geographical and weather variables to the accuracy of prediction. Table 1 explains the predictor groups of each model. All the models has the number of cholera case as an outcome.

Each model has a lag parameter l measured in day. The parameter means that we use the number of cholera cases of the current and previous $l-1$ days in a district of interest as a predictor for the model. It also means that we predict the number of cholera case of the district in next l days. For each model, we also use the past number of cholera cases of all neighbours of the district of interest and past weather information as additional predictors. Two districts are neighbours if they share parts of their administrative borders. Using SQL spatial queries in PostgreSQL/PostGIS we easily define neighbours of each district in Hanoi.

Instead of using statistical techniques for model building in many other studies, we adopt a machine learning approach. After a process of try and error, we found that the Random Forests (RF) regression method is the most suitable for our prediction models. RF regression is a supervised learning method. It learns on training data sets and predict on testing ones. Since all variables of the FS data set is time series, we apply the rolling forecasting origin techniques by Hyndman and Athanasopoulos [**hyndman2013**]. Using this technique, we first create an initial window that has s_1 data points as the first training data

set. The testing data set is next s_2 data point. Note that each data point in the training data set contains all predictors and the outcome, and each data point in the testing data set contains only predictors. The window is then shifting along the time axis until no data point left. The supervised model is built during the shifting and improved along the time axis. We set $s_1 = s_2 = l$ in all models.

Table 1. Description of predictors for CL, WI and GI models.

Predictor group	**Model**		
	CP	**WI**	**GI**
Weather information	- Mean temperature - Mean relative humidity - Precipitation - Sun hours - Wind speed - SOI	*Not used*	- Mean temperature - Mean relative humidity - Precipitation - Sun hours - Wind speed - SOI
Geographical information	- The number of cholera cases of a district D - The number of cholera cases of D's neighbour districts	- The number of cholera cases of a district D - The number of cholera cases of D's neighbour districts	*Not used*

Table 2 illustrates training and testing data sets of a CP model with lag parameter $l = 3$ of a district D. In the table, $w_1, w_2, ..., w_8$ are values of weather variables, $n_1, n_2, ..., n_8$ are values of cholera cases of all D's neighbours; and $d_4, d_5, ..., d_{11}$ are values of cholera cases of D. Indices of these variables indicate points in time. The model's first training data set is $\{w_1, w_2, w_3, n_1, n_2, n_3, d_4, d_5, d_6\}$ and first testing data set is $\{w_4, w_5, w_6, n_4, n_5, n_6\}$. The outcome of this model is $\{d_7, d_8, d_9\}$. Point in time to start training the CP model is 6.

Table 2. An example of sliding window training and testing for the CP model. In this example, lag parameter l is 3.

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}
n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8
Training data set 1			**Testing data set 1**				
	Training data set 2			**Testing data set 2**			
		Training data set 3			**Testing data set 3**		

5 Prediction results analysis

We built 29×3 RF regression models based on the FS data sets for 29 districts. Measures for regression models assessment are often root mean square error (RMSE) and adjusted coefficient of determination (adj-R^2) [**hyndman2013**]. We calculate RMSE and adj-R^2 for all 29×3 models. In the following subsections, we first compare effects of weather and geographical predictors based on CP, WI and GI models and the two measures RMSE, adj-R^2. We then perform statistical analysis to find relationship of prediction accuracy and prediction length; and evaluate the importance of weather variables in 29×3 RF models.

5.1 Effects of weather predictors and geographical predictors to the accuracy of prediction

To compare the effect of weather and geographical predictors to prediction accuracy, measured in RMSE and adj-R^2, we use Tukey method for 3, 7, 14 and 30-day in advance prediction.

RMSE comparison in Fig. 6 does not show any statistical difference in models' prediction accuracy: all the 95% confidence interval contains 0. In addition, p-values of RMSE comparison model are larger than 0.05. Thus using RMSE we cannot define which model among CP, WI and GI is the best. We will use adj-R^2 for models comparison.

