Water quality affects diatom diversity in Finnish rivers

Makiko Oda[*], Matti Leppänen[†], Anna Karjalainen[†], Satu-Maaria Karjalainen[†], Hiroshi Suito[**] and Timo Huttula[†]

[*] National Defense Medical College
3-2 Namiki, Tokorozawa, Saitama, 359-8513, Japan
oda@ndmc.ac.jp
[†]Finnish Environment Institute
Matti.T.Leppanen@ymparisto.fi, Anna.Karjalainen@ymparisto.fi,
Satu.Maaria.Karjalainen@ymparisto.fi, timo.huttula@ymparisto.fi
[**]Okayama University, Japan
suito@okayama-u.ac.jp

Abstract:
Land use has a great impact on water quality, which can affect diatom diversity and ecology. We assessed statistically these relationships using a spatial method. As the results, using classification and regression trees, we demonstrated low pH having a great impact on diatom diversity.

1. Introduction

Technical development has made a study based on spatial information much easier. This type of study technique is utilized in various disciplines such as Population Biology (Bellier et al., 2013) and Epidemiology (Takahashi et al., 2008). A study about water quality also uses spatial information. LaBeau et al. (2014) presented relationship between land use and phosphorus (P) and estimated the amount of P loading.

We studied the impact of water quality on diatoms using land use data. We found the spatial relationship between water quality and diatom utilizing Echelon analysis. Then, we assessed the impact of water quality on diatoms using classification and regression trees (CART).

In this paper, we explained the survey area, data and analysis methods followed by the results of Echelon analysis and CART. We also identified the substances that impact on diatoms.

2. Materials and methods

Figure 2.2 shows our study procedure. We analyzed land use data, water quality data and diatoms data utilizing Echelon analysis and CART.

2.1. Study area

The study area was the mid-western part of Finland (Figure 2.1). The water quality data were collected from the 13 catchment areas (rivers) with land use data and had altogether 29 sampling sites.

2.2. Land use data

Land use data was measured in 2006. Land use of each catchment area was classified in 43 types, and they were grouped into ten classes based on soil characteristics (Table 2.1) because of too detailed to estimate a missing value.

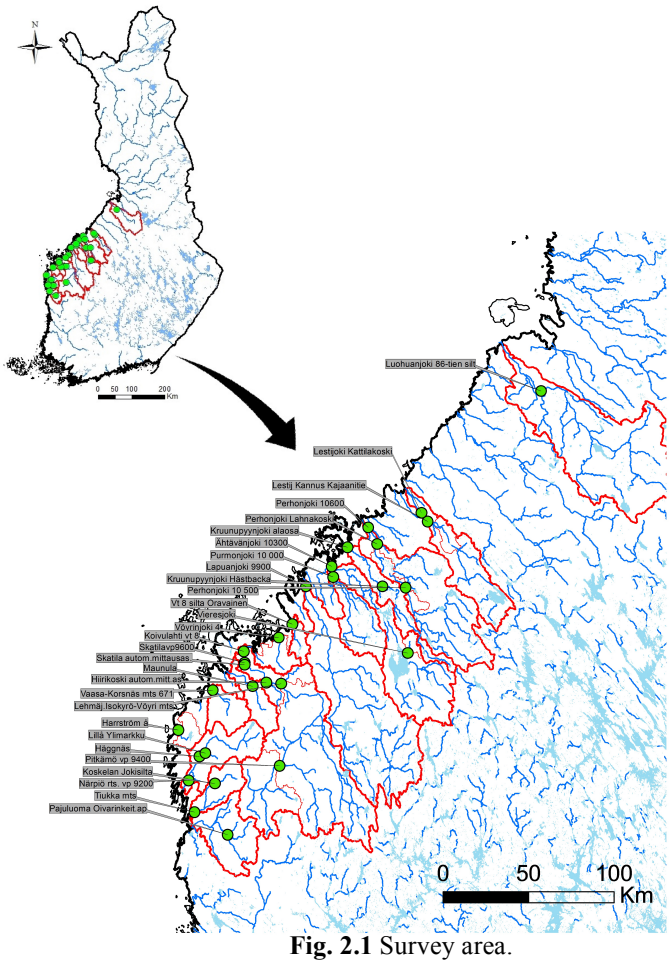

Fig. 2.1 Survey area.

