HELSINKI
INSTITUTE FOR
INFORMATION
TECHNOLOGY

A Study of Electrofishing Bias in Terms of Habitat and Abundance Using Information-Theoretic Tools

Kimmo Valtonen, Tommi Mononen, Petri Myllymäki, Henry Tirri,
Jaakko Erkinaro, Erkki Jokikokko, Sakari Kuikka and Atso Romakkaniemi

December 22, 2002

HIIT
TECHNICAL
REPORT
2002–5

A STUDY OF ELECTROFISHING BIAS IN TERMS OF HABITAT AND ABUNDANCE USING INFORMATION-THEORETIC TOOLS

Kimmo Valtonen, Tommi Mononen, Petri Myllymäki, Henry Tirri, Jaakko Erkinaro, Erkki Jokikokko, Sakari Kuikka and Atso Romakkaniemi

Helsinki Institute for Information Technology HIIT

Tammasaarenkatu 3, Helsinki, Finland

PO BOX 9800

FIN-02015 HUT, Finland

http://www.hiit.fi

HIIT Technical Reports 2002–5

ISSN 1458-9451

A study of electrofishing bias in terms of habitat and abundance using information-theoretic tools

Kimmo Valtonen, Tommi Mononen, Petri Myllymäki, Henry Tirri, Jaakko Erkinaro, Erkki Jokikokko, Sakari Kuikka, and Atso Romakkaniemi

Abstract: In electrofishing it is usually assumed that the abundance of fish at a site is strongly dependent on habitat type. In practice the yearly choices of sites are not perfectly representative of the distribution of habitat types in a river, so a bias is introduced into density estimates based on the observed densities. However, it is assumed that this bias is time-invariant, allowing the use of observed densities as relative values. In this work we study whether this is so using a general information-theoretic methodology in a probabilistic framework. Our methodology allows measuring the similarity of pre-existing biological knowledge and an empirical model learned from a set of new observations. It also enables a separate study of habitat sampling bias and habitat–abundance relationship over a time series. Given a set of restrictions on the eligibility of sites, as is usually the case in electrofishing, bias-minimal selections of sites to electrofish can also be provided. In our empirical studies we test our methodology on real-world data sets from two Gulf of Bothnia rivers, consisting of expert-made habitat site classifications coupled with observational electrofishing data on salmon. Our approach is general in the sense that there are no restrictions on the nature or construction method of the probabilistic models used. Furthermore, our methodology compares the models directly, instead of comparing artificial data sets generated using them.

1. Introduction

It is a basic assumption in the planning of electrofishing that there will always be a bias in the data, for various reasons: for example, electrofishing in some habitats is technically difficult, it is thought that some habitats are too hostile to support any fish, or there are too many sites to allow electrofishing with full representativeness. It is however assumed that this allowed bias is time-invariant, enabling the use of observed densities as relative values comparable over a time series.

Let us state briefly the premises adopted in this work. Our basic assumption is that the abundance of fish at a particular site depends on the habitat type of that site. By *abundance* we mean relative density, i.e. the density of a particular age group given a particular habitat, relative to the densities of that age group in other habitats, during a particular year. The actual absolute densities naturally depend on the absolute size of the population, which we assume to be dependent on other factors (such as the numbers of ascending adults in previous years) excluded from this analysis.

We furthermore make some assumptions of more technical nature. We assume that the habitat classifications are accurate and sufficient. In reality this is not exactly true: for example, we know that the electrofishing sites cover only a tiny portion (less than 1%) of the habitat-classified areas they are part of (see Fig. 1). The habitat classifications of the sites stay the same over time in our data. This enforces upon us an assumption of time-invariance, which might also be questioned. These technical assumptions depend solely on the nature of our data: given more accurate data, e.g. habitat-classification of electrofishing sites instead of larger areas, our models would reflect nature more accurately.

Given these assumptions, we study in this paper the interplay of habitat and abundance. We demonstrate a methodology which allows us to compare existing biological knowledge to empirical models built from new observations. What is more, our methodology enables us to study habitat sampling bias and the relationship of habitat and abundance both separately and together. We provide examples of application to various types of fisheries problems: both as a tool for data analysis and as a planning aid in the selection of sites to electrofish. In our empirical studies, we apply our methodology to real-world data from two Gulf of Bothnia salmon rivers, Simo and Tornio (Finnish side).

We start by describing in Chapter 2 our modeling approach. In Chapter 3 we define our methodology formally, proceeding to show examples of its application in Chapter 4 using artificial data. We describe our real-world data sets in Chapter 5, expounding our empirical work on them in Chapter 6. Finally, we discuss the results and outline future work in Chapter 7.

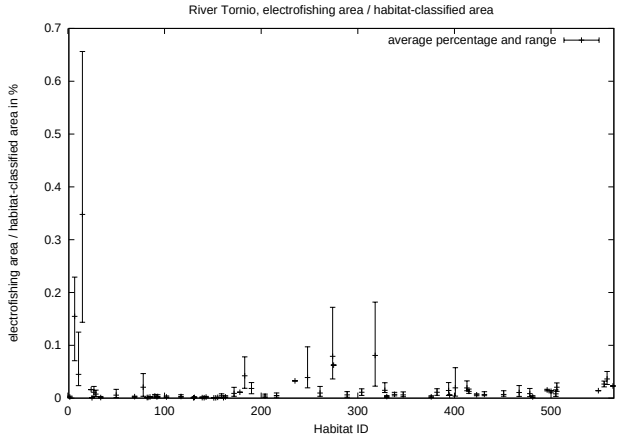
2. Modeling approach

Given biological knowledge (or a set of hypotheses) and a set of new data, a natural objective is to study how well the pre-existing knowledge describes the new observations. One way of tackling this problem is to build a model describing the new data, proceeding to compare the resulting empirical model to the knowledge. To enable comparison, a common language and structure for expressing both models is needed.

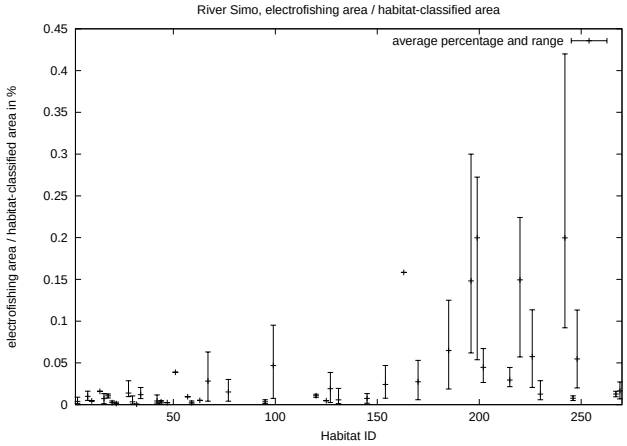
In this work we have picked probability theory as such a language, i.e. we assume that both the knowledge and the empirical model are probability distributions. Given the assumptions

Kimmo Valtonen, Tommi Mononen, Petri Myllymäki, and Henry Tirri. Complex Systems Computation Group (CoSCo), Helsinki Institute for Information Technology (HIIT), P.O. Box 9800, FIN-02015 HUT, Finland. http://cosco.hiit.fi/, Firstname.Lastname@hiit.FI
Jaakko Erkinaro, Erkki Jokikokko, Sakari Kuikka, and Atso Romakkaniemi. Firstname.Lastname@rktl.fi

Fig. 1. The ratio of electrofishing areas and the habitat-classified areas covering them, average and range across the available time series. (a) River Tornio. (b) River Simo.



