THE APPLICATION OF TRINDEX TO PREDICT HARMFUL ALGAL BLOOMS IN LAKE TORMENT (NOVA SCOTIA, CANADA)

Kateryna Hushchyna, Bao Quoc Vo, Tri Nguyen-Quang

Biofluids and Biosystems Modeling Lab (BBML)
Faculty of Agriculture, Dalhousie University,
39 Cox Road, B2N 5E3, Truro-Bible Hill, Nova Scotia, Canada
tri.nguyen-quang@dal.ca

Abstract. This paper introduces the Threshold Index (hereafter called TRINDEX) for the Harmful Algal Bloom (HAB) prediction in lake Torment (Nova Scotia province, Canada). TRINDEX was suggested via the logarithmic transformation and then the thresholds of bloom pattern were established by the discrimination test named the Receiver Operating Characteristic curve (ROC). The cohort studies will be also presented to show how accurate the bloom prediction when using TRINDEX in comparing to the real observations.

Key words: Harmful Algal Bloom (HAB), Threshold Index (TRINDEX), Cohort studies, Receiver Operating Characteristic (ROC) curve.

1. Introduction

Harmful algal blooms (HAB) have become more prominent and frequent event in Canadian fresh waters [1]. HAB can cause damaging physical (by an excess accumulation and following hypoxia in the water) or chemical (producing toxins) effects on various habitats living in the waterbody and surrounding areas. It is also well-known fact that HAB can affect negatively on the environment, human and animal lives through their food chain or via direct contacts with the water containing HAB. Eutrophication has been considered for a long time the main cause of HAB with the intensive growth of cyanobacteria in freshwater bodies. Therefore, the different indices mostly based on nutrients were largely developed to explain how eutrophication can lead to the HAB patterns. These indices included single or combined indexes, Single indexes, such as total phosphorus (TP), chlorophyll-a (Chl-a) or phycocyanin

(PC) concentration, Secchi disk measurements [2–7], are still very common in the practice of water management. However, they can poorly predict the phytoplankton growth in real conditions of the waterbody. Some combined indexes of HAB prediction are very specific to the local environmental conditions. Moreover, rare research works were focused on the determination of the thresholds for HAB happening, except for [8] which predicted the risk of cyanobacterial dominance, and from there the threshold could be identified.

Based on the Trophic Index proposed for marine water to characterize the trophic state of coastal marine water (Emilia-Romagna coast) of the Adriatic Sea by Vollenweider and co-workers [9], our index is suggested with modification and adaptation for freshwater media. Our suggestion, called TRINDEX, will show the estimation effectiveness for the variation of cyanobacterial dominance in phytoplankton communities. Moreover, TRINDEX can assist in determining the threshold for cyanobacterial bloom onset due to the fact that the typical pigment phycocyanin characterizing for development of cyanobacteria is incorporated.

In this paper, we introduce in the first step the new indicator used to estimate the potential bloom occurrence in freshwater bodies and then to determine the threshold for bloom happening. Next, we will use the cohort studies to evaluate the accuracy of TRINDEX in the bloom prediction. Samples and real bloom observations were obtained from lake Torment (Nova Scotia). Our utmost goal in this paper is to introduce a new and simple way to help predicting HAB patterns via the determination of bloom thresholds

and then to continue performing TRINDEX indices for a long term practical application in the water management and monitoring program of the areas affected by HAB.

2. Methodology

2.1. Study sites

Our research was conducted at the lake Torment in Nova Scotia province as above mentioned. The location and geographical information are displayed in Fig. 1 below.