Table 2.1 Ten classes of land use grouped based on soil.

Class of land use	Original land use
Artificial area	Density constructed urban areas, sparsely constructed urban areas, industrial and service areas, traffic areas, harbor areas, airport areas, area for land taking, disposal sites, under construction, parks and green areas in urban areas, summer cottages, sport areas
Agricultural area	Fields, fruit tree and berry production areas, stock feeding areas, small agricultural mosaic
Mineral soil area	Deciduous forests on mineral soil, evergreen forests on mineral soil, mixed forest on mineral soil, rare tree areas, cc <10-30% on mineral ground
Peat area	Deciduous forests of peat land, evergreen forests on peat land, mixed forest on peat land, rare tree areas, cc <10-30% on peat ground, peat production areas
Rocks area	Evergreen forests on rocks, mixed forest on rocks, rare tree areas, cc <10-30% on rock ground, rock ground

Wet area	Moor land with rare growing, inland wetlands on the land, inland wetlands in the water, open swamps, wetlands on the land area close to the sea, wetlands in the water close to the sea
Others	Natural grasslands, thin tree groups and moors, rare tree areas, cc <10%, rare tree areas over boreal forest, rare tree areas, cc <10-30% on below electricity line, sandy beaches
River	River
Lake	Lake
Sea	Sea

2.3. Water quality data

Water quality was surveyed from 2006 to 2012, and the number of sampling sites was 29. However, none of the sites had complete data (i.e. without missing variables). The Table 2.2 shows the percentages of missing variables pertaining to the water quality sampling sites.

Table 2.2 The percentages of missing variables.

Substance	Alkalinity	Al	Cd	COD	Particulate matter	Total S	Cr
Missing value rate (%)	51.3	76.9	79.9	59.0	70.7	100	80.2

Cu	Pb	Ni	Total Organic Carbon	Fe	Turbidity	Zn	SO4	Color	pH
80.2	80.2	79.9	92.3	94.1	63.0	80.2	95.6	51.3	49.1

2.4. Diatom data

Diatoms state was evaluated using two methods: number of type-specific species (TT_{40}) and percent model affinity (PMA). Both TT_{40} and PMA are between 0 and 1: close to 0 means the bad state, close to 1 means the best.

TT_{40} (Aroviita et al., 2008) is a taxon type specific index, which tells about how many taxa are found in the studied site compared to the reference site. The higher the index value, the higher the number of the same taxa in the sites. A low index value shows that there is a difference in the diatom community at the study site compared to a reference community.

PMA (Novak and Bode, 1992) describes the similarity of the reference and studied diatom community. If the index value is high, the communities in the studied and reference environment area are the same. If the value is low, the community in the studied site differs from the one in the reference site.

These data were surveyed in summer from 2007 to 2012, and the number of survey areas

varied from year to year. 2010 was the highest number of survey with 17 areas, and the lowest was 2008 with no survey taken place.

2.5. Models for land use data and water quality

Water quality data had multiple missing values; therefore, we made the complete data using two kinds of method; near sampling site data and estimating generalized linear model (GLM). We preferably used the sampling site data from the same catchment area, near sites and the same river. Nevertheless, if data had a missing value, we estimated them using GLM according to equation 2.1.