(a)



(b)

of Chapter 1, we are thus interested in modeling $P(H, A)$, the joint distribution of habitat and abundance, where H denotes habitat type and A abundance of salmon. From basic laws of probability it follows that this joint distribution can be written down using two different structures:

$$(1) \qquad P(H, A) = P(H)P(A \mid H)$$
$$(2) \qquad = P(A)P(H \mid A).$$

When learning an empirical non-causal model from a data set, both structures are viable. Biological knowledge, however, is much easier to express using (1), due to causal intuition. Hence, we choose it as our structure.

We choose this simple model in favour of more complex ones, both to simplify exposition and because we assume that habitat is the dominating factor affecting abundance. Moreover, the main aim of this paper is to study the interplay of habitat and abundance, expounding the merits of an information-theoretic methodology while focusing on habitat selection bias. Other factors that abundance might depend on are beyond the scope of this work.

As an illustrative example of the simplifying assumptions made, the abundance of each age group is modeled separately, resulting in a set of models

$$\mathcal{M} = \{P_1(H, A_1), \ldots, P_k(H, A_k)\},$$

one for each of the k age groups. Thus, we assume that A_i, the abundance of age group i, does not affect the abundances of other age groups. (We will drop the age group index from now on to simplify notation.) We also ignore any dependencies between abundances of age groups at different points in time. This means that, for example, the abundance of age 1+ fish is assumed to be independent of the abundance of 0+ fish in the previous year. It must be stressed that whether such dependencies are taken into account or not is irrelevant with respect to the essence of our methodology. The tools exhibited in this paper deal with probabilistic models in general, regardless of their structural complexity. They would work just as well on more intricate models.

Our goal is to find a set of measures on probabilistic models, enabling the study of the following objectives:

(1) The amount of habitat sampling bias over a time series. Have the yearly site choices been representative of the river with respect to habitat type?

(2) Variance of habitat sampling bias over time. Has the bias stayed constant?

(3) Bias-minimality of the yearly choices of sites in the available time series. How far have they been from an optimally representative set of the same size (given a set of restrictions on our choices)?

(4) Variance over time in the conditional distribution of abundance given a habitat. Has the relationship of habitat and abundance stayed the same?

(5) An analysis of possible changes in the joint distribution of habitat and abundance over a time series, and recognition of whether the changes are due to a change in habitat sampling or to a change in the relation of habitat and abundance.

Objectives (1) - (3) deal with habitat sampling bias. Objectives (4) - (5) study whether habitat sampling is the only factor affecting the joint distribution of habitat and abundance.

An important aspect of our approach is that it leaves completely open the way the probabilistic models are constructed. They can be based on biological knowledge, and/or learned from data, using any preferred methodology. Therefore we can e.g. evaluate how much general biological knowledge, obtained from other studies, may help in estimation.

Let us briefly describe the traditional approach to this problem for comparison purposes. In several cases, a simulator is constructed representing "truth" (our $P(H, A)$ above), and then artificial data is generated from it, i.e. a data set of imaginary

observed densities at sites of different habitat types. The real-world data is then compared to this artificial data set using the tools of classical statistics. We will discuss in detail in Chapter 7 the ways our approach differs from the traditional one.

3. Methodology

Since our models are probability distributions, we need a means of measuring their similarity. With this in mind, we first describe some basic information-theoretic concepts, and then discuss the particular type of empirical modeling adopted in this work.

In the following we assume that our domain is discrete, i.e. our variables have either nominal or ordinal values. The habitat variable can be seen as an example of a nominal variable: there need not exist any order on the set of habitat types. Abundance on the other hand can be viewed as either continuous or discrete (but ordered). From the management point of view, an ordered but discrete value set for abundance suffices, since it allows qualitative judgements about the system.

3.1. Measuring the divergence of probabilistic models

For a general introduction to information theory, see [1]. Let X be a random variable with alphabet $\mathcal{X}$ and probability mass function $P(x)$, $x \in \mathcal{X}$. The *relative entropy* between two distributions $P(X)$ and $Q(X)$ is defined as

$$D(P(X) \parallel Q(X)) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}, \tag{3}$$

where $Q(X)$ is another probability mass function. Relative entropy is also called *Kullback-Leibler distance* [3]. Properly speaking, relative entropy is neither a distance or a measure, since it is asymmetric ($D(P \parallel Q) \neq D(Q \parallel P)$ in general) and does not satisfy the triangle inequality. Hence, we will use the term *divergence* in this presentation.

$D(P \parallel Q) \geq 0$ for all distributions P and Q, and $D(P \parallel Q) = 0$ if and only if $P(X) \equiv Q(X)$, that is, if the two distributions are the same. We will use the convention $0 \log \frac{0}{Q(X)} = 0$ for the cases when $P(X) = 0$, on the grounds that $\lim_{x \to 0} x \log x = 0$.

In intuitive terms, $D(P \parallel Q)$ is a measure of the distance between two distributions P and Q, i.e. it measures the inefficiency of assuming that the distribution is Q when the "true" distribution is P (hence the asymmetric nature.) You can also see relative entropy as the expected logarithm of the likelihood ratio, i.e. the exponent of the expected error in assuming the distribution is Q, when it in fact is P. Example 3.1 illustrates the suitability of relative entropy as a tool for the study of objectives (1) - (3).

Example 3.1. Let H describe the habitat type of a site. Let us for simpleness of exposition assume that it has only two values: *poor* and *good*. Let $P(H)$ be our "true" model for the distribution of habitat types in a river. Instead of using $P(H)$ we employ distribution $Q(H)$, however. $D(P(H) \parallel Q(H))$, the relative entropy (divergence) of $P(H)$ and $Q(H)$, is shown in Fig. 2(a).

The definition of *conditional relative entropy* is

$$\begin{aligned} & D(P(Y \mid X) \parallel Q(Y \mid X)) \\ = & \sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} P(y \mid x) \log \frac{P(y \mid x)}{Q(y \mid x)}, \end{aligned} \tag{4}$$

defining the divergence of two conditional distributions $P(Y \mid X)$ and $Q(Y \mid X)$, where Y is a random variable with alphabet $\mathcal{Y}$, and Q is a probability mass function. Example 3.2 shows how conditional relative entropy enables us to study objective (4).

Example 3.2. As in Example 3.1, let H describe the habitat type of a site. In addition to having a distribution $P(H)$ over the habitat types of sites in a river, we also have two conditional distributions $P(A \mid H)$ and $Q(A \mid H)$ describing abundance at a site given its habitat type. To keep things simple, we assume that A has only two values, *scarce* and *abundant*. An example of $P(A \mid H)$ and $Q(A \mid H)$ is shown in Table 1. E.g. according to distribution $P(A \mid H)$ the probability of there being a lot of fish when the habitat is of *poor* type is 0.1. Intuitively put, $Q(A \mid H)$ differs from $P(A \mid H)$ in being more optimistic about abundance in *poor* habitats.

In order to study the general case, let us first assume that $P(A \mid H)$ and $Q(A \mid H)$ always agree with respect to $P(\cdot \mid H = good)$ as in Table 1. What they do disagree about is the abundance of fish in *poor* habitats. Let us study the graph of divergence for different degrees of disagreement.

