Juvenile salmon patch identification and comparison using Echelon analysis

Makiko Oda[*], Saija Koljonen[**], Fumio Ishioka[†], Petteri Alho[††], Hiroshi Suito[†], Timo Huttula[**] and Koji Kurihara[†]

[*] National Defense Medical College
3-2 Namiki, Tokorozawa, Saitama, 359-8513, Japan
oda@ndmc.ac.jp
[**]Finnish Environment Institute
Saija.Koljonen@ymparisto.fi, timo.huttula@ymparisto.fi
[†] Okayama University, Japan
fishioka@law.okayama-u.ac.jp, suito@okayama-u.ac.jp, kurihara@ems.okayama-u.ac
[††] Department of Geography and Geology, University of Turku
mipeal@utu.fi

Abstract:

We studied a habitat patch identification technique for juvenile Atlantic salmon utilizing the suitability index and Echelon analysis. After identification of patches, we evaluated each identified patch using the likelihood ratio statistic so that the best suitability areas for salmon were determined. Laser and sonar are interferometric measuring systems we used for our salmon data. As a result of comparison between these two data sets, we found that sonar was more optimistic than laser for measuring suitability index.

1. Introduction

Various patch identification methods for habitats have been published (Fortin et al., 1995; Plotkin et al., 2002). A patch is defined as a spatially homogeneous area where at least one variable has similar attributes either of category or quantitative value (Fortin et al., 2005). Therefore, a patch was adopted in studies for both plants and animals. Patches can be identified using various variables such as tree or animal abundance and percentage coverage of trees. However, a solid method of identifying patches has not been established yet.

In the present study, we identified juvenile salmon patches using Echelon analysis. Oda et al. (2012) suggested the technique for identification of patches in a forest using Echelon analysis. We presented that the technique can also be used for an animal, juvenile salmon, and evaluated the identification patches.

In this paper, we first explained the patch identification method we used, and identified juvenile salmon patches based on the suitability index. We then assessed these identified

patches using the two different patch evaluation methods called laser and sonar, and finally compared the data derived from these two patch evaluation methods for measuring suitability index.

2. Survey site and data

The subarctic river Utsjoki is a tributary of a highly productive Atlantic salmon (*Salmo salar*, L.) river, Tana River. We modeled a stream stretch, which is known to have a strong Atlantic salmon juvenile population. The modeling procedure was undertaken for two different datasets with different accuracy: laser and sonar.

We applied life stage-specific habitat preference criteria (HPC) for depth, velocity and substrate preference by Atlantic salmon juvenile (Mäki-Petäys et al., 2002, 2004, unpublished data at http://www.rktl.fi/www/uploads/pdf/raportti284.pdf) to acquire a combined suitability index (CSI) value (product of habitat preference values). For this analysis we used only one discharge situation (20 m^3s^{-1}) recognized as a normal summer flow.

3. Echelon analysis

Echelon analysis (Myers et al., 1997; Kurihara et al., 2000; Ishioka et al., 2007) is a method to show phase structure in Echelon dendrogram based on response variables and neighboring information.

3.1. Echelon analysis procedure

A phase structure of data can be shown using Echelon dendrogram (Figure 3.1): The surface value is shown such as contour lines in the left picture in Figure 3.1. The middle and right pictures show a lateral view. The right picture is Echelon dendrogram.

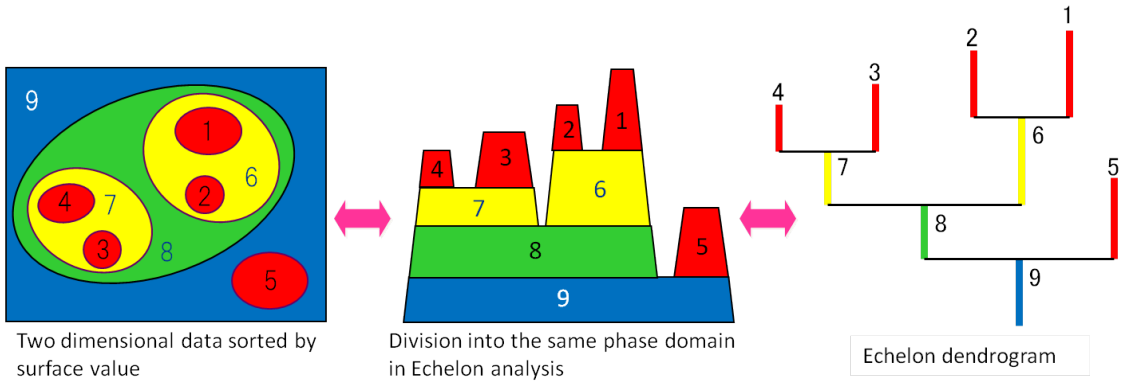

Fig. 3.1 The diagram of Echelon analysis.