$$Y_j = \beta_0 + \sum_i \beta_i X_i + \varepsilon \qquad (2.1)$$

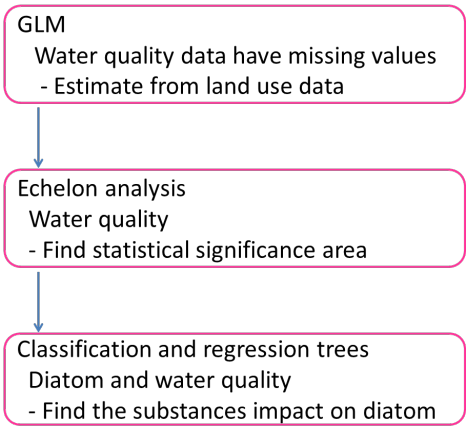

Fig. 2.2 Schematic diagram.

where Y_i is a variable like alkalinity and X_i are year, season and each land area like artificial area or agriculture area of Table 2.1.

However, the estimation the substance like metals was not successful due to too many missing values. Therefore, we estimated the missing values of alkalinity, COD, turbidity and pH.

2.6. Spatial model for water quality

Establishing an Echelon dendrogram makes it easy to detect statistically significant areas. For additional information, see a section 3 of "Juvenile salmon patch identification and comparison using Echelon analysis".

We detected hotspot areas from 3D data based on time series data, therefore time (season and year) and areas information were included in significantly high areas. Statements of significance are based on $p \leq 0.05$.

2.7. Models for water quality and diatoms

Classification and regression trees (CART) were obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree (Loh, 2011).

A variable was divided based on impact on the response variable. In this case, response variables were the diatom indexes and explanatory variables were the water quality data.

3. Results

3.1 Hotspot areas of alkalinity, COD, pH and turbidity

We detected hotspot areas of alkalinity, COD, pH and turbidity using an Echelon analysis. Each Echelon dendrogram is shown in Figures 3.1, 3.2, 3.3 and 3.4, and Tables 3.1, 3.2, 3.3 and 3.4 explain the hotspot areas. The red symbols represent the most likely significance areas and the orange symbols represent the secondary likely areas in each figure. The *p*-values of all the most likely hotspot areas and secondary areas were less than 0.5; therefore, the detected areas were the statistically significant areas. *K* in figures is the maximum number of areas in the most likely cluster. The secondary cluster can also be detected in the same manner. Here, *K*=27 (approximately 10% of the total) was adapted. However, the number of diatom samples was 47.

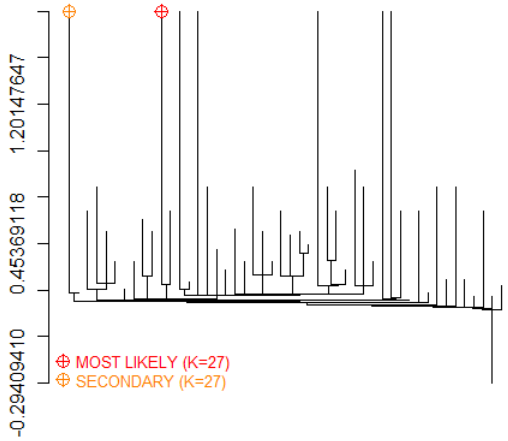

Fig. 3.1 The Echelon dendrogram of alkalinity (2006-2012).

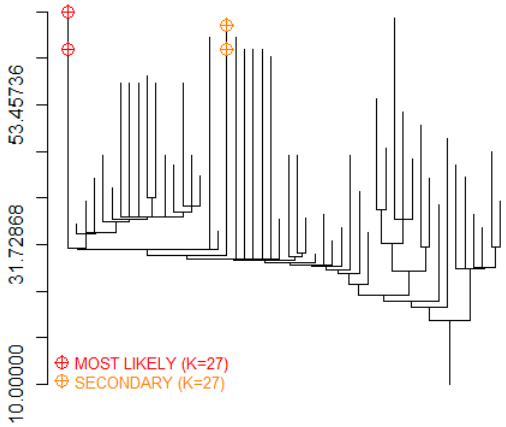

Fig. 3.2 The Echelon dendrogram of COD (2006-2012).