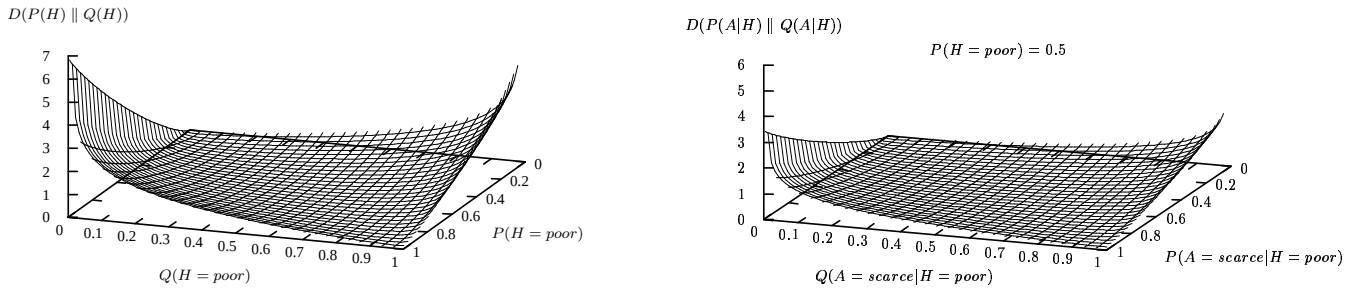
If $P(H = poor) = 0.5$, i.e. both habitat types are equally probable, $D(P(A \mid H) \parallel Q(A \mid H))$ is as shown in Fig. 2(b).

Let us compare this to a case where *poor* habitats are prevalent ($P(H = poor) = 0.9$), and to a case where they are rare ($P(H = poor) = 0.1$). The resulting conditional distribution divergences are shown in Fig. 3.

It can be seen that the "true" distribution of habitat types affects in a natural way the divergence of the conditional distributions. If a habitat type is in truth a rare one, any differences in the modeling of abundance at sites of that type add relatively little error. If, on the other hand, a type dominates a river, model divergence can potentially produce a significant amount of error. Thus, our measure behaves as desired.

Finally, putting together all of the above, the *relative entropy of a joint distribution* of two random variables X and Y has the

Fig. 2. (a) Example 3.1. The divergence of $P(H)$ and $Q(H)$. (b) Example 3.2. Divergence of $P(A \mid H)$ and $Q(A \mid H)$ when all habitat types are equally probable.



(a) (b)

Table 1. Two example distributions $P(A \mid H)$ and $Q(A \mid H)$.

	$P(A \mid H)$		$Q(A \mid H)$	
	H = poor	H = good	H = poor	H = good
A = scarce	0.9	0.3	0.6	0.3
A = abundant	0.1	0.7	0.4	0.7

following decomposition property

$$
\begin{aligned}
&D(P(X,Y) \parallel Q(X,Y)) \\
&\quad = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{Q(x,y)} \\
&\quad = \sum_{x,y} P(x,y) \log \frac{P(x)P(y \mid x)}{Q(x)Q(y \mid x)} \\
&\quad = \sum_{x,y} P(x,y) \log \frac{P(x)}{Q(x)} + \sum_{x,y} P(x,y) \log \frac{P(y \mid x)}{Q(y \mid x)} \qquad (5) \\
&\quad = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \\
&\qquad + \sum_{x} P(x) \sum_{y} P(y \mid x) \log \frac{P(y \mid x)}{Q(y \mid x)} \\
&\quad = D(P(X) \parallel Q(X)) + D(P(Y \mid X) \parallel Q(Y \mid X)).
\end{aligned}
$$

In intuitive terms, this means that if the true joint distribution of X and Y is P, but we use Q instead, we can study the divergence by studying the two constituent parts of the joint distribution separately, allowing us to tackle objective (5).

3.2. Empirical modeling

Since we have chosen to model our variables as discrete, the multinomial distribution is a natural choice for a model class. In a multinomial distribution a random variable X has a set of r_X discrete values $\mathcal{X} = \{x_1, \ldots, x_{r_X}\}$. With the structure we have chosen this entails $\mathcal{H} = \{h_1, \ldots, h_{r_H}\}$ and $\mathcal{A} = \{a_1, \ldots, a_{r_A}\}$, where $\mathcal{H}$ is our set of distinct habitat types and $\mathcal{A}$ is our set of abundance categories.

Our empirical models are thus defined by a set of parameters

$$
\begin{aligned}
\Theta = (&\theta_{h_1}, \ldots, \theta_{h_{r_H}}, \\
&\theta_{a_1|h_1}, \ldots, \theta_{a_{r_A}|h_1}, \\
&\vdots \\
&\theta_{a_1|h_{r_H}}, \ldots, \theta_{a_{r_A}|h_{r_H}}),
\end{aligned}
$$

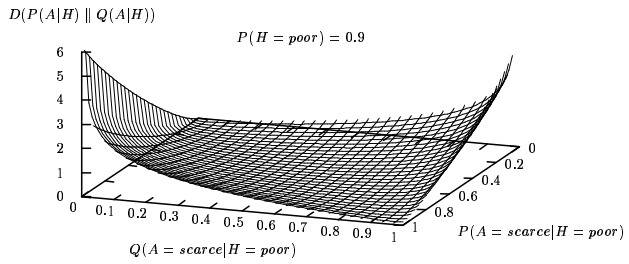
where θ_{h_j} is the probability of a site being of type h_j and $\theta_{a_k|h_j}$ is the probability of observing abundance category a_k at sites of habitat type h_j.

Given a data set, we choose the set of parameter values giving the highest probability to the observations. These *maximum likelihood* parameters $\hat{\Theta}$ are in the case of our model class of multinomial distributions normalized frequencies, i.e. $\hat{\theta}_i = n_i / \sum_j n_j$, where n_i is the number of times event i occurs in the data.

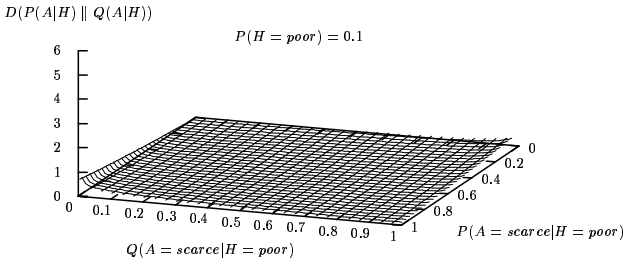
Example 3.3. Let $\mathcal{H} = \{h_1, h_2\}$ and $\mathcal{A} = \{a_1, a_2\}$, and our set of observations be $[[h_1, a_1], [h_1, a_2], [h_1, a_1], [h_2, a_2]]$. Now our maximum likelihood parameters are $\hat{\theta}_{h_1} = 3/4, \hat{\theta}_{h_2} = 1/4, \hat{\theta}_{a_1|h_1} = 2/3, \hat{\theta}_{a_2|h_1} = 1/3, \hat{\theta}_{a_1|h_2} = 0/1, \hat{\theta}_{a_2|h_2} = 1/1$.

Concerning empirical modeling in general, it should be kept in mind that the maximum likelihood parameters $\hat{\Theta}$ can fail to generalize. If the models in the chosen model class are too complex, there is a danger of overfitting the model to the particular set of observations studied. Although we have fixed our structure as (1), excessive complexity can still enter through the choice of possible habitat types and the categorization of abundance. In the case where we compare the empirical model to biological knowledge, we can assume that the knowledge is already encoded in an optimal, generalizing way (assuming

Fig. 3. Example 3.2. Divergence of $P(A \mid H)$ and $Q(A \mid H)$. (a) Sites of *poor* type are prevalent. (b) Sites of *poor* type are rare.



(a)



(b)

that the new observations have not been used in the construction of biological knowledge). Choosing the same categorizations for our empirical model should thus avoid overfitting to the observations. If both models are empirical ones, however, an overfit-avoiding criterion is needed. The MDL principle [4] is a possible information-theoretic choice in that case. We will not go further in that direction in this work, however, because our focus is on comparing knowledge and empirical models, not on the important and complicated issue of empirical modeling per se.

