Enhancing Cholera Outbreaks Prediction Performance in Hanoi, Vietnam Using Solar Terms and Resampling Data

Nguyen Hai Chau

Faculty of Information Technology
VNUH University of Engineering and Technology, Hanoi, Vietnam
chaunh@vnu.edu.vn

Abstract A solar term is an ancient Chinese concept to indicate a point of season change in lunisolar calendars. Solar terms are currently in use in China and nearby countries including Vietnam. In this paper we propose a new solution to increase performance of cholera outbreaks prediction in Hanoi, Vietnam. The new solution is a combination of solar terms, training data resampling and classification methods. Experimental results show that using solar terms in combination with ROSE resampling and random forests method delivers high area under the Receiver Operating Characteristic curve (AUC), balanced sensitivity and specificity. Without interaction effects the solar terms help increasing mean of AUC by 12.66%. The most important predictor in the solution is Sun's ecliptical longitude corresponding to solar terms. Among the solar terms, `frost descent` and `start of summer` are the most important.

Keywords: Cholera outbreaks prediction, Solar terms, Resampling.

1 Introduction

Cholera - an acute diarrhea disease - remains a global threat, especially in developing countries. The World Health Organization (WHO) estimates that every year, about 3-5 millions people are affected by cholera worldwide [1]. Prediction of cholera outbreaks helps mitigating their consequences. At present, cholera outbreaks prediction is still a difficult problem. Researchers have found relationships of cholera outbreaks and environmental factors [2] in Bangladesh, China and Vietnam. The factors are used as predictors for cholera outbreaks prediction models.

In Bangladesh, researchers have found that the number of cholera cases strongly associated with local temperature and sea surface temperature (SST) [3]; local weather, southern oscillation index (SOI) and flooding condition [4]; and ocean chlorophyll concentration (OCC) [5].

In China, statistical evidences showed that precipitation, temperature and location altitude, relative humidity, atmospheric pressure [6]; SST, sea surface height (SSH) and OCC [7] are linked to number of cholera cases.

In Nha Trang and Hue of Vietnam, precipitation [8]; SST and river height [5] are correlated with number of cholera cases. Recently, there are attempts to predict cholera outbreaks in Hanoi, Vietnam [9,10] using machine learning models with local weather and SOI data. In these models, temperature and relative humidity are the most important factors to predict cholera outbreaks.

A common difficulty for prediction of cholera outbreaks in Hanoi is an imbalanced data set [9,10]. To deal with this difficulty, Le et al. [9] only use monthly data for cholera-present years of 2004 and 2007-2010. This approach makes good prediction models, but only for cholera-present years. For cholera-free years, these models are not suitable. Another approach is to use the number of cholera cases in the past days of a district and its geographical neighbours to make daily prediction models [10]. These models performance is good during outbreaks but still moderate when predicting the first cases of outbreaks.

In this paper, we propose a new model to enhance the performance of daily cholera outbreaks prediction in Hanoi. We resample our training data sets to overcome disadvantages of imbalanced data set. Furthermore, we use additional season information to increase prediction performance of classification models.

2 Study Area and Data Sets

2.1 Study Area

Our study area is Hanoi, the capital city of Vietnam. Hanoi is located at $21°01'$N, $105°51'$E. In 2016, Hanoi's population is about 7.5 millions. Hanoi's weather is warm humid subtropical, classified as *Cwa* in Köppen climate system.

2.2 Cholera Data Set

We obtained a raw cholera data set from Hanoi Grant 01C-08/-8-2014-2 project [9]. The data set consists of observed cholera cases in Hanoi from Jan 01, 2001 to Dec 31, 2012. Each observation contains date, patient's name, age, gender and home address. From the raw data set we aggregated and created a derived data set where each observation has only date and number of cholera cases in Hanoi (`epi` variable). The number of cholera cases is transformed to `yes`/`no` levels.

2.3 Local Weather Data Set

The local weather data set contains daily weather data from Jan 01, 2001 to Dec 31, 2012. Each record of the data set contains average temperature in Celsius degree, average relative humidity in percentage, daily sun hours, daily average wind speed in m/s and daily precipitation in mm. Corresponding variables names are `tavg`, `havg`, `sun`, `wind` and `precip`, respectively.

2.4 Southern Oscillation Index Data Set

We obtain SOI data from a website of Queensland government, Australia [11]. The data set contains daily SOI measurement (`soi` variable) from 1991 to date.