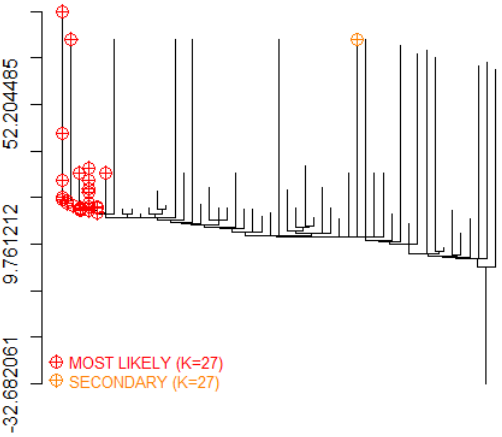

Fig. 3.3 The Echelon dendrogram of turbidity (2006-2012).

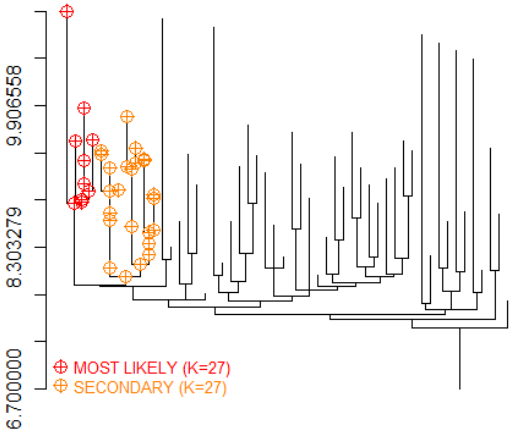

Fig. 3.4 The Echelon dendrogram of low pH (2006-2012).

Table 3.1 The Hotspot areas of alkalinity (2006-2012).

	p-value	season	year	sampling site
hotspot	0.001	summer	2006	Purmoj.10k
secondary	0.001	summer	2012	Purmoj.10k

Table 3.2 The Hotspot areas of COD (2006-2012).

	p-value	season	year	sampling site
hotspot	0.001	summer	2012	Purmoj.10k
		fall	2012	Purmoj.10k
secondary	0.001	summer	2011	Purmoj.10k
		fall	2011	Purmoj.10k

Table 3.3 The Hotspot areas of turbidity (2006-2012).

	p-value	season	year	sampling site
hotspot	0.001	spring	2006	Purmoj.10k
		spring	2006	Lapua 9900
		spring	2006	Vt8Oravainen
		spring	2006	Koivulahti vt8
		spring	2006	Maunula
		spring	2006	VaasaKorsnäs
		spring	2006	Harrström a
		spring	2006	Närpiö9200
		spring	2006	Tiukka
		summer	2006	Purmoj.10k
		summer	2006	Vt8Oravainen
		summer	2006	Harrström a
		fall	2006	Lapua 9900
		fall	2006	Vt8Oravainen
		fall	2006	Koivulahti vt8
		fall	2006	VaasaKorsnäs
		fall	2006	Närpiö9200
		spring	2007	Vt8Oravainen
		spring	2007	Maunula
		spring	2007	Harrström a
		spring	2007	Närpiö9200
		spring	2007	Tiukka
		summer	2007	Vt8Oravainen
		summer	2007	Harrström a
		fall	2007	Vt8Oravainen
secondary	0.001	summer	2012	Purmoj.10k

Table 3.4 The Hotspot areas of low pH (2006-2012).