Lake Torment (LT) locates in the Kings County (Nova Scotia province, Canada). The lake depth varies from 1.0 to 6.5 meters and around the lake there is the residential area including 250 cottages and leisure activities. In additions, the lake is surrounded by a forest and one Christmas tree farm nearby. According to Kings Report [10], LT is dystrophic with brown water, low pH, low carbonate level and especially high organic content. HAB are identified annually, even ramps up with high density lately. The research from Nguyen-Quang and collaborators [11] shows that cyanobacteria strains in LT ecosystem contain mostly *Dolichspermum flos-aquae* and they came along randomly in the time frame from June to November. *Microcystis* sp. also appeared in the fall season under the mixed bloom forms with *Dolichspermum flos-aquae*.
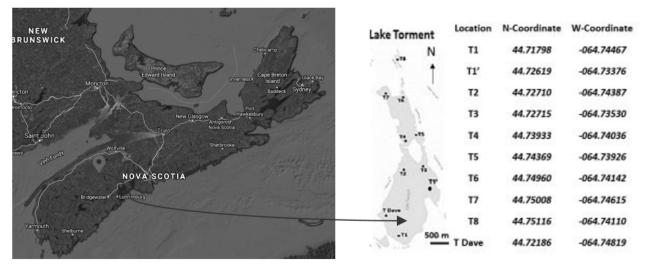

Lake Torment	Location	N-Coordinate	W-Coordinate
	T1	44.71798	-064.74467
	T1'	44.72619	-064.73376
	T2	44.72710	-064.74387
	T3	44.72715	-064.73530
	T4	44.73933	-064.74036
	T5	44.74369	-064.73926
	T6	44.74960	-064.74142
	T7	44.75008	-064.74615
	T8	44.75116	-064.74110
	T Dave	44.72186	-064.74819

Fig. 1. The study site (lake Torment) and sampling locations

2.2. Field Sampling Process and Lab Analysis

Water sampling was done with the bi-weekly or monthly frequency starting from May through to November during the 2016–2018 period with 170 samples in total (statistical analysis will be shown in Table 2 below). Water was taken at 0.5 m from the surface and at 1m from the bottom at 10 different locations. DO was measured by YSI probe (Professional Plus, Hoskin scientific LTD, USA) directly at the field. Other parameters (phosphate (PO_4), nitrate (NO_3), Chl-a and *PC*) were analysed at our Lab, Dalhousie University.

Water samples were filtered through the GF/A Whatman filters which were used for the determination of pigment concentrations. Next, filtered water was used for nutrients analyses. The used filters were extracted in 90 % acetone (for Chl-a) and in the phosphate buffer saline (for PC). For better extraction, a sonication (50 % amplitude for 30 seconds) and a two-step centrifugation (at room temperature with 3500 *g* for 10 minutes in the beginning followed by second centrifugation at 4 ℃, 13000 *g* for 1.5 hours) were held. Pigment concentrations from Chl-a and PC were measured in Turner 10AU Fluorometer (Turner Designs, USA) based on the calibration standard curve for both. Dissolved fraction of phosphate and nitrate were measured by a photometer using a tablet reagent system [12].

2.3. Mathematical formulation

Our goal for formulation in this paper is just to present a summary of all main points related to the TRINDEX conception and calculation. We do not show here details of formula development. Readers are referred to [13] for all details of TRINDEX development.

To deal with the non-normal distribution of most of the environmental data, the logarithmic transformation is a mathematically appropriate way to 'transform' random data into a normal distribution form as below.

$$TRINDEX = \frac{k}{n} \times \sum_1^n \left[\frac{\text{Log}M_i - \text{Log}L_i}{\text{Log}U_i - \text{Log}L_i} \right] \qquad (1)$$

where, M_i – measured parameter i; L_i – lower limit (concentration) of the considered parameter i;

U_i – upper limit (concentration) of the considered parameter i; k – factor standing for the maximum value of considered range (0.10), so $k=10$ by default. n – total number of parameters M_i we expect to consider.