We call the Echelon numbers 1, 2, 3, 4 and 5 as a “peak”, the Echelon numbers 6, 7 and 8 as a “foundation” and the Echelon number 9 as a “root”.

Each cell has a value as presented in Figure 3.2, and each cell has neighboring information. For example, cell [C3] is located at the center of its adjacent cells of [B2], [C2], [D2], [B3], [D3], [B4], [C4] and [D4].

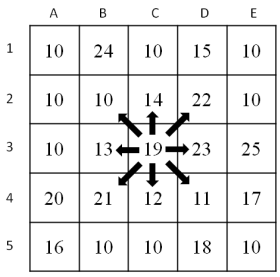

Fig. 3.2 5×5 array data.

1. Establish peaks

   At first, the cell [E3], with the maximum value in the 5×5 array data, is identified (Figure 3.3). Among adjacent cells of the cell [E3], the largest value, the cell [D3] is identified. The third ([D2]) and fourth ([C3]) values are identified in the same manner. The cell [C3] has the adjacent cell [B4], which is larger than the cell [C3]. Therefore, the first peak consists of three cells ([D2], [D3], [E3]). Other peaks {G(2), G(3), G(4)} are found in same manner (Figure 3.4).

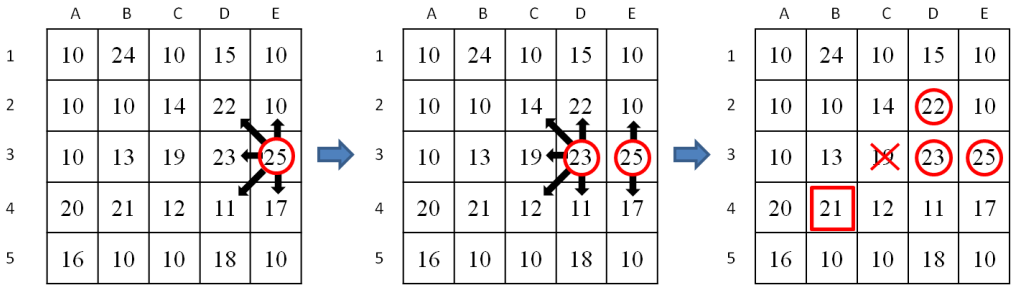

Fig. 3.3 Established peaks procedure.

2. Establish foundations

   Excluding the four peaks identified earlier, the cell [C3] is the maximum value in 5×5 array data based on its spatial location relationship (Figure 3.5), followed by the cell [E4]. Next, the cell [C2] becomes foundation between the G(2) and foundation [E4], [A5] and [D1]. The rest of cells are included in same [C2] foundation, since these are inevitably adjacent to peaks and the foundations.

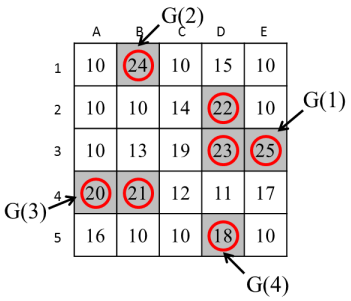

Fig. 3.4 Established peaks.