2.5 Solar Terms Data Set

A solar term [12,13] is an ancient Chinese concept describing a point of change in seasonal cycles. Solar terms are 15° separated along the apparent path of the Sun on the celestial sphere and are used in lunisolar calendars to synchronize with the seasons.

Although originated from China, solar terms are used in Japan, Korea and Vietnam. In Vietnam, solar terms are named *tiết khí*. Solar terms' Gregorian date and other corresponding climatology information from 1947 to date is available at Hong Kong Observatory [14]. We obtain solar terms information of years 2001 to 2012 from the observatory and use it as additional season data. The data includes solar terms names, corresponding Gregorian date and ecliptical longitude (EC) in degree. We describe solar terms in Table 1.

Table 1. List of solar terms.

Solar term	**Vietnamese**	**Sun's EC**	**Gregorian date**
Start of spring	Lập xuân	315°	Feb 3–5
Rain water	Vũ thủy	330°	Feb 18–20
Awakening of insects	Kinh trập	345°	Mar 5–7
Vernal equinox	Xuân phân	0°	Mar 20–21
Clear and bright	Thanh minh	15°	Apr 4–6
Grain rain	Cốc vũ	30°	Apr 19–21
Start of summer	Lập hạ	45°	May 5–7
Grain full	Tiểu mãn	60°	May 20–22
Grain in ear	Mang chủng	75°	Jun 5–7
Summer solstice	Hạ chí	90°	Jun 21–22
Minor heat	Tiểu thử	105°	Jul 6–8
Major heat	Đại thử	120°	Jul 22–24
Start of autumn	Lập thu	135°	Aug 7–9
Limit of heat	Xử thử	150°	Aug 22–24
White dew	Bạch lộ	165°	Sep 7–9
Autumnal equinox	Thu phân	180°	Sep 22–24
Cold dew	Hàn lộ	195°	Oct 8–9
Frost descent	Sương giáng	210°	Oct 23–24
Start of winter	Lập đông	225°	Nov 7–8
Minor snow	Tiểu tuyết	240°	Nov 22–23
Major snow	Đại tuyết	255°	Dec 6–8
Winter solstice	Đông chí	270°	Dec 21–23
Minor cold	Tiểu hàn	285°	Jan 5–7
Major cold	Đại hàn	300°	Jan 20–21

3 Data Preprocessing

We merge the cholera, local weather, and SOI into one (refers as MDS) with reference to the `date` variable as the primary key. A sample of the MDS with `date` variable is in Table 2. We use MDS for cholera outbreaks prediction. The `epi` variable, indicating cholera status of a particular date (`yes` or `no`), is the outcome. Predictors are `tavg, havg, precip, sun, wind` and `soi` variables. The `date` variable is not used directly. We derive seasonal information from `date` to make new variables. The new variables are `week` (week number in a year), `month` (month) and `solarterm, ec` (solar terms and corresponding Sun's ecliptical longitude, available at Hong Kong observatory [13,14]). For convenience in reference we abbreviate `solarterm`, `ec` variables as `solar`.

Table 2. A sample of MDS with `date` variable in a cholera outbreak of 2004.

date	tavg	havg	precip	sun	wind	soi	epi
2004-05-01	28.20	82	0.00	4.40	2.30	17.48	no
2004-05-02	28.60	84	0.00	3.70	2.30	-3.74	yes
2004-05-03	28.90	82	0.00	5.20	2.50	-15.84	yes
2004-05-04	24.00	83	67.00	0.10	1.80	3.16	yes
2004-05-05	21.80	71	0.00	0.00	1.80	18.25	yes
2004-05-06	22.30	77	0.00	0.80	1.30	11.13	yes

There are 24 solar terms. Therefore the `solarterm` and `ec` variables in MDS have a lot of null values. We process null values for `solarterm` and `ec` as follows: Each null value of `solarterm` is set to the closest solar term in the past. Not null values of `ec` (or reference values) are $0, 15, 30, \ldots, 360$ (refer to Table 1). Each null value of `ec` in a date d is set to its closest reference value in the past in a date d_0 plus $d - d_0$. A sample of MDS with seasonal variables is in Table 3.