$$TRINDEX_1 = \frac{10}{4} \times \left[\frac{LogPC - LogL_{PC}}{LogU_{PC} - LogL_{PC}} + \frac{LogD\%O - LogL_{D\%O}}{LogU_{D\%O} - LogL_{D\%O}} + \frac{LogPO_4 - LogL_{PO_4}}{LogU_{PO_4} - LogL_{PO_4}} + \frac{LogNO_3 - LogL_{NO_3}}{LogU_{NO_3} - LogL_{NO_3}} \right] \quad (2)$$

$$TRINDEX_2 = \frac{10}{5} \times \left[\frac{LogChla - LogL_{Chla}}{LogU_{Chla} - LogL_{Chla}} + \frac{LogPC - LogL_{PC}}{LogU_{PC} - LogL_{PC}} + \frac{LogD\%O - LogL_{D\%O}}{LogU_{D\%O} - LogL_{D\%O}} + \frac{LogPO_4 - LogL_{PO_4}}{LogU_{PO_4} - LogL_{PO_4}} \right.$$
$$\left. + \frac{LogNO_3 - LogL_{NO_3}}{LogU_{NO_3} - LogL_{NO_3}} \right] \quad (3)$$

$$TRINDEX1 = 2.5 \times \left[\frac{LogPC + 4.4}{4.7} + \frac{LogD\%O + 2}{4.0} + \frac{LogPO_4 + 2}{2.2} + \frac{LogNO_3 + 2}{2.2} \right] \quad (2')$$

$$TRINDEX2 = 2 \times \left[\frac{LogPC + 4.4}{4.7} + \frac{LogChl-a + 4.3}{4.5} + \frac{LogD\%O + 2}{4.0} + \frac{LogPO_4 + 2}{2.2} + \frac{LogNO_3 + 2}{2.2} \right] \quad (3')$$

The quantity (logUi – logLi) is defined by the difference of upper and lower limits. When these limits are determined, all values out of this range should be excluded. Therefore, to have an appropriate range to cover different trophic conditions, we used limits of detection (LOD) as the lower limit and the maximum value obtained in the measurements of the considered variable for the upper limit.

Our experimental data related to HAB for lake Torment is not normally distributed. Using the log transformation as mentioned above is to convert them into the normal distribution and TRINDEX can be then processed. The statistical software R combined with SPSS are used to carry out all steps including the next ones.

2.4. Determination of the bloom thresholds

As the pigment Chl-a can be found in all of algal species including the green and blue-green algae, as well as in other microplants, the pigment PC can presumably be a better alternative parameter to reflect the cyanobacterial presence in all phytoplankton community. Therefore, we suggest that when PC concentration in the lake water is greater than 0.03 ± 0.002 (mg/L), it can be considered as onset of bloom (PC criteria based on [5]), equivalent to the cell count 20,000 cells/mL of cyanobacteria, and PC will be inserted into the TRINDEX to evaluate the bloom patterns of cyanobacteria.

The onset of bloom could be a visible bloom or scum situation, but this might not be stable. The surface bloom at the onset status can be observed unstably appeared and disappeared in a short period of time (critical phase) while the supercritical phase of blooms can show a more stable situation where blooms or scums can last visibly for longer periods (many hours or many days). The onset status can lead to a 'stable bloom'

Two scenarios of TRINDEX 1 & 2 are referred to the research of Hushchyna and Nguyen-Quang [13]. Precisely, TRINDEX1 consists of PC, D %O, PO$_4$ and NO$_3$, while TRINDEX2 includes all parameters present in TRINDEX1 plus the pigment Chl-a.

if ambient conditions allow them to develop, or alternatively completely vanish, also due to the ambient conditions.

As in clinical practice [14], a 'yes or no' decision is usually required for 'diseased or non-diseased' situation, herein two states for the bloom: 'yes – bloom occurrence and no – no bloom' are also defined.