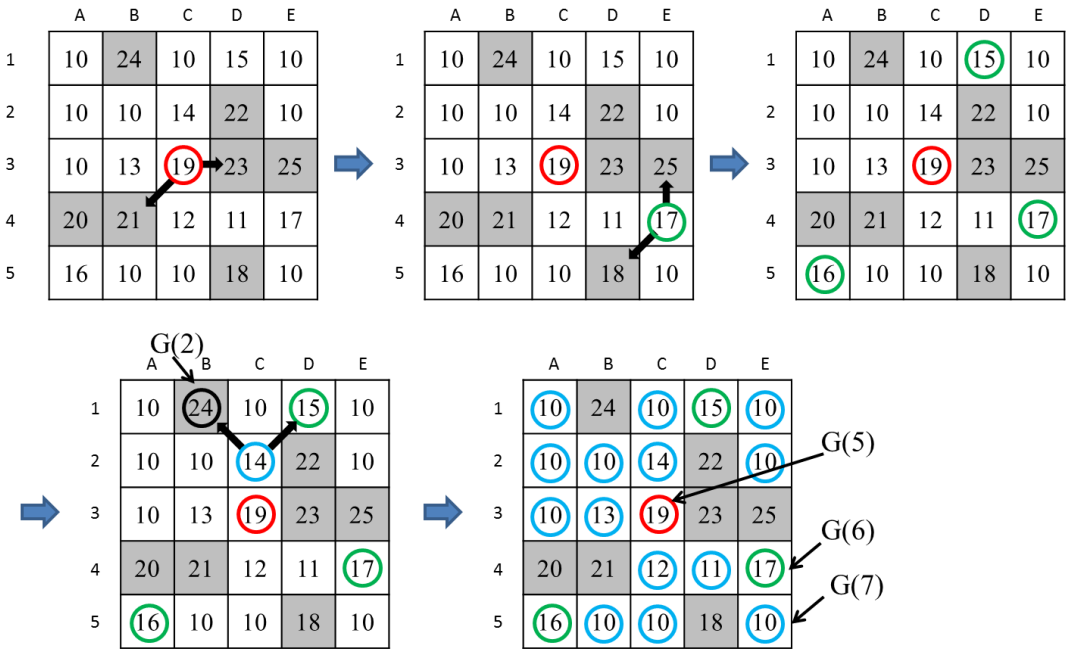

Fig. 3.5 Established foundations.

3. Establish Echelon dendrogram

Echelon dendrogrm can be established when peaks and foundations are identified in 5×5 array data. Four identified peaks: G(1) – G(4) are linked to foundations in order to draw a dendrogram (Figure 3.6). The foundation G(5) links the peaks G(1) and G(3), and the cell E[4] links the peak G(4) and foundation G(5). The cells [A5] and [D1] belong to the foundation of cell [E4]. These three cells establishes the foundation G(6). The rest of cells are the root G(7).

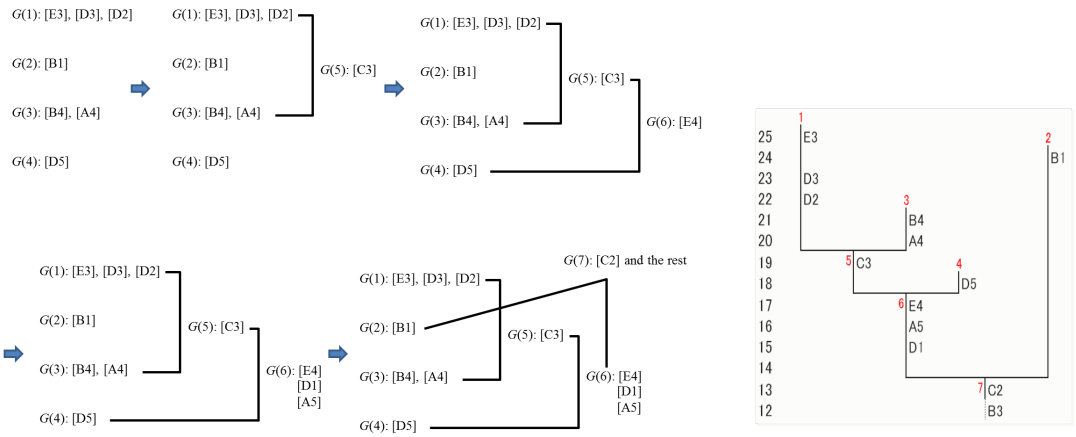

Fig. 3.6 Echelon dendrogram procedure.

2.2. Hotspot Detection

Establishing an Echelon dendrogram makes it easy to detect statistical significance among groups like G(1). This is an advantage of Echelon analysis. We call this significant area "hotspot". There is a subset area Z in the area G. p_1 is the population probability with an attribute within the area Z, and p_2 is the population probability with an attribute outside the area Z. The probabilities of all individuals with attributes are mutually independent. The hypothesis is as follows:

$$H_0 : p_1 = p_2 = p \qquad v.s. \qquad H_1 : p_1 > p_2$$

Where $n(G)$ is the total population in the area G, $n(Z)$ is the population within the area Z, $c(G)$ is the number of attributes in the area G, and $c(Z)$ is the number of the attributes within the area Z (Figure 3.7). Poisson model is used.