Table 3. The same sample as in Table 2 with seasonal variables `week, month` and `solar`. A reference value of `ec` is 45 and corresponding `solarterm` is `start of summer`.

week	month	ec	solarterm	tavg	havg	precip	sun	wind	soi	epi
18	5	41	grain rain	28.20	82	0.00	4.40	2.30	17.48	no
18	5	42	grain rain	28.60	84	0.00	3.70	2.30	-3.74	yes
18	5	43	grain rain	28.90	82	0.00	5.20	2.50	-15.84	yes
18	5	44	grain rain	24.00	83	67.00	0.10	1.80	3.16	yes
18	5	45	start of summer	21.80	71	0.00	0.00	1.80	18.25	yes
19	5	46	start of summer	22.30	77	0.00	0.80	1.30	11.13	yes

4 Design and Analysis of Experiments

4.1 Design of Experiments and Measure Selection

We need to perform classification tasks to predict the `epi` outcome variable of a day $d+1$ from predictors `tavg, havg, precip, sun, wind, soi` of day d in the MDS data set. The MDS is imbalanced with 4.2% positive (yes or present of cholera cases) observations. Classification methods on imbalanced data set often give very high specificity (or true negative rate) and low sensitivity (or true positive rate) [15]. Common approaches to deal with imbalanced data sets are resampling the train data set, using class-weighted classification methods and collect additional data if possible [15,16]. In this paper, we use a new approach that combines additional solar terms data and resampling methods.

To assess performance of classifiers, researchers use common measures including precision, recall, F1, area under the Receiver Operating Characteristic curve (refers as AUC) and Cohen's Kappa. Jeni et al. [17] found effects of imbalanced data sets to the measures. They defined skewness of an imbalanced data set as

$$skew = \frac{negative\ observations}{positive\ observations} \tag{1}$$

and found that F1 and area under the Precision-Recall curve (APR) drop when $skew > 1$ and becomes larger; Kappa drops when $skew \neq 1$ and becomes more different. The AUC is virtually not attenuated while *skew* changes. Therefore in this research, we choose the AUC, a non-sensitive measure to skewness, for assessment of classification models. The skew of MDS data set is 22.7.

Performance of cholera outbreaks prediction models is affected by three factors: resampling method, classification method and additional seasonal data. To compare the effects of factors to performance of models, we design a factorial experiment [18] with the above three factors. Levels of each factor are in Table 4. In the table, resampling factors are named after corresponding methods: `none, up, down, smote, rose` [16]. Classification methods are general linear model (`glm`), k-nearest neighbours (`knn`), C5.0 (`C5.0`) and random forests (`rf`). Seasonal information levels are `none, week, month, solar` meaning use MDS only, MDS and `week`, MDS and `month` and MDS and `solar`, respectively. We do not combine different types of seasonal information because they are all derived from `date` variable. The factors combination is 80. For each combination we build a prediction model and assess its performance using AUC measure.

Table 4. Factors and their levels.

Factor	**Levels**
Resampling	none, up, down, smote, rose
Method	glm, knn, C5.0, rf
Seasonal	none, week, month, solar

To build classification models, we first rearrange MDS with seasonal factors listed in Table 4 to obtain seasonMDS data sets. Each seasonMDS is randomly divided into training and testing data sets. The training set has 70% observations of the seasonMDS and the testing is the rest. The training data set is resampled following the above resampling methods. We then apply the above classification methods to build models on the training data set and test their performance on the test data set. This procedure is repeated 100 times for cross-validation. We collect the models performance measures to a PERF data set to compare models performance. The measures include AUC, sensitivity and specificity. We run all experiments in an R environment [19] and use **caret** package [20] for prediction modeling. In the next section, we analyze the result statistically.

4.2 Statistical Analysis of Models Performance

The purpose of our statistical analysis in this section is twofold. Firstly, we find which factors among `seasonal`, `sampling` and `method` influence AUC measure taking into account their possible interactions. Secondly, we compare mean of AUC of groups forming by combination of the three factors to find the best. The best groups must have high AUC, balanced sensitivity and specificity.

We use the analysis of variance (ANOVA) method [21] to analyze effects of the factors. To perform the ANOVA, we build a linear regression model named LM1 on the PERF data set. In the LM1 model, the three factors `seasonal`, `sampling` and `method` are independent variables. The AUC measure is a dependent variable. The LM1 model is

```
LM1 <- lm(auc ~ seasonal*sampling*method, data=PERF)
```

as writing in R syntax. The model takes interaction effects of the three factors into account. ANOVA test result of LM1 is in Table 5.

Table 5. ANOVA test for effects of factors and their interactions to mean of AUC.