Remark 3.4. A technical problem inherent in the calculation of $D(P(X) \parallel Q(X))$ is the possibility of $Q(X)$ being zero for some $x \in \mathcal{X}$, causing the divergence to explode to infinity when $P(x) > 0$ (If $P(X) = 0$, we use the convention given in Chapter 3.1). To avoid this, the models should be constructed so that they offer nonzero support for all possible events. In our empirical models in this work, we pretend having observed prior to our actual measurements a small ($\ll 1$) and equal number of all possible kinds of events $[h_j, a_k]$.

4. Application of methodology

Armed with the necessary tools, we will now demonstrate how to use them to study our objectives. Let our data consist of a time series of electrofishing data collected at a set of sites S. Data for year y_i consists of density measurements at a subset of sites S_{y_i} ($S_{y_i} \subseteq S$). We assume each site has been assigned a habitat type $h_j \in \mathcal{H}$, where $\mathcal{H}$ is the set of possible habitat types. We also assume we have observations of abundances $a_{ij} \in \mathcal{A}$, where $\mathcal{A}$ is the set of abundance categories, and a_{ij} is the observed abundance category at sites of habitat type h_j in year y_i.

4.1. Studying habitat sampling bias

We will now demonstrate a set of procedures for meeting objectives (1) - (3), which deal with habitat sampling bias.

Let $P(H)$ denote the distribution of habitat types across all of the electrofishing sites of a river, regardless of whether they have ever been electrofished from or not, and let $P_{y_i}(H)$ stand for the observed habitat type distribution of a particular year y_i, that is, the distribution of habitat types in the set of sites that were electrofished during year y_i, and year y_i only. $P(H)$ represents our biological knowledge about the habitat distribution of a river, and $P_{y_i}(H)$ is an empirical model built from the data for a particular year. $P(H)$ can, for example, be based on previous studies, be a hypothesis, or be obtained from existing data via the river-global distribution of habitat types assigned to the sites.

We can now study the bias incorporated in the choice of sites for each particular year y_i in the data by means of $D(\cdot \parallel \cdot)$. By looking at $D(P_{y_i}(H) \parallel P(H))$ we can measure the amount of error we make by assuming that the sites picked for each year are representative of the entire river, i.e. that the distribution of habitats in the set of sites chosen for electrofishing in a particular year reflects their distribution in the entire river.

Example 4.1. Let H be as in Example 3.1. All of a river's 10 electrofishing sites have been habitat-classified as shown in Table 2(a). Let our biological knowledge be $P(H = poor) = 0.7$ and $P(H = good) = 0.3$, and our data on observed densities as shown in Table 2(b). We can see that each year more sites are electrofished from, until at year y_5 electrofishing occurs at all sites. The resulting yearly empirical habitat distributions are shown in Table 2(c).

We can now calculate our bias in the choice of sites for each year using $D(P_{y_i}(H) \parallel P(H))$. Fig. 4(a) shows that even though the number of sites electrofished increases each year, the bias increases as well up to year y_4 according to our measure. This fits with our intuition, since in year y_2 *good* sites are slightly under-represented compared to *poor* sites, and going towards y_4 this under-representativeness increases. At year y_5 the bias is zero, and the empirical model agrees exactly with our knowledge.

We can also study the *optimality* of yearly selections given a set of restrictions on our choice of sites. As noted earlier, electrofishing usually has a built-in bias, because some sites and habitat types are always left unfished. In addition, there is also a limit on the number of electrofished sites. In our real-world data the limit lies at 11%.

Our measure allows us to study for each possible number of electrofished sites the best possible $D(P_{y_i}(H) \parallel P(H))$ for a given restriction on our choice of sites. This means that we can find an optimal subset of $S_e \subseteq S$, the eligible sites, to electrofish. This is made computationally easy by the convexity of $D(\cdot \parallel \cdot)$, enabling us to use a hill-climbing search algorithm. Note that there are N,

$$N = \binom{|S_e|}{|S_{y_i}|}, \tag{6}$$

Table 2. Example 4.1. (a) The types of the sites. (b) The time series of observed densities. (c) Yearly empirical habitat type distributions.

S	H
s_1	poor
s_2	good
s_3	poor
s_4	good
s_5	poor
s_6	poor
s_7	poor
s_8	poor
s_9	good
s_{10}	poor

(a)

Year		s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
y_1	Age class 0		6.2	1.0		1.6					
	Age class 1		3.2	0.3		0.9					
	Age class 2		1.5	0		0.2					
y_2	Age class 0	1.1	3.5	0.4	3.0	0.9					
	Age class 1	0.3	1.8	0.1	2.1	0.4					
	Age class 2	0.1	0.7	0	1.1	0.1					
y_3	Age class 0	1.3	4.2	0.4		0.9		0.6	0.9		1.1
	Age class 1	0.2	1.6	0.1		0.4		0.3	0.4		0.9
	Age class 2	0	0.5	0		0.1		0	0.1		0.2
y_4	Age class 0	1.7		0.5		0.8	0.9	0.7	1.4	3.3	1.6
	Age class 1	0.4		0		0.6	0.8	0.2	0.6	2.3	1.2
	Age class 2	0.1		0		0.2	0.1	0.3	0.2	0.2	0.3
y_5	Age class 0	1.6	4.0	0.3	3.7	1.0	0.5	0.5	0.9	3.0	1.5
	Age class 1	0.4	1.2	0	1.4	0.7	0.4	0.3	0.6	2.5	1.2
	Age class 2	0.1	0.6	0	0.2	0.2	0.1	0.1	0.2	0.5	0.2

(b)

	H = poor	H = good
P_{y_1}	0.67	0.33
P_{y_2}	0.6	0.4
P_{y_3}	0.86	0.14
P_{y_4}	0.88	0.12
P_{y_5}	0.7	0.3

(c)

possible ways of picking site samples of the same size as that of year y_i, so a brute-force search is impracticable for quite small sample sizes already. For example, let us assume a river has 50 habitat-classified areas. Let us say we wish to pick 20% of the sites for electrofishing. If we assume that 10 of the sites will never be chosen ($|S_e| = 40$ and $|S_{y_i}| = 10$), from (6) we see that there are 847,660,528 ways of choosing 10 sites, even though 50 is quite a small number of sites. By comparison, river Tornio has 565 sites.

Example 4.2. Let the observed data be as in Example 4.1. We wish to see whether we could have made a more representative choice of sites in years y_1, y_2, y_3 and y_4. (Note that in this case all sites are eligible, i.e. $S_e = S$.) Going over all possible sample sizes, we can calculate a way of picking sites that minimizes our bias for each size. See Table 3 for one particular series of optimal choices and Fig. 4(b) for a comparison against the choices made in Example 4.1. We can see now that even though our bias was non-zero at years y_1 and y_2, we could not have done any better with samples of those sizes.

Our measure can thus be used as a planning aid: given a subset of sites/habitat types to choose from and a selection percentage, our methodology can offer a set of bias-minimal selections of sites. A person planning to electrofish in a river can also pick a set of sites freely, and see how far from the optimum her selection lies.

Naturally, the concept of bias-minimality is conditional on the particular habitat assignment, biological knowledge and empirical model adopted, but note how we only require that the type assignment defines a probabilistic sample space and the models are probability mass functions.

Our tool can also be used to study different habitat hypotheses about a river: it allows one to see whether site sampling has been representative of the hypothesis under consideration. Finally, dropping the time-invariance of the types assigned to sites would require no modification to the measure; all it would entail is that the knowledge $P(H)$ would differ from one year to another.