With reference to GI-CP and WI-CP means and confidence interval in Fig. 7, we see that the CP models with highest adj-R^2 are the best. The GI models having the smallest adj-R^2 are the worst. It means that the number of cholera case in neighbours of a district strongly influences the number of cholera cases in that district. Fig. 8 compares 3-day in advance prediction accuracy of CP, WI and GI models for Badinh, Dongda, Socson, Thanhxuan and Unghoa districts.

5.2 Relationship of prediction accuracy and prediction length

As mentioned above, the CPs are best models. We then apply CP models for final prediction of cholera cases on 29 districts of Hanoi, compare prediction results to observed data for $l = 3, 7, 14, 30$ and calculate adj-R^2 measures. Details of adj-R^2 measures are in table 3. To observed change of prediction accuracy versus length of prediction, we built a multiple linear regression model. The model's predictors are districts and the number of day to predict in advance. Outcome of the model is the adj-R^2 measure. Details of the model is in Listing 1.1.

Listing 1.1. Model for relationship of prediction accuracy and prediction length.

```
## Call:
## lm(formula = adjrsq ~ district + day, data = tmp)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.16773 -0.06615 -0.01686  0.05959  0.25176
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.2846737  0.0538533   5.286 9.34e-07 ***
## districtbavi       -0.1765346  0.0740029  -2.386   0.0193 *
## ...
## districtunghoa     -0.1753860  0.0740029  -2.370   0.0200 *
## day                -0.0076452  0.0009427  -8.110 3.17e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1047 on 86 degrees of freedom
## Multiple R-squared:  0.6179, Adjusted R-squared:  0.4891
## F-statistic: 4.796 on 29 and 86 DF,  p-value: 7.651e-09

confint(fit, 'day')
##              2.5%        97.5%
## day -0.009519223 -0.005771201
```

The multiple linear regression model shows that, if everything else remains and prediction length increases by one day, the adj-R^2 measure will decrease by 0.0076 with $[-0.0095, -0.0057]$ 95% confidence interval. The p-value is very small (3.17^{-12}) indicating the statistical significant of the multiple linear model. The model's explained variation is 49%. Fig. 9 compare accuracy of CP models for 3, 7 and 14-day in advance prediction.

5.3 Importance of weather variables

From the 29 CP models built by RF regression method, we easily extract the importance of weather variables. The boxplots in Fig. 10 show that the daily mean temperature, relative humidity are the most important weather variables with approximately 50% importance in comparison to other weather variables and cholera case variables. The precipitation and sun hour variables are about 30-35% importance. The wind speed and SOI are less than 20% importance.

6 Conclusions

We have successfully built supervised learning models based on RF regression method for short-term prediction of cholera outbreaks in Hanoi from 2001 to 2012. To the best of our knowledge, this research is probably the first one that uses a machine learning approach to predict cholera outbreaks in Hanoi, Vietnam. The models built in this paper show that geographical factors of Hanoi, here are the numbers of cholera cases of a district of interest and its geographical neighbours, are very important ones to predict cholera outbreaks besides the weather variables. Among weather variables, daily mean temperature and daily mean relative humidity are the most important. SOI is least important weather factor to cholera in Hanoi. We also found that the prediction accuracy

Table 3. Adj-R^2 of each district for difference prediction length using CP models.

	Prediction length			
District	30-day	14-day	7-day	3-day
badinh	0.0092	0.0448	0.3126	0.3592
bavi	-0.0001	0.0015	-0.0002	0.0186
caugiay	0.0107	0.0623	0.2323	0.4142
chuongmy	0.0015	0.1117	0.0745	0.2343
danphuong	0.0003	0.0088	0.0177	0.2331
donganh	-0.0002	0.0434	0.0085	0.0476
dongda	0.0491	0.1325	0.4133	0.6214
gialam	-0.0002	0.0007	0.0509	0.0831
hadong	0.0106	0.0314	0.2248	0.3998
haibatrung	0.0085	0.0719	0.209	0.4544
hoaiduc	0.0037	0.0386	0.2188	0.5297
hoankiem	0.024	0.0739	0.2466	0.4122
hoangmai	0.0254	0.0458	0.1821	0.3372
longbien	0.007	0.0121	0.1152	0.2316
melinh	-0.0001	-0.0002	0.0666	0.0017
myduc	-0.0002	0.0171	0.0169	0.3762
phuxuyen	-0.0001	0.0003	0.0335	0.0281
phuctho	0.0005	0.0214	0.0248	0.013
quocoai	0.0012	0.0161	0.0553	0.0785
socson	0.002	0.007	-0.0002	0.0642
sontay	-0.0002	-0.0002	-0.0002	-0.0002
tayho	0.0086	0.0735	0.2706	0.4905
thachthat	0.0005	-0.0002	0.0051	0.15
thanhoai	0.0094	0.0439	0.1588	0.2535
thanhtri	0.0152	0.0159	0.0386	0.0967
thanhxuan	0.0277	0.0551	0.287	0.4969
thuongtin	0.0196	0.1813	0.1243	0.3211
tuliem	0.0137	0.0699	0.2345	0.4197
unghoa	-0.0002	0.0137	0.0008	0.01