Four possible outcomes can result for each trial: true positive, true negative, false positive and false negative. At this point, the cut-off area will be introduced as the area which measures the discrimination, i.e. the ability of the TRINDEX test to correctly classify those with, or without the 'disease', as a binary variable. That is equivalent to bloom occurrence (yes) or no bloom (no) respectively. The ROC (Receiver Operating Characteristic) analysis is a binary discriminator test which assesses the predictive power of a binary classification system to evaluate a model in a decision-making process and it helps to identify the threshold. On ROC curve, however, there are numerous observed cut-off points, thus pointing out the best cut-off point will strengthen the reliability for the model.

The point with coordinates (0,1) is where the sensitivity and specificity's values are equal to 1.0. The best cut-off point (or optimal threshold point) can be determined by the minimum distance from the point (0,1) to the ROC curve [15] (see Fig. 2). Mathematically, the distant between this point (0,1) to any points on ROC curve is calculated as below.

$$d = \sqrt{(1 - S_n)^2 + (1 - S_p)^2}, \quad (4)$$

where, S_n denotes the sensitivity and S_p is a specificity.

Based on this formula (4), we can infer the determination of the best cut-off point as below

$$\text{cutoff point} = d_{min} = \min\left(\sqrt{(1 - S_n)^2 + \left(1 - S_p\right)^2} \right) \quad (5)$$

Once the ROC curve is drawn, the area under the curve (AUC) can be used to evaluate the overall

performance of the discrimination test (Table 1). AUC may take values ranging from 0.5 (no discrimination) to 1 (perfect discrimination).
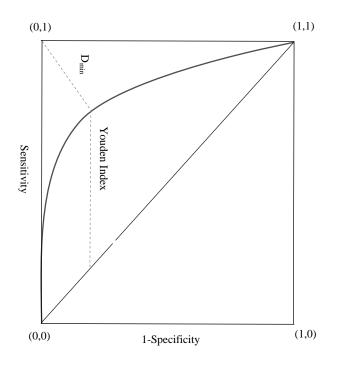

Fig. 2. Finding best cut-off point (optimal threshold) for ROC curve

Table 1

The AUC criteria to evaluate the accuracy of diagnostic test

AUC Value	0.9–1.0	0.8–0.9	0.7–0.8	0.6–0.7	0.5–0.6
Evaluation	Excellent (A)	Good (B)	Fair (C)	Poor (D)	Fail (F)

Another approach we can use to estimate the effectiveness of our test is the Youden index J [16] which can be defined as follows.

J = max {sensitivity c + specificity c - 1}   (6)

where c ranges over all possible criterion values.

The Youden index J, ranging between 0 and 1, is commonly used to measure overall diagnostic effectiveness [17]. When J values close to 1, they indicate the effectiveness is relatively good, and close to 0 showing the limited effectiveness.

2.5. Cohort studies applied for TRINDEX in Torment

Cohort studies [18] are a type of medical research used to investigate the causes of disease and to establish links between risk factors and health outcomes. The word cohort means a group of people and has the roots from military strategies. These types of studies look at

groups of people, which can be prospective (forward-looking) or retrospective (backward-looking).

The prospective studies are planned and carried out over a future time period while the retrospective cohort studies look at data that already exist and try to identify risk factors for particular conditions.

Inspired from this, we assume the risk factors for our problem are represented by our TRINDEX1 or TRINDEX2; and the health outcomes are the bloom situations: no bloom (no disease) or blooming (disease). For more details of cohort studies, refer to [18].

A cohort study follows up two or more groups from exposure to outcome. In its simplest form, a cohort study compares the experience of a group exposed to some factor with another group not exposed to the factor. If the former group has a higher or lower frequency of an outcome than the unexposed, then an association between exposure and outcome is evident [18]. TRINDEX1 is chosen for the analysis of our cohort studies in this research to predict the probability of blooms in lake Torment. We use R with the package 'epiR' to perform our studies.

3. Results and Discussions

3.1. Descriptive statistics and normal distribution test for TRINDEX

A general look on some basic statistical features of TRINDEX1 and TRINDEX2 is shown as below.