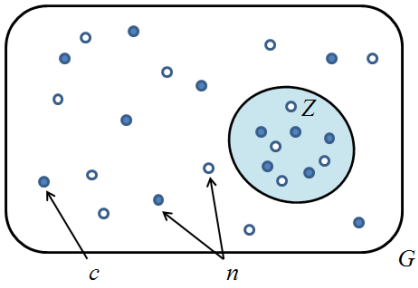

Fig. 3.7 Population n and the number of the attribute c in area G. Comparing $p_1 = c(Z)/n(Z)$ and $p_2 = c(G)/n(G)$.

The probability of the number of points is $c(G)$ in the area G is

$$\frac{\exp[-p_1 n(Z) - p_2 n(\bar{Z})][p_1 n(Z) + p_2 n(\bar{Z})]^{c(G)}}{c(G)!} \qquad (3.1)$$

The density at location x in the area G is

$$\frac{p_1 n(x)}{p_1 n(Z) + p_2 n(\bar{Z})} \quad \text{if} \quad x \in Z \qquad (3.2)$$

$$\frac{p_2 n(x)}{p_1 n(Z) + p_2 n(\bar{Z})} \quad \text{if} \quad x \in \bar{Z} \qquad (3.3)$$

The likelihood function of Poisson model is

$$\begin{aligned} L(Z, p_1, p_2) &= \frac{\exp[-p_1 n(Z) - p_2 n(\bar{Z})][p_1 n(Z) + p_2 n(\bar{Z})]^{c(G)}}{c(G)!} \\ &\times \prod_{x_i \in Z}^{n_i \in Z} \frac{p_1 n(x)}{p_1 n(Z) + p_2 n(\bar{Z})} \prod_{x_i \in Z^c}^{n_i \in Z^c} \frac{p_2 n(x)}{p_1 n(Z) + p_2 n(\bar{Z})} \qquad (3.4) \\ &= \frac{\exp[-p_1 n(Z) - p_2 n(\bar{Z})}{c(G)!} p_1^{c(Z)} p_2^{c(\bar{Z})} \prod_{x_i}^{n} n(x_i) \end{aligned}$$

Let $x(Z)$ be a random variable for the number of attributes within the area Z. Anywhere under the area Z,

$$x(A) \sim Poisson(p_1 n(A \cap Z) + p_2 n(A \cap \bar{Z})) \qquad (3.5)$$

Under the null hypothesis,

$$x(A) \sim Poisson(pn(A)) \tag{3.6}$$

To maximize the likelihood function, the maximum likelihood function over the area Z is calculated. The maximum likelihood estimators $\hat{p}_1 = c(Z)/n(Z)$ and $\hat{p}_2 = c(\overline{Z})/n(\overline{Z})$ are estimated.

$$L(Z) = \frac{\exp[-c(G)]}{c(G)!}\left(\frac{c(Z)}{n(Z)}\right)^{c(Z)}\left(\frac{c(\overline{Z})}{n(\overline{Z})}\right)^{c(\overline{Z})}\prod_{x_i}^{n} n(x_i) \tag{3.7}$$

The likelihood ratio $\lambda(Z)$ is the maximum value in the subset area within the area G to detect hotspots.

$$\lambda(Z) = \frac{\max_Z L(Z)}{L_0} = \frac{\left(\frac{c(Z)}{n(Z)}\right)^{c(Z)}\left(\frac{c(\overline{Z})}{n(\overline{Z})}\right)^{c(\overline{Z})}}{\left(\frac{c(G)}{n(G)}\right)^{c(G)}} \tag{3.8}$$

L_0 is the following likelihood function under the null hypothesis,

$$L_0 := \sup_p \frac{\exp[-pn(G)]}{c(G)!} p^{c(G)} \prod_{x_i}^{n} n(x_i) = \frac{\exp[-c(G)]}{c(G)!}\left(\frac{c(G)}{n(G)}\right)^{c(G)}\prod_{x_i}^{n} n(x_i) \tag{3.9}$$

The test statistic $\lambda(Z)$ can be

$$\lambda(Z) = \left(\frac{c(Z)}{e(Z)}\right)^{c(Z)}\left(\frac{c(\overline{Z})}{e(\overline{Z})}\right)^{c(\overline{Z})} \tag{3.10}$$

where $e(Z)$ is the expectation of the attribute within the area Z, and $e(G)$ is equal to $c(G)$.