of our models, measured in adj-R^2, will decrease by 0.0076 if prediction length increases by one day.

Acknowledgement

The authors wish to thank Assoc. Prof. Nguyen Ha Nam from VNU University of Engineering and Technology for his discussions and recommendations during experiments.

References

1. M. Ali, A. L. Lopez, Y. A. You *et al.*, *The global burden of cholera*, Bulletin of the World Health Organization **90** (3), 2012, 209–218A.
2. D. A. Sack, R. B. Sack, G. B. Nair, A. K. Siddique, *Cholera*, Lancet **363** (9404), 2004, 223–233.
3. B. M. Nguyen, J. H. Lee, N. T. Cuong *et al.*, *Cholera outbreaks caused by an altered Vibrio cholerae O1 El Tor biotype strain producing classical cholera toxin B in Vietnam in 2007 to 2008*, Journal of clinical microbiology **47** (5), 2009, 1568–1571.
4. The World Health Organization, *Cholera*, Geneva, Switzerland, 2003.
5. L. A. Kelly-Hope, W. J. Alonso, V. D. Thiem*et al.*, *Temporal trends and climatic factors associated with bacterial enteric diseases in Vietnam 1991-2001*, Environmental health perspectives **116** (1), 2008, 7–12.
6. M. Emch, C. Feldacker, M. S. Islam, M. Ali, *Seasonality of cholera from 1974 to 2005: a review of global patterns*, International journal of health geographics, 2008, 7–31.
7. J. Mendelsohn, T. Dawson, *Climate and cholera in KwaZulu-Natal, South Africa: the role of environmental factors and implications for epidemic preparednessi*, International journal of hygiene and environmental health **211** (1-2), 2008, 156–162.
8. R. Reyburn, D. R. Kim, M. Emch, A. Khatib, L. von Seidlein, M. Ali, *Climate variability and the outbreaks of cholera in Zanzibar, East Africa: a time series analysis*, The American journal of tropical medicine and hygiene **84** (6), 2011, 862–869.
9. M. S. Islam, M. A. Sharker, S. Rheman *et al.*, *Effects of local climate variability on transmission dynamics of cholera in Matlab, Bangladesh*, Transactions of the Royal Society of Tropical Medicine and Hygiene **103** (11), 2009, 1165-1170.
10. R. S. Kovats, M. J. Bouma, S. Hajat, E. Worrall, A. Haines, *El Nino and health*, Lancet **362** (9394), 2003, 1481–1489.
11. The World Health Organization, *Using Climate to Predict Infectious Disease Outbreaks: A Review*, Geneva, Switzerland, 2004.
12. The World Health Organization, Control GTFoC. *Cholera country profile: Vietnam*. Geneva, Switzerland, 2008.
13. The World Health Organization, *Outbreak news: Severe acute watery diarrhoea with cases positive for Vibrio cholerae, Vietnam. Releve epidemiologique hebdomadaire / Section d'hygiene du Secretariat de la Societe des Nations = Weekly epidemiological record /Health Section of the Secretariat of the League of Nations*, **83** (18), 2008, 157–158.
14. M. Ali, D. R. Kim, M. Yunus, M. Emch, *Time Series Analysis of Cholera in Matlab, Bangladesh, during 1988-2001*, Journal of Health, Population and Nutrition **31** (1), 2013, 11–19.
15. R. C. Reiner, A. A. King, M. Emch, M. Yunus, A. S. G. Faruque, M. Pascual, *Highly localized sensitivity to climate forcing drives endemic cholera in a megacity*, Proc. Natl. Acad. Sci. USA **109**, 2012, 2033–2036.
16. Xu Min , Cao ChunXiang, Wang DuoChun, Kan Biao, Jia HuiCong, Xu YunFei, Li XiaoWen, *District prediction of cholera risk in China based on environmental factors*, Chinese Science Bulletin **58** (23), 2013, 2798–2804.
17. Min Xu, Chunxiang Cao, Duochun Wang, and Biao Kan, *Identifying Environmental Risk Factors of Cholera in a Coastal Area with Geospatial Technologies*, Int. J. Environ. Res. Public Health **12**, 2015, 354–370.