4.2. Studying the interplay of habitat and abundance

In Chapter 2 we defined our model for the joint distribution of abundance and habitat as $P(H, A) = P(H)P(A \mid H)$. To inspect the interplay of habitat and abundance in a new set of observations, we study $D(P_{y_1,\ldots,y_i}(H, A) \parallel P(H, A))$, where $P(H, A)$ is our biological knowledge, expressed as the joint distribution of habitat and abundance, and $P_{y_1,\ldots,y_i}(H, A)$ is our cumulative empirical model, i.e. the empirical joint distribution at year y_i, based on the data up to and including year y_i, describing a new data set.

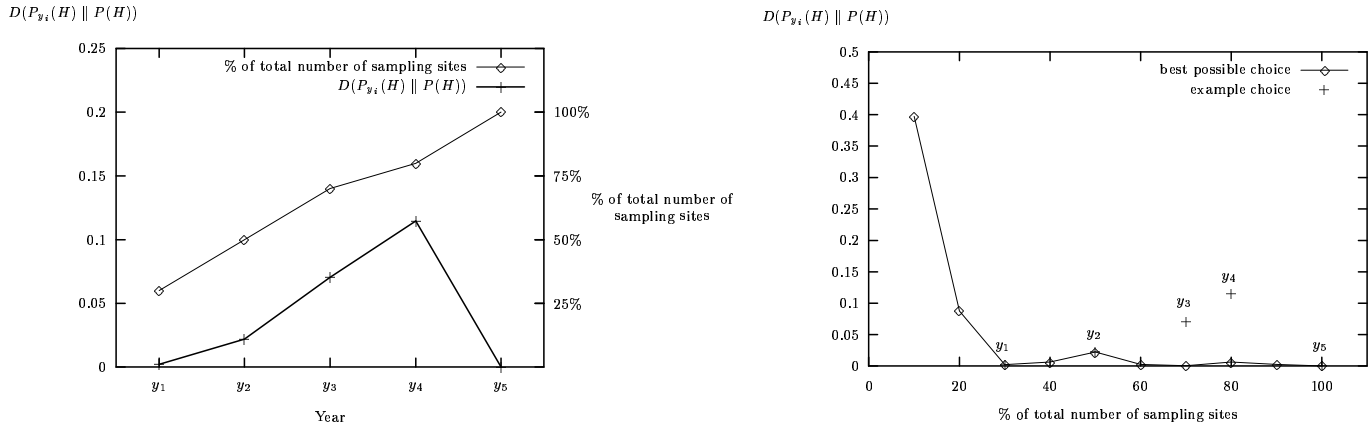
Using (5), the joint distribution divergence can be decomposed as follows:

$$\begin{aligned} &D(P_{y_1,\ldots,y_i}(H, A) \parallel P(H, A)) = \\ &\qquad D(P_{y_1,\ldots,y_i}(H) \parallel P(H)) \\ &\qquad + D(P_{y_1,\ldots,y_i}(A \mid H) \parallel P(A \mid H)). \end{aligned} \tag{7}$$

That is, the cost of assuming that our pre-existing knowledge $P(H, A)$ describes well the new observations up to year y_i can be seen as the sum of two costs: the cost of assuming our habitat sampling has been representative and the cost of assuming our knowledge and the data agree on the relationship of habitat and abundance. If the divergence of our joint distributions changes at some point in time, we can see whether this is due to a change in the choice of habitat types chosen for electrofishing, to a change in the relationship of habitat and abundance in the data, or to a change in both.

Example 4.3. Let H be the habitat type variable of Example 3.1. We have electrofishing data for a time series of 5 years from the electrofishing sites of Example 4.1. Our biological knowledge about the sites' habitats, $P(H)$, is as in Example 4.1, and the habitat assignments of sites are also the same (see

Fig. 4. (a) Example 4.1. Time-variance of habitat sampling bias compared to percentage of sites sampled. (b) Example 4.2. Best-case bias vs. actual choices



(a) (b)

Table 3. A set of possible bias-minimal ways of choosing a given number of sites with the habitat types of Example 4.2. "*" signifies inclusion.

% of total number of sites	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
10	*									
20	*	*								
30	*	*	*							
40	*	*	*		*					
50	*	*	*	*	*					
60	*	*	*	*	*	*				
70	*	*	*	*	*	*	*			
80	*	*	*	*	*	*	*	*		
90	*	*	*	*	*	*	*	*	*	
100	*	*	*	*	*	*	*	*	*	*

Table 2(a)). We have a different series of observations, however. Table 4(a) shows the yearly statistics of habitat observations. From the table we can calculate e.g. that $P_{y_1,\ldots,y_4}(H = good) = 0.6875$.

The observations of abundances given a habitat are shown in Table 4(b), from which we can calculate e.g. that for age class 0, $P_{y_1,\ldots,y_4}(A = scarce \mid H = good) = 0.25$. $P(A \mid H)$, our pre-existing knowledge about the distribution of abundance given a habitat, is shown in Table 4(c).

We can now calculate $D(P_{y_1,\ldots,y_i}(H, A) \parallel P(H, A))$ as described above, shown in Fig. 5(a) for all age classes . The decomposition to $D(P_{y_1,\ldots,y_i}(H) \parallel P(H))$ and $D(P_{y_1,\ldots,y_i}(A \mid H) \parallel P(A \mid H))$ is shown in figures Fig. 5(b) and Fig. 5(c).

Studying Fig. 5 we see that age 2+ data seem to diverge from our pre-existing biological knowledge the most, and none of the data converge to our knowledge. Turning now to constituent parts, the figure indicates that habitat sampling bias has increased from year y_2 onwards. Going back to the data, we see that this is because sites of good type have been overrepresented from then on. Looking at the habitat–abundance relationship divergence we see that age 0+ data actually come close to our knowledge from year y_4 onwards, whereas the data for age 1+ fish exhibit a steady habitat–abundance relationship, which however disagrees with our pre-existing biological knowledge. And finally, the abundance of 2+ fish in the data (given a habitat type) seems to diverge clearly from our knowledge beginning with year y_4, regardless of the increased habitat sampling bias.

When interpreting the results, an important thing to keep in mind is that the conditional distribution $P_{y_1,\ldots,y_i}(A \mid H)$ has $|\mathcal{A}| \cdot |\mathcal{H}|$ parameters, compared to $|\mathcal{H}|$ for $P_{y_1,\ldots,y_i}(H)$, i.e. there are more parameters to be estimated from the same amount of data. Accordingly, if habitat sampling seems to stabilize, but the relationship of habitat and abundance still fluctuates, this might be due to several reasons, if we only have a short time series of data. It might provide evidence of some factor besides our habitat system affecting abundance, or our habitat type system might classify sites non-optimally with respect to abundance (i.e. some habitat types in our system might differentiate levels of abundance poorly). These two cases can be studied by changing either the model structure or the type system. However, in the case of a short time series, it could also be that $P_{y_1,\ldots,y_i}(A \mid H)$ is complex enough to require more data

Table 4. Example 4.3. (a) Numbers of observed habitats per type for each year. (b) Observed abundances given habitat for each year. (c) Pre-existing biological knowledge about the relationship of habitat and abundance, assumed to be equal for all age classes.