Table 2

Descriptive statistics

Indices	TRINDEX1	TRINDEX2
Sample number	170	170
Mean	3.93	4.09
Median	3.92	3.96
Standard Deviation	1.0198	1.0705
Skewness	0.956	1.464

In Table 2 for the case of TRINDEX1, mean and median values are nearly equal to each other (3.93 and 3.92), the skewness is in the range from -1.0 to +1.0 (0.956). Therefore, this variable is likely to be normal distributed. However, in the case of TRINDEX2, though mean and median are nearly equal to each other (4.09 and 3.96), the skewness is not in the range from -1.0 to +1.0 (1.464). Hence this variable seems unlikely to be normal distributed. The histograms of distribution for

both cases are displayed in Fig. 3. For both cases TRINDEX1 and 2, their histograms show the bell-shaped curve. However, to confirm if these two datasets of TRINDEX1 and TRINDEX2 are normally distributed, the *Q-Q Tests* for both cases TRINDEX1 and 2 are conducted.

According to the Q-Q plot tests, if the datasets are closely distributed along a line, it can be concluded that they are with a normal distribution. In contrast, if the data points are scattered far from the line, it is

obvious that the dataset are not under the normal distribution form. The Q-Q plot for TRINDEX 1 and 2 are displayed in Fig.4 which show the majority of data points of both TRINDEX1 and TRINDEX2 disperse very closely along the diagonal line in each case. Therefore, it can be concluded that the logarithm transformation has converted our field data from lake Torment (which are not normally distributed) into two datasets under the quasi-normal distribution form of TRINDEX 1 and TRINDEX2.
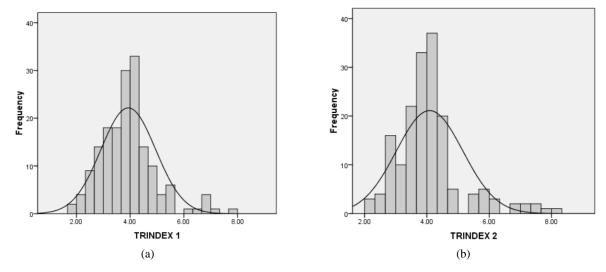

(a)   (b)

Fig. 3. Histogram of distribution for (a) TRINDEX 1 and (b) TRINDEX 2 in lake Torment

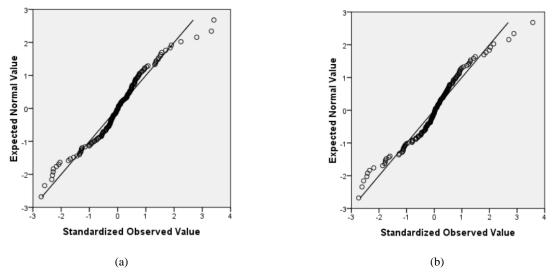(a)   (b)

Fig. 4. Normal Q-Q Plot of TRINDEX1 (a) and TRINDEX2 (b)

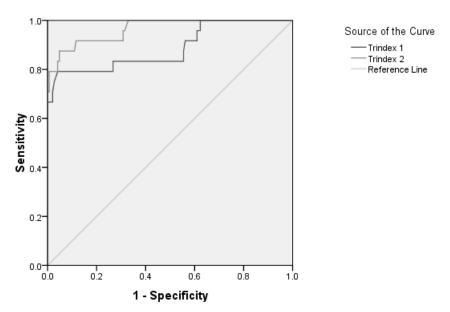### 3.2. ROC Curves and threshold values to predict bloom

With the assistance of SPSS software, our ROC curves for TRINDEX1 and TRINDEX2 are built as shown in Fig. 5. The yellow line represents the reference below it the curve is not valid. The blue

and green lines (standing for TRINDEX1 and TRINDEX2, respectively) illustrate the trade-off between sensitivity (the blooming patterns really occurring, or true positive) and (1 – specificity), i.e. non-blooming cases are identified as positive, or false positive.