18. M. Emch, C. Feldacker, M. Yunus *et al.*, *Local Environmental Predictors of Cholera in Bangladesh and Vietnam*, The American Journal of Tropical Medicine and Hygiene **78** (5), 2008, 823–832.
19. *Daily SOI data set of the Queensland, Australia*, available online at `https://www.longpaddock.qld.gov.au/seasonalclimateoutlook/southernoscillationindex/soidatafiles/DailySOI1887-1989Base.txt`
20. R. Hyndman, G. Athanasopoulos, *Forecasting: principles and practice*, Otexts, 2013.

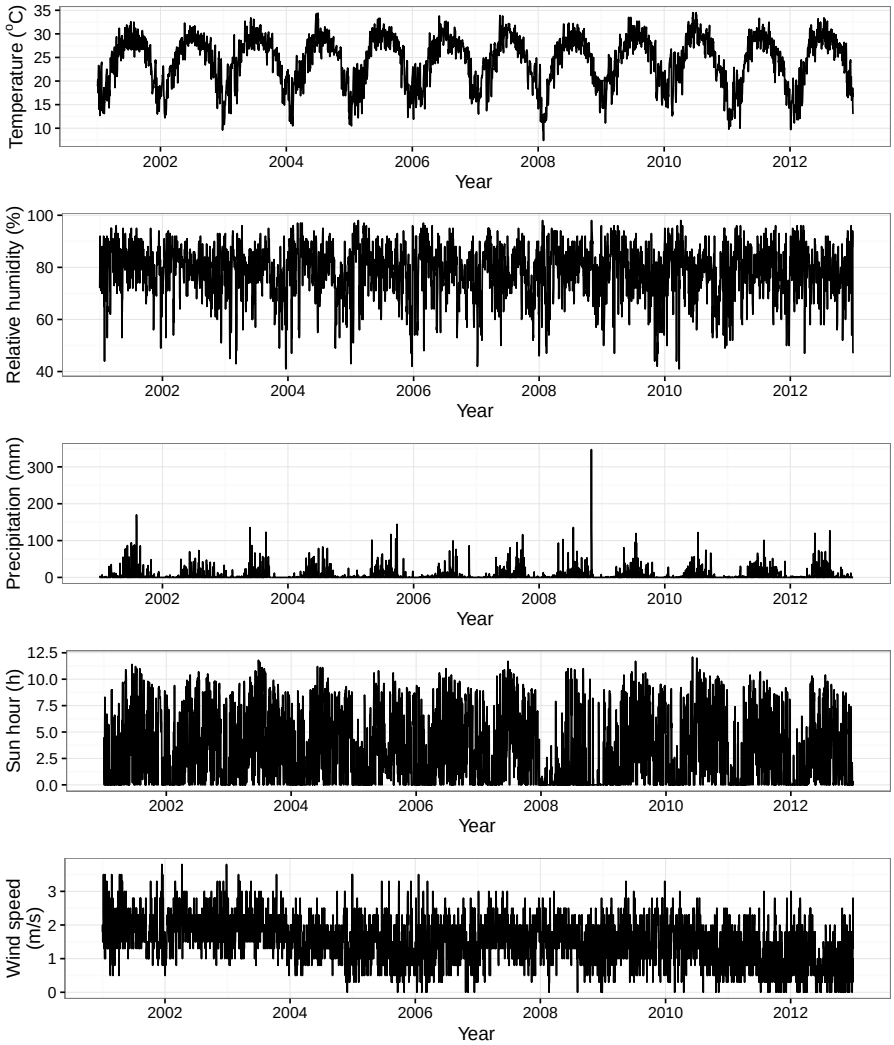

Fig. 3. Daily local weather data of Hanoi from 2001 to 2012 as measured by Lang weather station. From top to bottom: Mean temperature, mean relative humidity, precipitation, sun hour and wind speed.