2.3 Hotspot detection procedure using Echelon dendrogram

The hotspot detection of Echelon analysis is as follow:

1. Draw the Echelon dendrogram for target data.
2. Scan the areas from upper Echelon to the bottom, based on the hierarchical structure determined in Step 1.
3. Detect the hotspot, which takes the maximum log likelihood ratio; $\log\lambda(Z)$.

The significance of the hotspot candidate is evaluated using Monte Carlo simulation.
A p-value based on Monte Carlo simulation can be obtained as follow:

1. Generate a random data set under the null hypothesis when we condition with conditions on the total number of attribute $c(G)$.
2. Calculate the $\log\lambda(Z)$ from the simulated data.
3. Repeat a process of Step 1 and 2 multiple times.

4. Define as $p = R/(\sharp SIM + 1)$

where R is the rank of the test statistics from the real data set among all data sets and $\sharp SIM$ is the number of simulated data sets that have been generated.

Echelon analysis detects two hotspot candidates: most likely cluster and secondary cluster. The most likely cluster is the highest $\log\lambda(Z)$, and the secondary cluster $\log\lambda(Z)$ is the second highest.

4. Patch identification method

4.1. Patch identification method

First, an area with significantly high suitability was detected based on the Echelon analysis hotspot detection method (Figure 4.1 and 4.2, Table 4.1) and salmon suitability index. The vertical line of Echelon dendrogram is the suitability index. If the number of areas is too large, we can set the maximum number of hotspot areas in advance. This maximum number is K. K is the maximum number of areas in most likely cluster. The secondary cluster can also be detected in the same manner. Here, K=500 (approximately 10% of the total) was adapted (Figure 4.3). Detected areas included not good suitability (<0.5). Therefore, we proposed another patch identifying method (Figure 4.4). The top ten patches of maximum suitability within each patch area were shown (Figure 4.5 and 4.6).

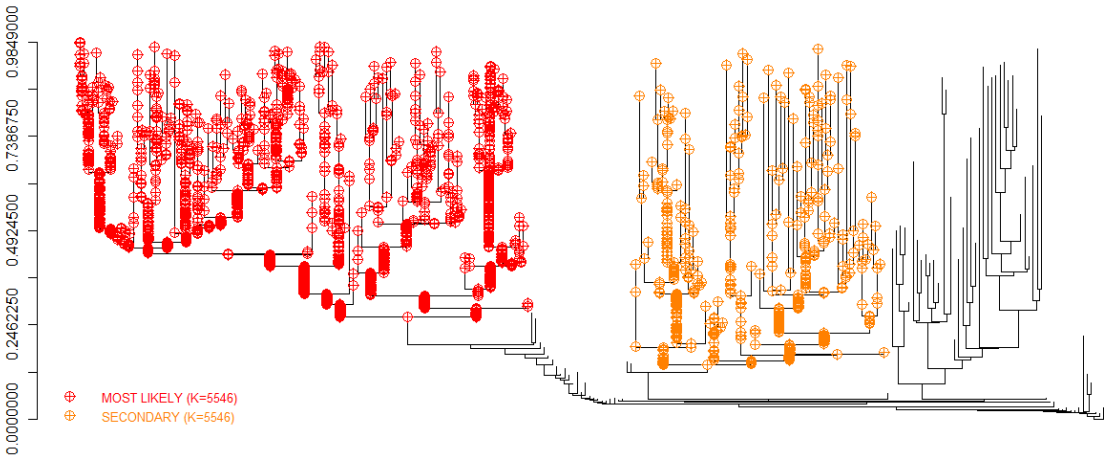

Fig. 4.1 Echelon dendrogram (laser) marked significantly high areas.

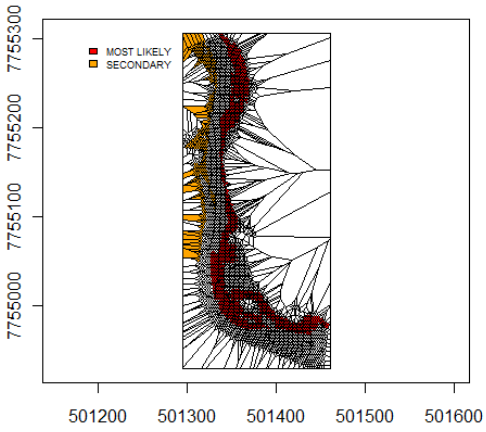

Fig. 4.2 Significantly high areas (laser).