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
seasonal	3	16.6507	5.5502	4692.1734	0.0E+00
sampling	4	4.8779	1.2195	1030.9513	0.0E+00
method	3	19.8978	6.6326	5607.1960	0.0E+00
seasonal:sampling	12	0.8860	0.0738	62.4214	1.5E-145
seasonal:method	9	5.6684	0.6298	532.4556	0.0E+00
sampling:method	12	11.1310	0.9276	784.1803	0.0E+00
seasonal:sampling:method	36	1.6713	0.0464	39.2479	2.4E-250
Residuals	7920	9.3684	0.0012		

In Table 5, all *p*-values corresponding to the factors (or variation sources) in the last column (**Pr(>F)**) are much smaller than 0.05 indicating that all the factors, their pairwise interactions (`seasonal:sampling`, `seasonal:method`, `sampling:method`) and interaction of all three (`seasonal:sampling:method`) have effects to AUC measure and the effects are all statistical significant. By comparing mean squares in the fourth column (**Mean Sq**), we see that `method` has highest value of 6.63, then `seasonal` (5.55), `sampling` (1.22), `sampling:method` (0.93). The interaction `seasonal:sampling:method` has lowest value of 0.05. It means that the `method` and `seasonal` variables have highest effects and the interaction `seasonal:sampling:method` has lowest effect to mean of AUC.

We compare mean of AUC taking into account the interactions by calculating adjusted mean of AUC of the groups and producing interaction plots. We use **phia** [22] and **ggplot2** [23] packages to perform this task. Fig. 1 shows effects of pairwise interactions to mean of AUC. We refer plots in the figure by their coordinate $(row, column)$ with reference to the top left plot. Each plot among (1,1), (2,2) and (3,3) has one curve. These curves describe main effects (no interaction) of `seasonal`, `sampling` and `method` to mean of AUC. As seen in the plots, `solar` of seasonal, `smote` of sampling and `rf` of method are the most influence to mean of AUC. Using another linear model with no interaction

```
LM2 <- lm(auc ~ seasonal+sampling+method, data=PERF)
```

we quantify main effects of the factors, described Table 6. In the table all p-values in the last column is much smaller than 0.05. It is statistically significant that `solar`, `week` and `month` of `seasonal` factors increase mean of AUC by 12.66%, 8.25% and 7.88%, respectively. These values are in the second column of Table 6. Main effects of other factors are interpreted in the same manner.

Table 6. A linear model describing main effects of the factors to mean of AUC.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.6310	0.0022	283.7711	0.0E+00
seasonalmonth	0.0788	0.0019	41.5728	0.0E+00
seasonalsolar	0.1266	0.0019	66.7829	0.0E+00
seasonalweek	0.0825	0.0019	43.4977	0.0E+00
samplingdown	0.0497	0.0021	23.4324	1.7E-117
samplingrose	0.0614	0.0021	28.9797	1.1E-175
samplingsmote	0.0715	0.0021	33.7317	2.6E-233
samplingup	0.0518	0.0021	24.4478	2.3E-127
methodC5.0	-0.0098	0.0019	-5.1705	2.4E-07
methodknn	-0.0875	0.0019	-46.1373	0.0E+00
methodrf	0.0521	0.0019	27.4557	1.2E-158

The off-diagonal plots in Fig. 1 describe pairwise interactions of the factors. Plot (1,3) shows interactions of `method` and `seasonal`. We notice that combination of `rf` (method) and `solar` (seasonal) delivers the highest mean of AUC. If we change method to `knn`, mean of AUC will change different amounts depending on seasonal data. When seasonal data is `none` or `solar`, the mean of AUC will decrease the same amount of approximate 9%. However when it is `week` or `month`, the mean of AUC will decrease an amount of approximate 20%. These effects are shown in the plot (1,3) by almost parallel segments of `solar` and `none` and nonparallel segments of `solar` and `month/week` with reference to `knn` and `rf` points on the x-axis. Interpretation of other interactions is similar.