	p-value	season	year	sampling site
hotspot	0.001	fall	2006	Lestj.Kattilakosk
		fall	2006	Perhonj10600
		fall	2006	Kruununpyyn. ala
		fall	2006	Ähtävänj10300
		fall	2006	Purmoj.10k
		fall	2006	Lapua 9900
		fall	2006	Vt8Oravainen
		spring	2007	Kruununpyyn. ala
		spring	2007	Purmoj.10k
		spring	2007	Lapua 9900
		spring	2007	Vt8Oravainen
secondary	0.001	fall	2006	Maunula
		fall	2006	VaasaKorsnäs
		fall	2006	Harrström a
		fall	2006	Närpiö9200
		fall	2006	Tiukka
		spring	2007	Maunula
		spring	2007	Harrström a
		spring	2007	Närpiö9200
		summer	2007	Närpiö9200
		fall	2007	Purmoj.10k
		fall	2007	Lapua 9900
		fall	2007	Vt8Oravainen
		fall	2007	Maunula
		fall	2007	VaasaKorsnäs
		fall	2007	Harrström a
		fall	2007	Närpiö9200
		fall	2007	Tiukka
		spring	2008	Purmoj.10k
		spring	2008	Lapua 9900
		spring	2008	Vt8Oravainen
		spring	2008	Koivulahti vt8
		spring	2008	Maunula
		spring	2008	Harrström a
		spring	2008	Närpiö9200

Many hotspot areas included the year 2006, however diatoms were not surveyed in this year. Therefore, we tried Echelon analysis again excluding year 2006. Each Echelon dendrogram is shown in Figures 3.5, 3.6, 3.7 and 3.8, and Tables 3.5, 3.6, 3.7 and 3.8 explain the hotspot areas.

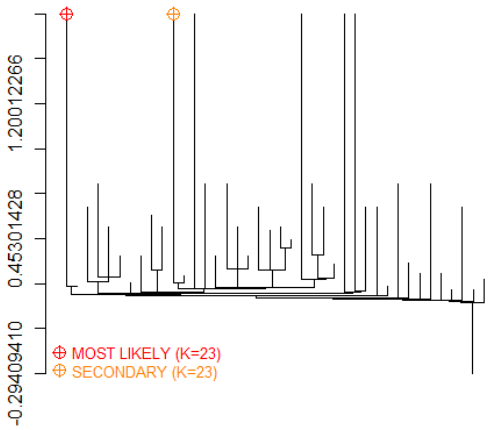

Fig. 3.5 The Echelon dendrogram of alkalinity (2007-2012).

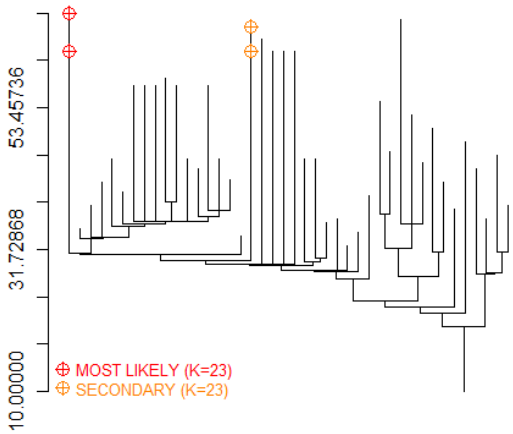

Fig. 3.6 The Echelon dendrogram of COD (2007-2012).

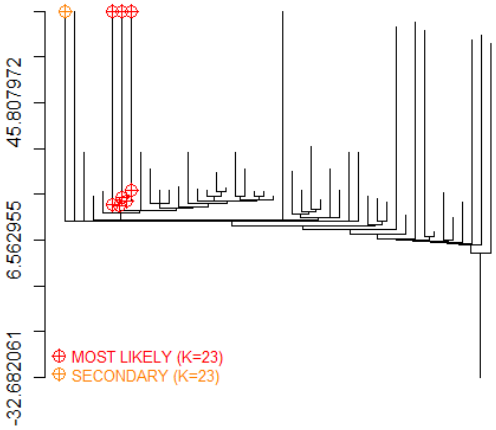

Fig. 3.7 The Echelon dendrogram of turbidity (2007-2012).