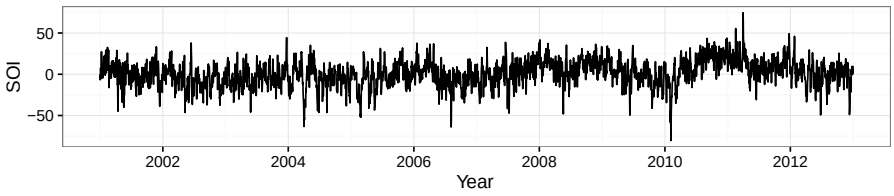

Fig. 4. Daily SOI from 2001 to 2012.

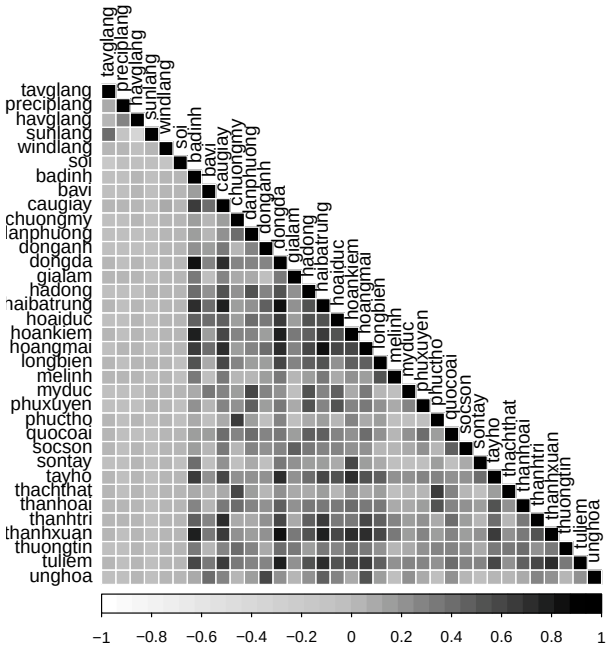

Fig. 5. Correlogram of the FS data set. Temperature, relative humidity, precipitation, sun hour, wind speed and SOI variables are denoted as `tavglang`, `havglang`, `preciplang`, `sunlang`, `windlang` and `soi`, respectively. Other variables, named after districts names, are daily cholera cases corresponding to districts.