Based on formula 5, TRINDEX1 will have the cut-off point value of 4.815 (d_{min}= 0.212) and TRINDEX2 will have the cut-off point value of 4.605 (d_{min} = 0.134). In both cases of TRINDEX 1 and TRINDEX 2, at those cut-off points, the sensitivity is lower than the specificity (79.2 % vs. 95.9 % for TRINDEX1, and 87.5 % vs. 95.2 % for TRINDEX 2) which lead to the conclusion that these two TRINDEXES are likely to be appropriate for diagnosing the "disease-bloom" in the water reservoirs.

Besides, we can also determine Youden Index based on formula 6. Youden index for TRINDEX1 case is 0.73, while it equals to 0.78 in the case of TRINDEX2. These Youden values are close to 1.0 in both cases showing the good effectiveness of these 2 TRINDEXES.

As the result in Table 3, both TRINDEX1 and 2 have low p value (<0.0001). According to Table 1, the ROC curve for TRINDEX2 shows an excellent AUC (0.956 or 95.6 %), whereas the curve for TRINDEX1 has also a good AUC (0.887 or 88.7 %) (95 %CI: 0.830 to 0.930; $p < 0.0001$).

ROC Curve

Fig. 5. ROC Curves of TRINDEX1 and TRINDEX2

Table 3

Area under the curve – AUC

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95 % Confidence Interval	
				Lower Bound	Upper Bound
TRINDEX 1	0.888	0.046	0.000	0.799	0.977
TRINDEX 2	0.965	0.019	0.000	0.928	1.002

* The test result variable(s): TRINDEX1, TRINDEX2 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

[a] Under the nonparametric assumption

[b] Null hypothesis: true area = 0.5

3.3. Cohort studies

Cohort studies are processed for lake Torment as previously mentioned. The matrix to be considered in the study is defined by the number of bloom cases predicted by TRINDEX1 based on PC criteria, versus the observed cases of blooms during the period 2016-2018.

According to our data analysis, cases of real observations for blooms and no-blooms are shown in following Table 4.

Table 4

Cases of blooms in lake Torment (2016–2018) by field observations versus predicted cases by TRINDEX1

Torment lake patterns	Real cases visually observed on the field	Cases predicted by TRINDEX 1
Bloom	30	24
No bloom	140	146

Table 5 displays the results from our cohort studies. In Table 5, the *Outcome+* (bloom) and *Outcome–* (no bloom) parts show the summary of entrance data before going to calculation. The *Incident risk (Inc. risk)* is defined as the risk possibility (very high bloom

possibility). The *Expose+* is case of real observations while the *Expose–* is case predicted by using TRINDEX1.

Table 5

Results from Cohort studies by R

	Outcome+	Outcome-	Total	Inc Risk *	Odds
Exposed+	30	140	170	**17.6**	0.214
Exposed-	24	146	170	**14.1**	0.164
Total	54	286	340	15.9	0.189
Point estimates and 95 % CIs:					
Inc risk ratio Odds ratio Attrib risk*	1.25 (0.76, 2.05) 1.30 (0.73, 2.34) **3.53 (-4.23, 11.29)**				

Test that OR = 1: χ^2 (1) = 0.793 Pr> χ^2 = 0.37
Wald confidence limits
CI: confidence interval
* Outcomes per 100 population units (%)

For lake Torment, the incident risk if we use TRINDEX1 (high-risk probability of blooms) is 14.1 % while the real observations give 17.6 % (bold numbers).

The difference between prediction by TRINDEX1 and observations is 3.53 % (defined as Attribute Risk) with the 95 %CI is (-4.23 % to 11.29 %).

In other words, in using TRINDEX1 for the bloom prediction, we could get the results with a probability of risk not far from the real observations (14.1 % versus 17.6 %) and this difference varies in average from -4.23 % to 11.29 % (95 %CI).