Table 4.1 Significantly high areas (laser).

	Most likely	**Secondary**
The number of area	1707	489
logλ	383.07	19.49
p-value	0.001	0.001

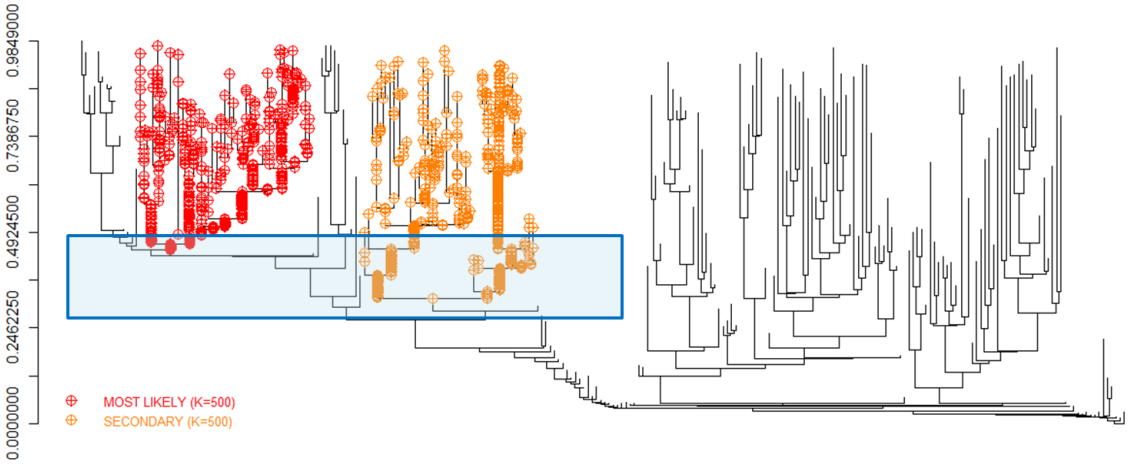

Fig. 4.3 Echelon dendrogram (laser) marked significantly high areas (*K*=500).

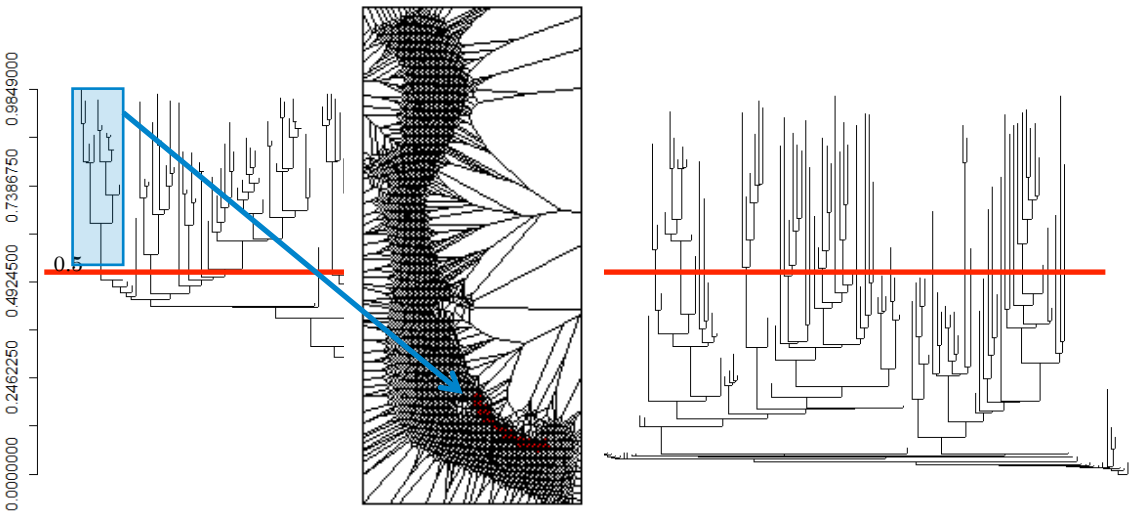

Fig. 4.4 Proposed patch identifying method.

4.2. Patch evaluation

Some identified patches are too small, therefore a patch size criterion or patch evaluation were needed. Eq. (3.10) can be useful as patch evaluation. The result is shown in Table 4.2 and 4.3. Obviously, logλ of the too small patch was small, therefore these logλ were utilized to evaluate patches.