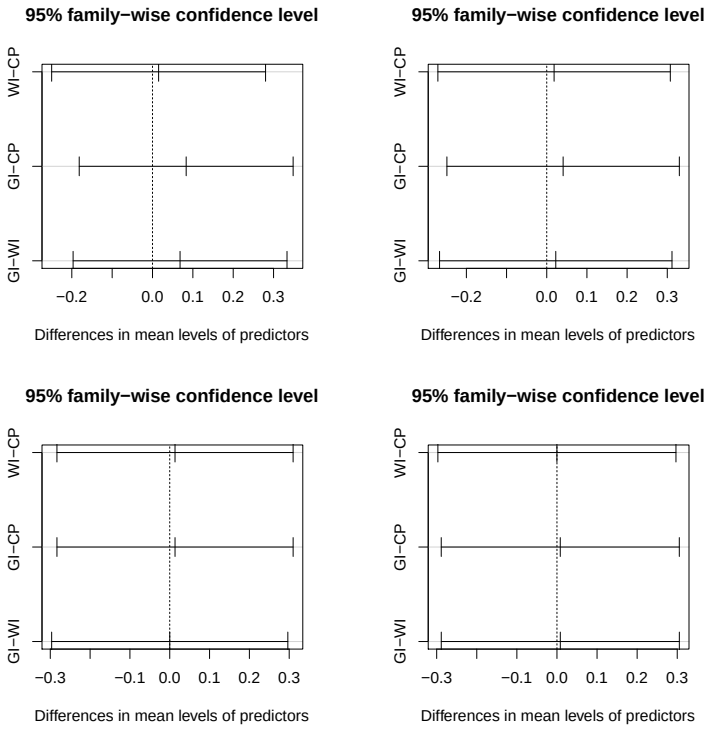

Fig. 6. Tukey multiple comparison of RMSE of CP, WI and GI models. From left to right, top to bottom: Comparison for 3, 7, 14 and 30-day in advance prediction.

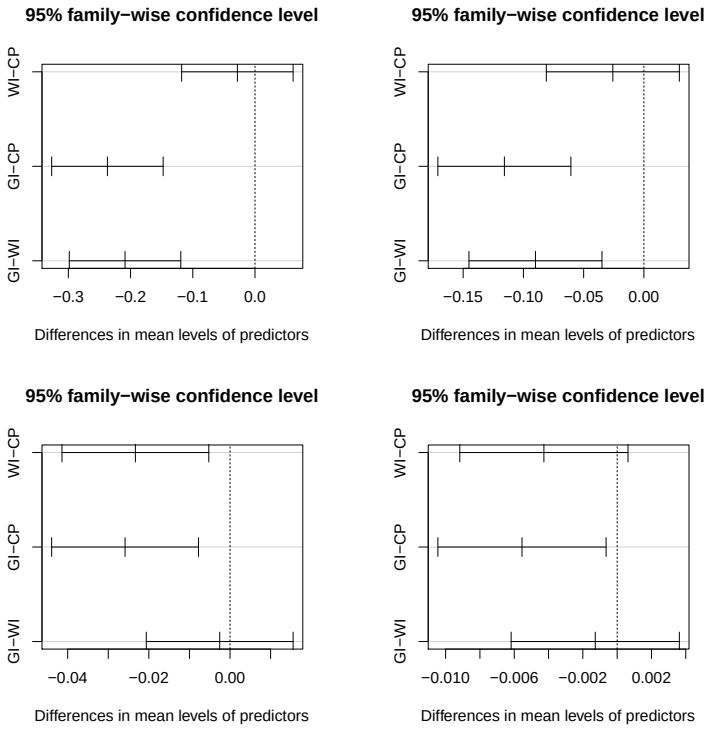

Fig. 7. Tukey multiple comparison of adj-R^2 of CP, WI and GI models. From left to right, top to bottom: Comparison for 3, 7, 14 and 30-day in advance prediction.