To find groups those have highest means of AUC we calculate adjusted mean of AUC. The adjusted means of the best groups are in Table 7. Results in the table show that using `solar` and `rf` is recommended for high AUC. Sampling

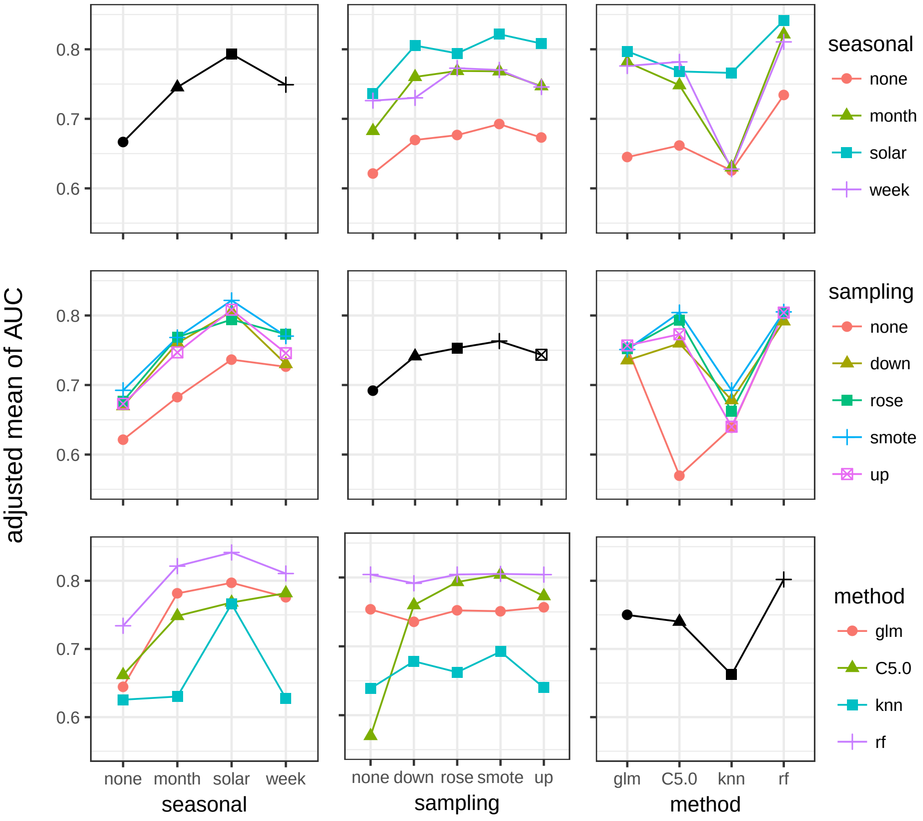

Figure 1. Influence of pairwise interactions to mean of AUC (color online).

method is a choice among `up`, `smote` or `rose`. The `rose` gives balanced of sensitivity (0.746) and specificity (0.778), and therefore is chosen. The combination of (`week`, `smote`, `C5.0`) is slightly worse than (`solar`, `rose`, `rf`). However it has unbalanced sensitivity (0.525) and specificity (0.898) and is not chosen.

4.3 Importance of Variables

We have chosen (`solar`, `rose`, `rf`) as the best combination. The importance of variables in this combination is simply extracted from the random forests output in the factorial experiment (Sec. 4.1). To compare importance of variables statistically, we build yet another linear model for an ANOVA test. The model's outcome is the importance; and the predictor is a factor having `tavg`, `havg`, `precip`, `sun`, `wind`, `soi`, `ec`, `solarterm` as levels. Test result is statistical significant and shows that the top important variables are `ec`, `soi`, `tavg` and `havg`. Among the solar terms, `start of summer` and `frost descent` are the most important. Table 8 lists the importance of these variables.

Table 7. List of five groups with highest mean of AUC.

seasonal	sampling	method	auc	sensitivity	specificity
solar	up	rf	0.855	0.143	0.989
solar	smote	rf	0.850	0.601	0.861
solar	rose	rf	0.840	0.746	0.778
solar	smote	C5.0	0.838	0.564	0.867
week	smote	C5.0	0.837	0.525	0.898

Table 8. Importance of variables and solar terms in percentage.

	mean	95% confidence interval
ec	97.51	[96.67, 98.35]
soi	90.29	[89.45, 91.13]
tavg	75.54	[74.70, 76.38]
havg	74.54	[73.70, 75.38]
frost descent	15.39	[14.55, 16.23]
start of summer	13.47	[12.64, 14.31]

5 Conclusions

We have developed a new solution to enhance performance of cholera outbreaks prediction in Hanoi. Our solution is a combination of solar terms data, ROSE resampling of training data set and random forests method. The solution delivers high AUC (0.84), balanced sensitivity and specificity.

Without taking interactions of the factors into account, the solar terms data helps increasing mean of AUC by an amount of 12.66%. The most important variables in the solution are `ec`, `soi`, `tavg` and `havg`. Among the solar terms, `frost descent` and `start of summer` are the most important.