Table 4.2 The number of cell and logλ (laser).

ID	1	2	3	4	5	6	7	8	9	10
The number of cell	10	102	770	4	79	2	12	81	22	10
logλ	42.61	22.92	208.08	0.78	17.51	0.51	2.91	13.22	5.58	1.92

Table 4.3 The number of cell and logλ (sonar).

ID	1	2	3	4	5	6	7	8	9	10
The number of cell	158	80	61	2	9	267	140	24	2	19
logλ	44.79	23.18	15.56	0.64	2.84	75.44	33.18	6.81	1.02	6.13

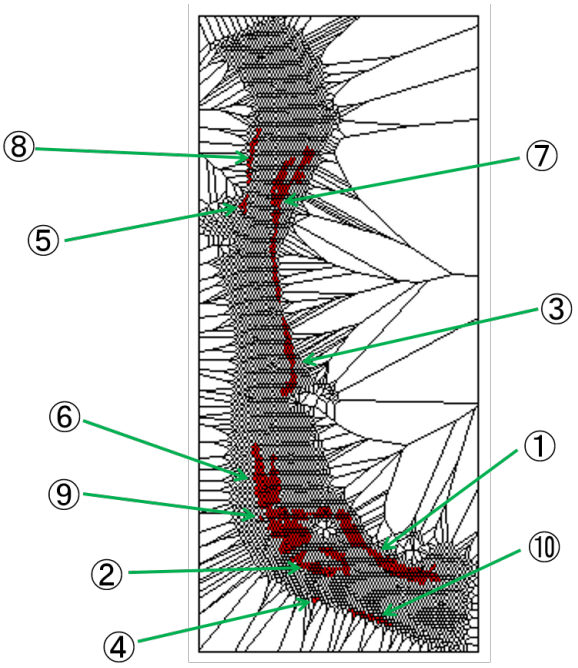

Fig. 4.5 Identified ten patches (laser).

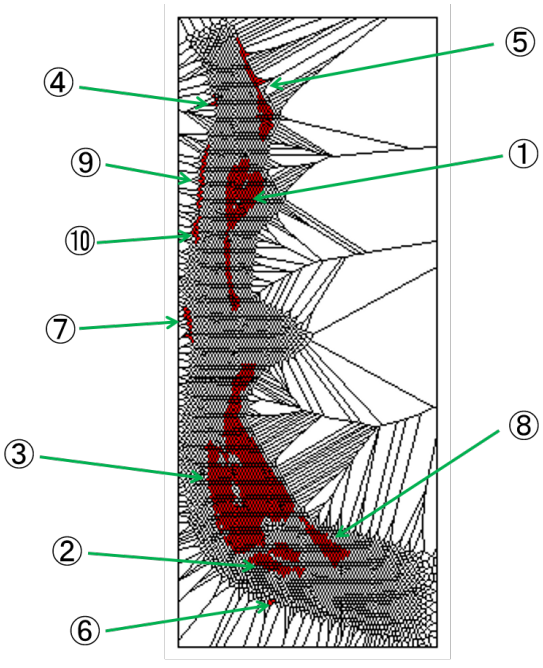

Fig. 4.6 Identified ten patches (sonar).

5. Comparing two kind of data: laser and sonar

logλ are indexes evaluated only in same area. Laser data and sonar cannot be compared using logλ. An advantage of Echelon analysis is easy scan. We can compare two kinds of data by making a dendrogram from both laser and sonar. A conventional method is detecting significance area in each area, it cannot compare some areas. New method can assess same area as one area and one dendrogram (Figure 5.1). The hypothesis is as follows:

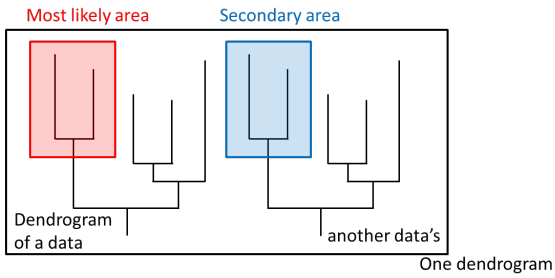

Fig. 5.1 The Echelon dendrogram with two kind of data method.

H_0 : All area's suitability indexes are equal.
H_1 : not H_0

The result is Figure 5.2. The first and forth hotspot were sonar data, second and forth were laser data. This result might have been related that the laser measurement accuracy was higher than the sonar.