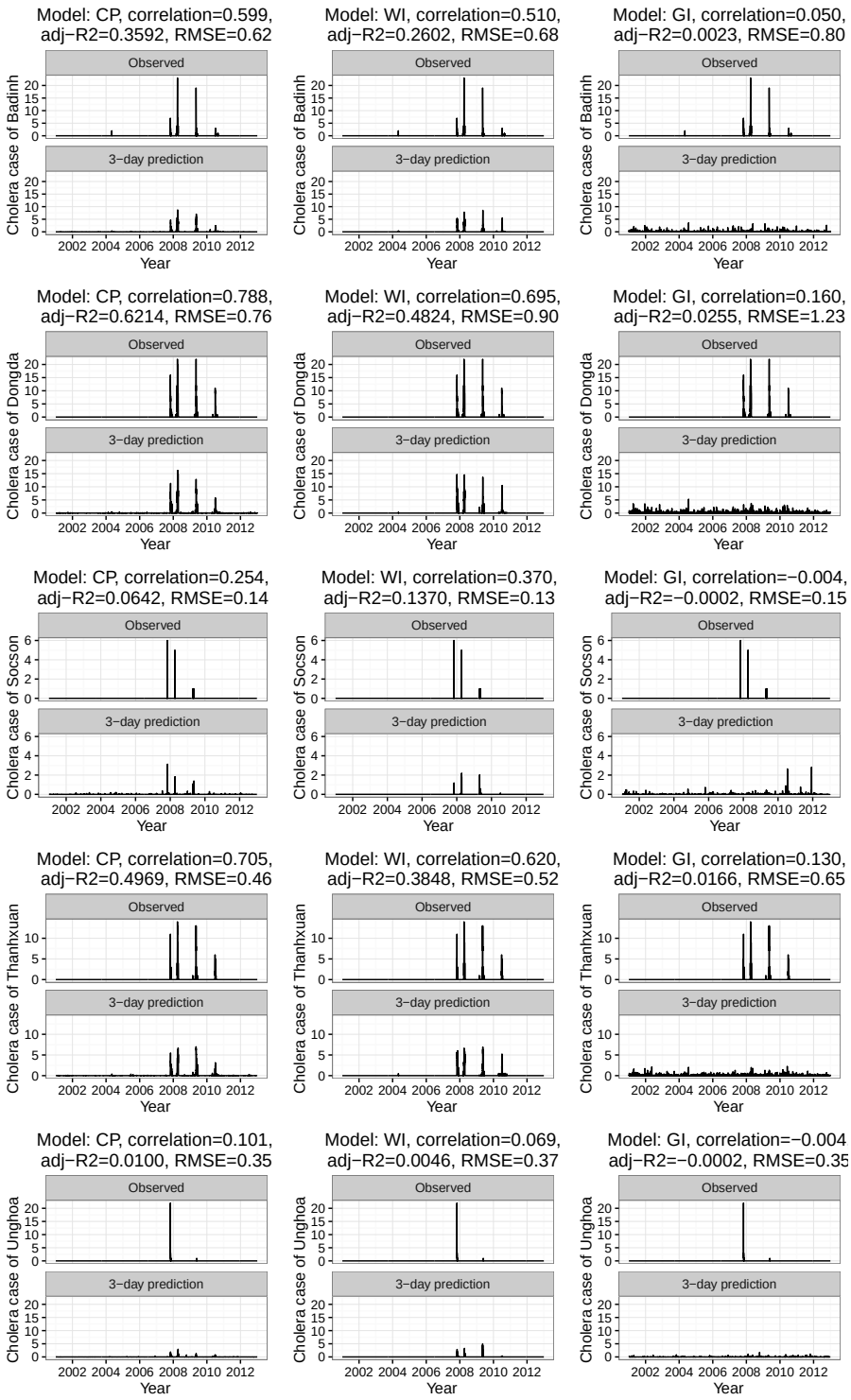

Fig. 8. Comparison of 3-day in advance prediction accuracy of CP, WI and GI models for Badinh, Dongda, Socson, Thanhxuan and Unghoa districts.