To our best knowledge, this research is the first that uses solar terms in a cholera outbreaks prediction. Our experiment results show that solar terms are worth considering as predictors for cholera modeling in Hanoi.

References

1. Jutla, A., Whitcombe, Hasan, N., Haley, B., Akanda, A., Huq, A., Alam, M., Sack, R., Colwell, R.: Environmental factors influencing epidemic cholera. Am. J. Trop. Med. Hyg. 89 (3), 597–607 (2013)
2. Martinez, P.P., Reiner, R.C.Jr., Cash, B.A., Rodó. X. et al.: Cholera forecast for Dhaka, Bangladesh, with the 2015-2016 El Niño: Lessons learned. PLoS ONE 12 (3), e0172355 (2017)
3. Ali, M., Kim, D.R., Yunus, M., Emch, M.: Time series analysis of cholera in Matlab, Bangladesh, during 1988-2001. Journal of Health, Population and Nutrition 31 (1), 11–19 (2013)
4. Reiner, R.C., King, A.A., Emch, M., Yunus, M., Faruque, A.S.G, Pascual, M.: Highly localized sensitivity to climate forcing drives endemic cholera in a megacity. Proc. Natl. Acad. Sci. USA 109, 2033–2036 (2012)

5. Emch, M., Feldacker, C., Yunus, M. et al.: Local environmental predictors of cholera in Bangladesh and Vietnam. The American Journal of Tropical Medicine and Hygiene 78 (5), 823–832 (2008)
6. Xu, M., Cao, C.X., Wang, D.C., Kan, B., Jia, H.C., Xu, Y.F., Li, X.W.: District prediction of cholera risk in China based on environmental factors. Chinese Science Bulletin 58 (23), 2798–2804 (2013)
7. Xu, M., Cao, C.X., Wang, D.C., Kan, B.: Identifying environmental risk factors of cholera in a coastal area with geospatial technologies. Int. J. Environ. Res. Public Health 12, 354–370 (2015)
8. Kelly-Hope, L.A., Alonso, W.J., Thiem, V.D. et al.: Temporal trends and climatic factors associated with bacterial enteric diseases in Vietnam 1991-2001. Environmental health perspectives 116 (1), 7–12 (2008)
9. Le, T.N.A, Ngo, T.O., Lai, T.H.T, Le, H.Q., Nguyen, H.C., Ha, Q.T.: An experimental study on cholera modeling in Hanoi. Proceedings of Asian XI Conference on Intelligent Information and Database Systems 2016, 230–240 (2016)
10. Nguyen, H.C., Le, T.N.A.: Using local weather and geographical information to predict cholera outbreaks in Hanoi, Vietnam. Advances in Intelligent Systems and Computing 453, Springer, 195–212 (2016)
11. Daily Southern oscillation index data set of the Queensland, Australia, `https://www.longpaddock.qld.gov.au/seasonalclimateoutlook/southernoscillationindex/soidatafiles/DailySOI1887-1989Base.txt`
12. Qian, C., Yan, Z., Fu, C.: Climatic changes in the twenty-four solar terms during 1960-2008. Chinese Science Bulletin. Atmospheric Science 57 (2-3), 276–286 (2012)
13. Hong Kong Observatory's solar term introduction, `http://www.weather.gov.hk/gts/time/24solarterms.htm`
14. Hong Kong observatory's climatology for the 24 solar terms: `http://www.weather.gov.hk/cis/statistic/ext_st_vernal_equinox_e.htm?element=0&operation=Submit`
15. Kuhn, M., Johnson, K.: Applied predictive modeling. Springer (2013)
16. He, H., Mai Y.: Imbalanced learning: foundations, algorithms and applications. John Wiley & Sons, Inc. (2013)
17. Jeni, L.A., Cohn, J.F., Torre, F.D.L.: Facing imbalanced data recommendations for the use of performance metrics. In: ACII '13 Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (2013)
18. Montgomery, D.C.: Design and analysis of experiments, 8th ed. John Wiley & Sons, Inc. (2013)
19. Micheaux, P., Drouilhet, R., Liquet, B.: The R software: Fundamentals of programming and statistical analysis. Springer (2013)
20. **caret** package, `https://cran.r-project.org/web/packages/caret/index.html`
21. Faraway, J.: Linear models with R, 2nd ed. CRC Press (2015)
22. **phia** package, `https://cran.r-project.org/web/packages/phia/index.html`
23. **ggplot2** package, `https://cran.r-project.org/web/packages/ggplot2/index.html`