	H = poor	H = good
y_1	2	3
y_2	1	2
y_3	1	3
y_4	1	3
y_5	1	3

(a)

Year		H = poor	H = good
y_1	Age class 0	scarce	abundant
	Age class 1	scarce	abundant
	Age class 2	scarce	abundant
y_2	Age class 0	scarce	abundant
	Age class 1	scarce	abundant
	Age class 2	scarce	abundant
y_3	Age class 0	scarce	abundant
	Age class 1	scarce	abundant
	Age class 2	abundant	scarce
y_4	Age class 0	abundant	scarce
	Age class 1	scarce	abundant
	Age class 2	abundant	scarce
y_5	Age class 0	scarce	abundant
	Age class 1	scarce	abundant
	Age class 2	abundant	scarce

(b)

$P(A \mid H)$		
	H = poor	H = good
A = scarce	0.9	0.25
A = abundant	0.1	0.75

(c)

to capture the shape of the distribution well enough. Nonetheless, this is a problem common to all methods of comparing models constructed from real-world data to a given model, so we lose nothing by adopting our methodology.

4.3. Studying the interplay of habitat and abundance without pre-existing biological knowledge

In practical modeling, the pre-existing biological knowledge might well be lacking, or we might intentionally want to incorporate biological knowledge only in the habitat classification phase, preferring to let the data decide on the habitat–abundance relationship. In this case, we have no $P(H, A)$ to compare the empirical model to.

To overcome this, we adopt the following technique. At each time step y_i we take the empirical model based on data collected so far to be the accumulated biological knowledge at that moment in time. We then measure the distance to the corresponding empirical model at the previous moment in time.

Formally put, we study

$$D(P_{y_1,\ldots,y_i}(H, A) \parallel P_{y_1,\ldots,y_{i-1}}(H, A)).$$

This procedure measures the convergence of our empirical model: as time unwinds, successive models should come close to each other, if our joint distribution is time-invariant. Naturally, in the case of convergence it does not follow that our $P_{y_1,\ldots,y_i}(H, A)$ would now be a "true" distribution. It only indicates that our habitat sampling bias has stayed constant and the relationship of habitat and abundance has stabilized. Note that this way of measuring reflects reality in several fisheries problems: we as observers are always located at a point y_j in time, having at our disposal the data collected up to that point, wanting to predict for year y_{j+1}. The divergence $D(P_{y_1,\ldots,y_i}(H, A) \parallel P_{y_1,\ldots,y_{i-1}}(H, A))$ measures in a sense the amount of information about the joint distribution that we would have gained at year y_{j+1}, had we added the measurement of that year to our data set. Our measure thus shows the momentary changes in the empirical distribution over a time series. (In other words, the series of divergences describe the learning process in a field study.)

Example 4.4. Let H be the habitat type variable of Example 3.1 once more. We have the same set of observations as in Example 4.3, shown in tables Table 4(a) and Table 4(b). This time we lack the biological knowledge $P(H, A)$, however.

We can now calculate $D(P_{y_1,\ldots,y_i}(H, A) \parallel P_{y_1,\ldots,y_{i-1}}(H, A))$ shown in Fig. 6(a) for all age classes.

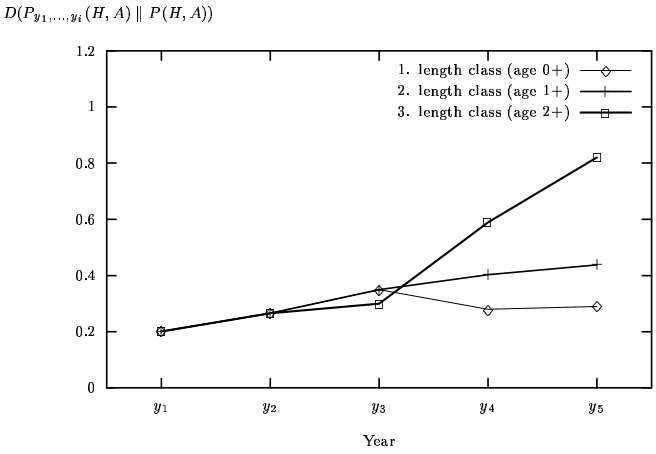
The decomposition to $D(P_{y_1,\ldots,y_i}(H) \parallel P_{y_1,\ldots,y_{i-1}}(H))$ and $D(P_{y_1,\ldots,y_i}(A \mid H) \parallel P_{y_1,\ldots,y_{i-1}}(A \mid H))$ is shown in figures Fig. 6(b) and Fig. 6(c).

The graphs show that our habitat sampling process fluctuates slightly at first, converging after year y_4. Overall, habitat sampling has a negligible effect on joint distribution divergence. Looking at the habitat–abundance relationship, we see how 0+ data have a quirk at year y_4, apparently returning to convergence at y_5. By comparison, 2+ data, which have a consistent change from y_3 on, have converged the least by y_5, whereas 1+ data, which stay the same across the time series, converge rapidly.

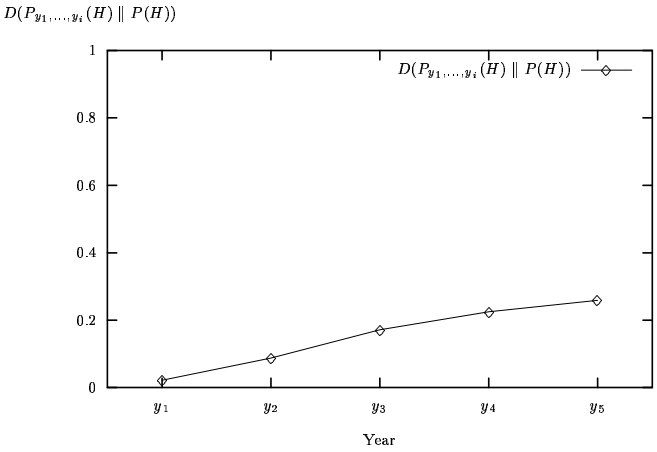
The technique we exhibited above is only one possible way of tackling the problem of having no pre-existing biological knowledge. When interpreting the results, it is helpful to keep in mind that $P_{y_1,\ldots,y_i}$ always has the same data as $P_{y_1,\ldots,y_{i-1}}$, added with the data of one year. Hence, as time goes on, the potential for difference gets smaller, depending somewhat on the method of learning the empirical models from the data. The absolute divergences at different points in time can thus not be used as absolute, directly comparable measures of change in the distribution. As time goes on, successive divergences come increasingly comparable, however.

One way of avoiding this characteristic would be to use $D(P_{y_{i-2T+1},\ldots,y_{i-T}}(H, A) \parallel P_{y_{i-T+1},\ldots,y_i}(H, A))$. In this case

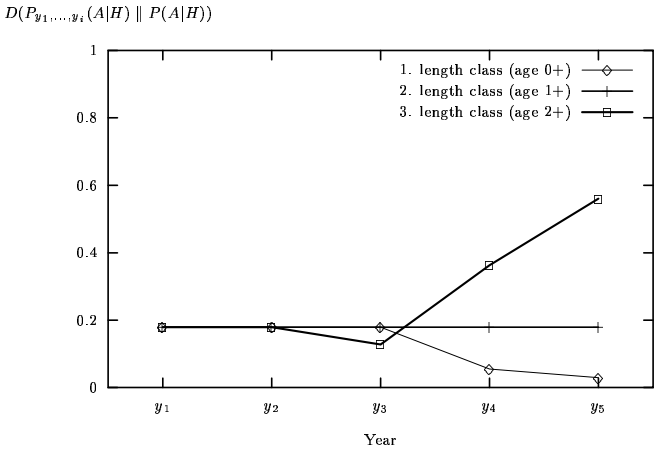
Fig. 5. Example 4.3. (a) Variance of $D(P_{y_1,\ldots,y_i}(H,A) \parallel P(H,A))$ over time for each age class. (b) Variance of habitat sampling bias over time. (c) Variance of abundance–habitat relationship for each age class.