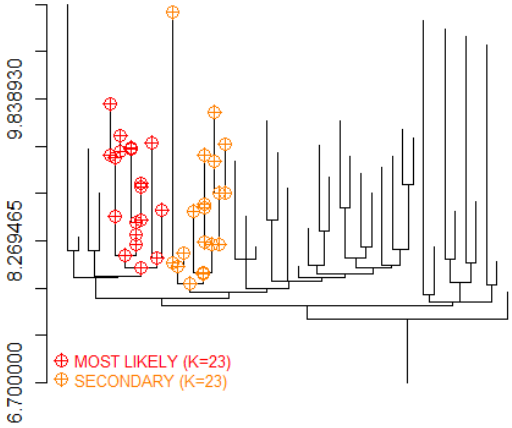

Fig. 3.8 The Echelon dendrogram of low pH (2007-2012).

Table 3.5 The Hotspot areas of alkalinity (2007-2012).

	p-value	season	year	sampling site
hotspot	0.001	summer	2012	Purmoj.10k
secondary	0.001	summer	2011	Purmoj.10k

Table 3.6 The Hotspot areas of COD (2007-2012).

	p-value	season	year	sampling site
hotspot	0.001	summer	2012	Purmoj.10k
		fall	2012	Purmoj.10k
secondary	0.001	summer	2011	Purmoj.10k
		fall	2011	Purmoj.10k

Table 3.7 The Hotspot areas of turbidity (2007-2012).

	p-value	season	year	sampling site
hotspot	0.001	spring	2007	Purmoj.10k
		summer	2007	Purmoj.10k
		fall	2007	Purmoj.10k
		spring	2008	Purmoj.10k
		summer	2008	Purmoj.10k
		fall	2008	Purmoj.10k
		spring	2009	Purmoj.10k
		summer	2009	Purmoj.10k
secondary	0.001	summer	2012	Purmoj.10k

Table 3.8 The Hotspot areas of low pH (2007-2012).

	p-value	season	year	sampling site
hotspot	0.001	spring	2007	Maunula
		spring	2007	Harrström a
		spring	2007	Närpiö9200
		summer	2007	Närpiö9200
		fall	2007	Purmoj.10k
		fall	2007	Lapua 9900
		fall	2007	Vt8Oravainen
		fall	2007	Maunula
		fall	2007	VaasaKorsnäs
		fall	2007	Harrström a
		fall	2007	Närpiö9200
		fall	2007	Tiukka
		spring	2008	Purmoj.10k
		spring	2008	Lapua 9900
		spring	2008	Vt8Oravainen
		spring	2008	Koivulahti vt8
		spring	2008	Maunula
		spring	2008	Harrström a
		spring	2008	Närpiö9200
secondary	0.001	fall	2008	Lestj.Kattilakosk
		fall	2008	Perhonj10600
		fall	2008	Purmoj.10k
		fall	2008	Lapua 9900
		fall	2008	Vt8Oravainen
		fall	2008	Koivulahti vt8
		fall	2008	Maunula
		fall	2008	VaasaKorsnäs
		fall	2008	Harrström a
		fall	2008	Närpiö9200
		fall	2008	Tiukka
		spring	2009	Perhonj10600
		spring	2009	Kruununpyyn. ala
		spring	2009	Ähtävänj10300
		spring	2009	Purmoj.10k
		spring	2009	Lapua 9900
		spring	2009	Vt8Oravainen

	spring	2009	Koivulahti vt8
	spring	2009	Maunula
	spring	2009	Harrström a
	spring	2009	Närpiö9200
	summer	2009	Kruununpyyn. ala

3.2 Water quality impact on diatom diversity

We assessed the impact of water quality on diatom diversity utilizing CART (Figure 3.9 and 3.10). We got different results on diatom indexes. In case of TT_{40}, the important variables were turbidity, alkalinity and pH. On the other hand, the important substances for PMA were pH and COD.