As the range of 95 % CI contains the zero value and going through from – to + ranges, this difference can be not statistically significant, i.e the bloom risk difference between them is just random. In other words, using the prediction by TRINDEX1 is quite close to the real observations. That means the TRINDEX1 prediction for lake Torment is quite accurate.

Conclusions

In this paper, the prediction of Harmful Algal Blooms (HAB) via TRINDEX uses a simple logarithmic transformation combined with the technique of classification/discrimination and in situ sampling. The TRINDEX is to introduce a new assessment tool for Cyanobacterial Bloom prediction. The state-of art test of discrimination (ROC Curve) was used to confirm the threshold values. From this, the forecasting for

bloom incidence can be processed. It can be said that at the first time, the cohort studies were introduced into the HAB research. They were used to establish links between risk factors (parameters involved) and bloom outcomes to estimate the accuracy of TRINDEX. Further research related to TRINDEX development and cohort studies for other waterbodies would be envisaged.

All our efforts rely on a close relationship between observations and simple model leading to developing a forecasting capability. Building a based observation-modelling framework would not only increase our knowledge of fundamental ecological relationships for constructing the better models but also establish the structure of an early warning system, the goal of which is the protection of livelihoods and water resources.

Acknowledgements

References

[1] Pick FR.: Canadian J. Fisheries and Aquatic Sci., 2016, 73, 1149.

[2] Ahn CY., Joung SH., Yoon SK., Oh HM.: J. Microbiology, 2007, 45(2), 98.

[3] Bartram J., Burch M., Falconer R., Jones G., Kuiper-Goodman T.: Situation assessment, planning and management [in:] Chorus I., Bartram J. (Eds), Toxic Cyanobacteria in Water: A guide to their public health consequences, monitoring and management. E & FN Spon, London, 1999, 183–210

[4] Brylinsky M.: Lake Utopia Water Quality Assessment. Final report, Dept. Environment of New Brunswick, Canada 2009.

[5] Brient L., Lengronne M., Bertrand E., Rolland D., Sipel A., Steinmann D., Baudin I., Legeas M., Le Rouzic B., Bormans M.: Journal of Environmental Monitoring, 2008, 10, 248.

[6] Chorus I.: Current approaches to cyanotoxin risk assessment, risk management and regulations in

different countries, Federal Environmental Agency, Germany 2012.

[7] Novotny V., Olem H.: Water Quality: Prevention, Identification, and Management of Diffuse Pollution, Van Nostrand Reinhold, New York 1994.

[8] Downing JA., Watson SB., McCauley E.: Canadian J. Fisheries and Aquatic Sci., 2001, 58 (10), 1905.

[9] Vollenweider RA., Giovanardi F., Montanari G., Rinaldi A.: Environmetrics, 1998, 9, 329.

[10] Marty J., Reardon M.: King county lake monitoring program 2015 season, Report, Municipality of the County of Kings, Canada 2016.

[11] Nguyen-Quang T., McLellan K., Hushchyna K., Murdock A. : Harmful Algal Bloom (HAB) Monitoring for Lake Torment and Armstrong Lake. A systematic investigation in 2016-2017 versus 2014 Kings County water quality results. Technical report, Canada 2017.

[12] YSI 9300 and 9500 direct-read photometer: User manual. YSI Incorporated, USA 2010.

[13] Hushchyna K. and Nguyen-Quang T.: Nature Environment and Pollution Technology, 2020. In press.

[14] Carter JV., Pan J., Rai SN., Galandiuk S.: Surgery, 2016, 59(6), 1638.

[15] Rajeeve K. and Abhaya I.: Indian Pediatrics, 2011, 48, 277.

[16] Youden WJ.: An index for rating diagnostic test. Cancer, 1950, 3: 32–35

[17] Schisterman EF., Perkins NJ., Liu A., Bondell H.: Epidemiology, 2005, 16 (1), 73.

[18] Grimes DA. and Schulz KF.: The Lancet 2002, 359 (9303), 341.