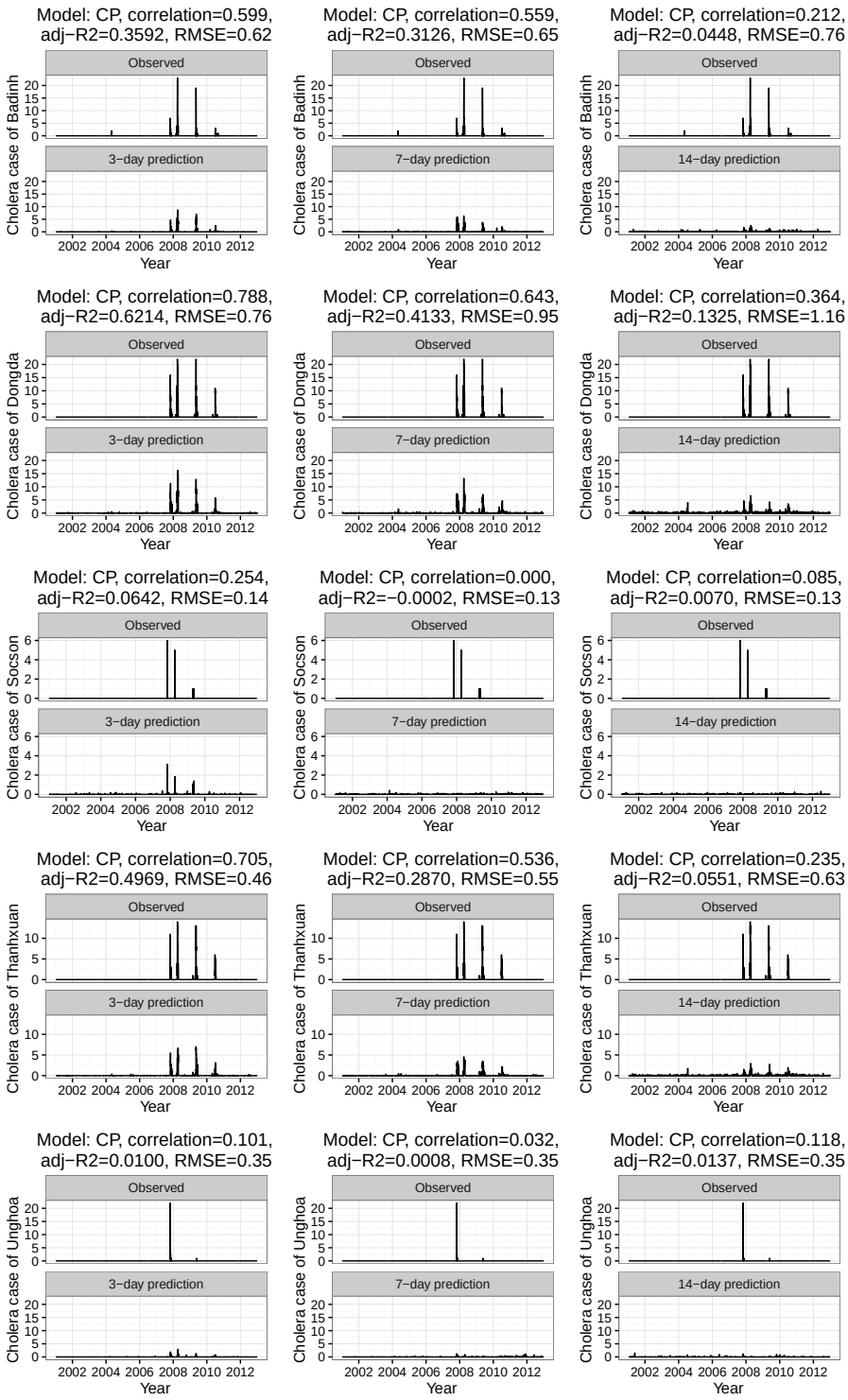

Fig. 9. Comparison of accuracy of CP models for 3, 7 and 14-day in advance prediction for Badinh, Dongda, Socson, Thanhxuan and Unghoa districts.

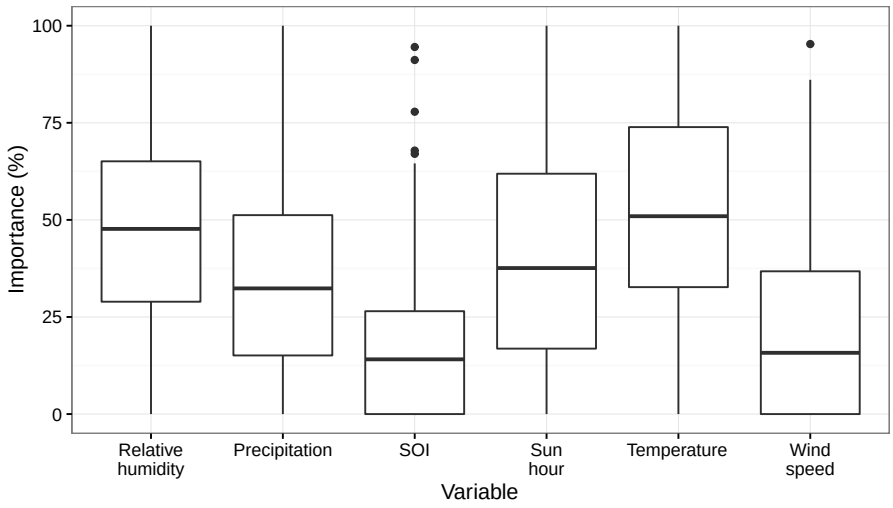

Fig. 10. Importance of weather variables to CP models.