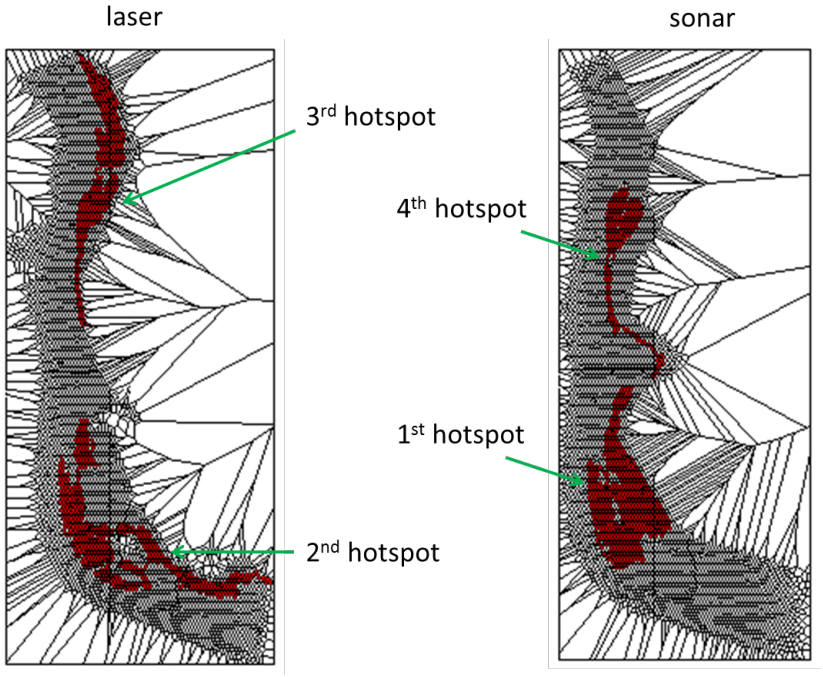

Fig. 5.2 Detected hotpot areas (p-value<0.5).

6. Discussion

In this study, we confirmed that the method of patch identification was effective in salmon habitat suitability. The identification of patches through Echelon dendrogram and salmon suitability index was more useful than detecting significant areas from spatial scan statistic. It is possible to show the best habitat for salmon by evaluating patches. However, in previous finding (Vadas and Orth, 2001), suitability index ≥ 0.75 were defined as optimal. We should further examine the critical suitability index. We also presented comparison method with two data (sonar and laser) using Echelon analysis. We want to establish the comparing method because it has much room to study.

References

[1] A. Mäki-Petäys, A. Huusko, J. Erkinaro and T. Muotka, "Transferability of habitat preference criteria of juvenile Atlantic salmon (*Salmo salar*) ", *Canadian Journal of Fisheries and Aquatic Science*s, **59**, pp. 218–228, 2002.

[2] A. Mäki-Petäys, J. Erkinaro, E. Niemelä, A. Huusko and T. Muotka, "Spatial distribution of juvenile Atlantic salmon (*Salmo salar*) in a subarctic river: size-specific changes in a strongly seasonal environment", *Canadian Journal of Fisheries and Aquatic Sciences*, **61**, 12, pp.2329–2338, 2004.

[3] 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.

[4] J. B. Plotkin, J. Chave and P. S. Ashton, "Cluster analysis of spatial patterns in Malaysian tree species", *The American Naturalist*, **160**, pp. 629–644, 2002.

[5] K. Kurihara, W. L. Myers and G. P. Patil, "Echelon analysis of the relationship between population and land cover patterns based on remote sensing data", *Community Ecology*, **1**, pp. 103-122, 2000.

[6] M.-J. Fortin and M. R. T. Dale, *Spatial Analysis: A guide for Ecologists*. New York: Cambridge, University Press, 2005.

[7] M.-J. Fortin and P. Drapeau, "Delineation of ecological boundaries: comparison of approaches and significance tests", *Oikos*, **72**, pp. 323-332, 1995.

[8] M. Oda, F. Ishioka, T. Masaki and K. Kurihara, "Forest partition using Echelon hierarchical structure", Bulletin of Data Analysis of Japanese Classification Society, **1**, 2, pp.17-31, 2012.

[9] R. L. Vadas, Jr and D. J. Orth, "Formulation of habitat suitability models for stream fish guilds: do the standard methods work?", *Transactions of the American Fisheries Society*, **130**, pp. 217-235, 2001.

[10] 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.