(a)



(b)



(c)

we assume that a time series of T years suffices to capture the empirical distribution. We then compare the empirical distribution of the last T years to the T years in time that preceded it. These two subsets of data are non-overlapping and of the same size. Most importantly, any measurements made using this measure at different time points y_i and y_j are comparable as absolute values. Naturally, this type of a measure can only capture completely changes occurring within a time frame of $2T$ years. This can be remedied by picking $T = \lfloor \frac{i}{2} \rfloor$ at each instant of time y_i. If this is done, different points in time are not exactly comparable, but the subsets of data compared at each point in time are. In this work we will not study these techniques any further, however, because they require a time series of such length that T can reasonably be expected to suffice for learning the empirical joint distribution model. Our real-world data only has time series of lengths 15 and 17, making these techniques inapplicable.

5. Real-world data sets

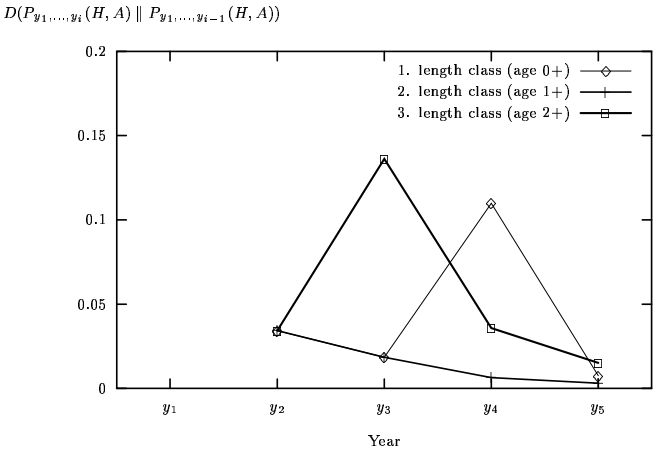
Let us now exhibit the nature of our data sets in some detail. We use habitat and electrofishing data from two Gulf of Bothnia salmon rivers, Simo and Tornio (the Finnish side). In this work only data on wild salmon are included.

5.1. Habitat data

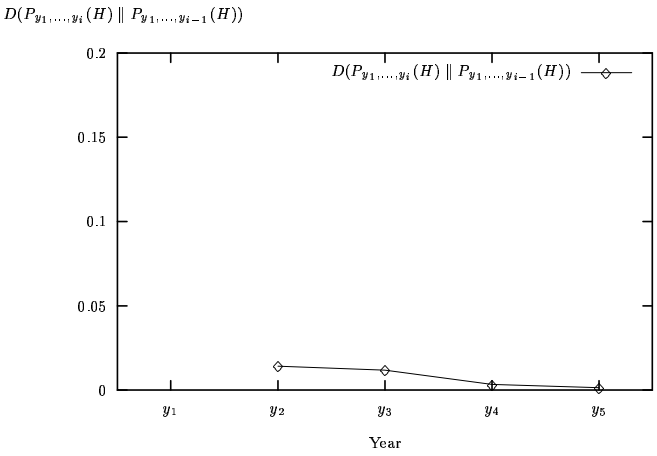
Table 5 shows our set of habitat variables. The variables are naturally grouped in the sense that e.g. all variables in the *Bottom* column are percentages which sum up to 100% for each particular site. To demonstrate the two-layered structure more clearly, Table 6 shows a hypothetical set of habitat data for variable groups *Current* and *Depth*.

In the following analysis, a variable set name such as *Depth* should be understood as the variable set {*Depth:* $<$ *20 cm*, *Depth: 20 - 50 cm*, *Depth: 50 - 100 cm*, *Depth:* $>$ *100 cm*}. Fig. 7 shows how the classification accuracy of electrofishing

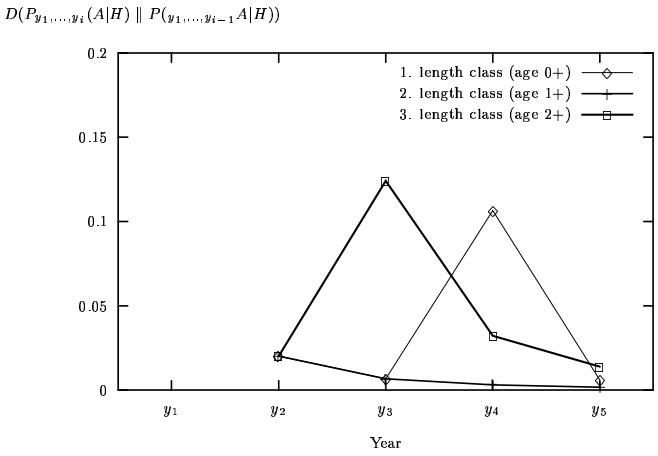
Fig. 6. Example 4.4. (a) Momentary changes in the empirical joint distribution for each age class. (b) Momentary changes in habitat sampling bias. (c) Momentary changes in abundance–habitat relationship for each age class.



(a)



(b)



(c)

sites differs significantly for the rivers: the data for Simo have a much coarser resolution.

5.2. Electrofishing data

In our data the electrofishing sites are subareas of the areas described by the habitat data. For each year in our time series we have electrofishing data from a subset of the total set of sites. The sampled sites and their number vary over time, especially in the beginning of our time series. See Fig. 8 for the yearly locations of electrofishing sites (to the accuracy of the location of the covering habitat area).

Each sampling site has been habitat-classified by domain experts as described in Chapter 5.1. This classification is "timeless", since these classifications are assumed to stay the same during all of our available time series.

The data sets contain electrofishing data on three levels:

1. The low level, where each record describes a single individual fish caught by electrofishing.
2. The intermediate level, where each record describes a single fishing run. That is, as electrofishing is carried out in 1 - 3 separate runs, we have a record for each of the individual runs at a specific site.
3. The high level, where each record summarizes the electrofishing data for an entire river for a single year. This data is irrelevant to this work, since individual sites (and thus their types) cannot be told apart.

The lowest level is not directly useful for the main problem here, since we are not interested in modeling a single fish. On the other hand, this type of data contains exact measurements such as length and weight instead of estimates. It also contains a relatively large number of samples (many thousands). An important observation is that this data is highly valuable

Table 5. An overview of the habitat variables in the data set. Each column consists of an interconnected group of variables.

Variable groups		
Bottom	Current	Depth
Bottom: Sand/mud/clay	Current: Stillwater pool	Depth: $<$ 20 cm
Bottom: Gravel, $<$ 2 cm	Current: Pool, visible current	Depth: 20 - 50 cm
Bottom: Stones, 2 - 10 cm	Current: Riffle	Depth: 50 - 100 cm
Bottom: Stones, 10 - 30 cm	Current: Rapid	Depth: $>$ 100 cm
Bottom: Boulders, $>$ 30 cm	Current: Strong rapid	
Bottom: Bedrock		

Table 6. An example of habitat data, groups *Current* and *Depth* only. Each row is a record of data describing a site, with the values within a variable group summing up to 100.

Current					Depth			
Stillwater pool	Pool, visible current	Riffle	Rapid	Strong rapid	$<$ 20 cm	20 - 50 cm	50 - 100 cm	$>$ 100 cm
0	20	30	50	0	0	0	100	0
25	50	25	0	0	0	0	40	60
65	30	5	0	0	0	60	30	10
100	0	0	0	0	90	10	0	0

in the sense that it can be used to classify fish based on their length.