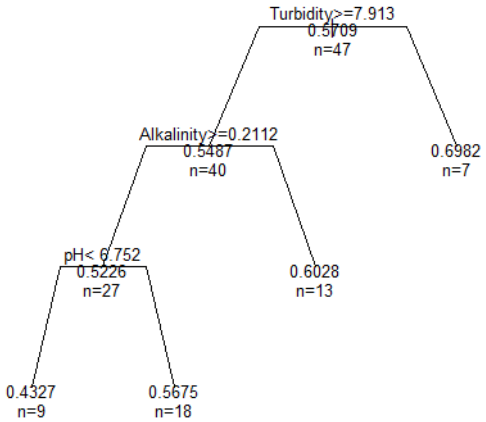

Fig. 3.9 Response variable: TT_{40}.

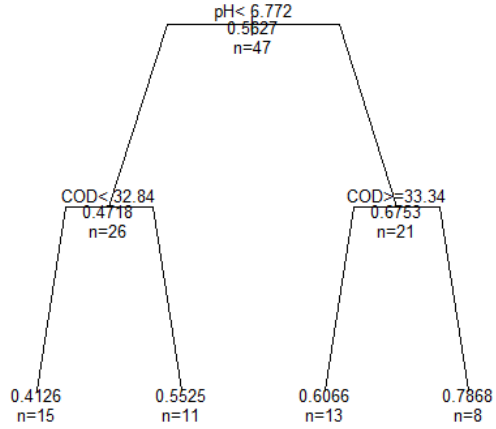

Fig. 3.10 Response variable: PMA.

4. Conclusions

In this study utilizing CART, it was found that the diatom diversity was impacted by multiple variables. On the other hand, we detected statistically significant areas of these variables using Echelon analysis. Water alkalinity and COD did not have a significant effect on the diatom indices. Instead, pH and turbidity had an effect on some sampling sites. However these areas did not match diatom survey area because of the difference of survey time. If the diatom had been surveyed at the same time and the same site of detected hotspot areas, we might have obtained more reliable results. However, we can also assume that the circumstances have not changed between the years. Further studies are needed to explore the role of the catchment area use on the diatom responses.

References

[1] E. Bellier, P. Monestiez, G. Certain, J. Chadœuf and V. Bretagnolle, "Reducing the uncertainty of wildlife population abundance: model-based versus design-based estimates", *Environmetrics*, **24**, pp. 476-488. 2013.

[2] F. Ishioka, K, Kurihara, H. Suito, Y. Horikawa and Y. Ono, "Detection of hotspots for three-dimensional spatial data and its application to environmental pollution data", *Journal of Environmental Science for Sustainable Society*, **1**, pp. 15–24, 2007.

[3] J. Aroviita, E. Koskenniemi, J. Kotanen and H. Hämäläinen, "A priori typology-based prediction of benthic macroinvertebrate fauna for ecological classification of rivers", *Environmental Management*, **42**, pp. 894–906, 2008.

[4] K. Takahashi, T. Yokoyama and T. Tango, "An introduction to disease mapping and disease clustering", *Journal of the National Institute of Public Health*, **57**, 2, pp. 86-92, 2008.

[5] M. A. Novak and E. W. Bode, "Percent model affinity: a new measure of macroinvertebrate community composition", *Journal of North American Benthological Society*, **11**, pp. 80–85, 1992.

[6] M. B. LaBeau, D. M. Robertson, A. S. Mayer, B. C. Pijanowski and D. A. Saad, "Effects of future urban and biofuel crop expansions on the riverine export of phosphorus to the Laurentian Great Lakes", *Ecological Modelling*, **277**, pp. 27-37, 2014.

[7] W. Myers, G. P. Patil and K. Joly, "Echelon approach to areas of concern in synoptic regional monitoring", *Environmental and Ecological Statistics*, **4**, pp. 131–152, 1997.

[8] Wei-Yin Loh, "Classification and regression trees", *WIREs Data Mining and Knowledge Discovery*, **1**, pp. 14-23, 2011.