At the intermediate level, a slight complication enters. During electrofishing, usually more than one fishing run is performed. However, the overall number of such runs, performed consecutively at the same site on the same day, varies. The most common number of runs is three, but sometimes there are fewer runs. Thus, we chose to always use the first run only, to have comparable data for all sites, rivers and years.

In fact, the data at the intermediate level are just a summing up of the lowest-level data, augmented by data on fishing runs that caught no fish. Therefore we created our own version of intermediate (fishing run) level data directly from the low-level data, adding to the result the unsuccessful fishing runs to avoid positive bias.

As the aim is to model the abundance of each age group separately, we need an age-classification system for our empirical models. Ready-made ages are provided in the data, but for part of the data the information is missing. Hence, we chose to classify the fish according to their length, using 7 and 11 cm as the split points. This provided us with a significantly higher amount of data, so we deemed it worthwhile. It also removed any bias that the ready-made age-classification might introduce to the system. These particular split points were determined by domain experts. To see how they correspond to the empirical length distributions of age-classified fish in our data sets, see Fig. 9. Note that the plot for river Tornio also shows how under-represented 0+ fish are in the aged subset of data for river Tornio, due to missing age labels for small fish.

Naturally, this length-class system does not correspond exactly to an age-class system: fish of the same length might have different ages, as Fig. 9 shows. Also, it is known that salmon grow faster and smoltify earlier in warmer environments, so a more southern river can have younger fish at a given length. But this is actually one of the advantages of our system: if we assume that maturity depends on size (which depends on age), it is reasonable to classify fish based on size, making fish of different ages but similar sizes (and thus presumably at the same stage of maturity) more comparable. Weight could in principle used as an alternative indicator of size as well, but the data was often missing, whereas length never was, so we chose to employ length only.

5.3. Categorization of the data

We will now describe the way we chose to categorize the available real-world data in our empirical modeling.

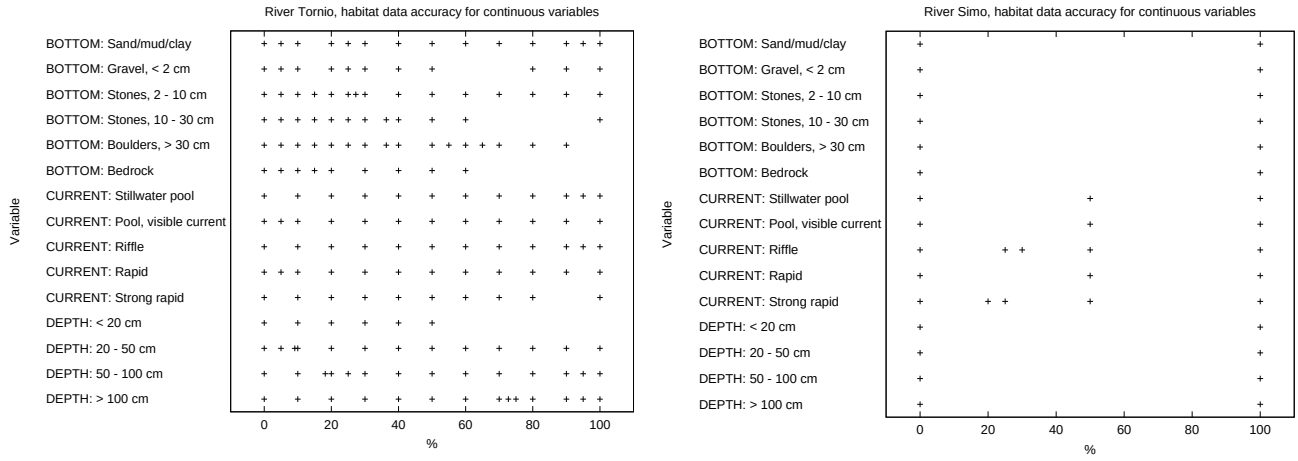
5.3.1. Habitat type systems

Recall from Chapter 5.1 that in our real-world data the electrofishing sites are described by a set of habitat variables V. Our methodology requires a type system. i.e. a disjoint and exhaustive partitioning of the space of all possible data vectors induced by V. In other words, we need a discrete classification of our set of sites.

A biologist or a fishery scientist might well want to define the type system on the basis of biological knowledge. In our empirical work here, we had no such given system available, and thus devised as unbiased a system as possible. This is not a feature of our methodology, however: any disjoint and exhaustive partitioning of the habitat data space will do. Our aim here was to find one which did not add any unwelcome bias to the system from the beginning. It should serve as a crude point of comparison. Remember also that from the modeling accuracy point of view, the essential thing here would be to habitat-classify the electrofishing sites themselves, due to their being such a tiny portion of the currently habitat-classified areas (see Fig. 1). Our current classification is probably quite inaccurate to start with.

Recall from Chapter 5.1 that our habitat data actually consists of two layers. Our set of variables is a set of n variable sets, $V = \{V_1, \ldots, V_n\}$, where each V_i is a set of n_i variables $\{v_{i1}, \ldots, v_{in_i}\}$. The values of the members of V_i sum up to 100 in each record (see Table 6).

Fig. 7. Habitat data accuracy. Each point denotes at least one occurrence of that particular value in the data. (a) River Tornio. (b) River Simo.



(a) (b)

As an example of an uninformed type system, we chose to employ equal-width discretization of our continuous habitat variables to K_h bins, independently of each other and the data. That is, if K_h is 3, values in the range $[0..33]$ get assigned to value 0, values in the range $(33..67]$ to value 1, and values in the range $(67..100]$ to value 2 for each variable v_{ij}.

Let us call H_i the random variable devised in this manner, consisting of a group of variables summing up to 100 and denoting a particular aspect of the habitat type of a site, e.g. depth. H_i is a vector-valued random variable taking as its values our discretized classifications, i.e. the value set of H_i is $\{0, 1, \ldots, K_h-1\}^{|V_i|}$. Because of the requirement of summing up to a constant, the possible values of H_i lie within a $|V_i|-1$-dimensional simplex in the $|V_i|$-dimensional space induced by V_i. Our discretization divides the simplex into ${K_h}^{|V_i|-1}$ regions of equal size. The following example illustrates this.

Example 5.1. Let us assume $V = \{V_1, V_2\}$. $V_1 = \{$Depth: $<$ 50 cm, Depth: 50 - 100 cm, Depth: $>$ 100 cm$\}$, $V_2 = \{$Current: pool, Current: riffle, Current: rapid$\}$. Each $v_{ij} \in V_i$ is a continuous variable with range $[0..100]$, and $\sum_j v_{ij} = 100$ in each record. Table 7(a) shows the classifications of three sites by experts using V. We choose $K_h = 2$. Our type system produces the classifications shown in Table 7(b).

Now, as H is vector-valued ($H = [H_1, H_2]$), we see that H_2 does not differentiate any of the three sites. Due to differences in the values of H_1, however, s_i and s_j get mapped to a different habitat type (since their discretized versions differ), whereas s_i and s_k are considered to be of the same habitat type, since they both have discretized value $[0, 0, 0]$ for H_1, even though their original classifications differ somewhat.

Fig. 10 shows the situation in visual terms.

Our goal, a single habitat type variable, is thus a vector-valued variable H with value set

$$[H_1, \ldots, H_n] \in \{0, 1, \ldots, K_h - 1\}^m,$$

where $m = \sum_i |V_i|$.

Even though we have not used any biological knowledge in the determination of the subregions of the simplex, the ones defined by our type system clearly ought to catch some of the characteristics of habitat, keeping in mind the semantics of our habitat variables: the variables within each group are more or less ordered (going from shallow to deep water in the case of *Depth* for example).

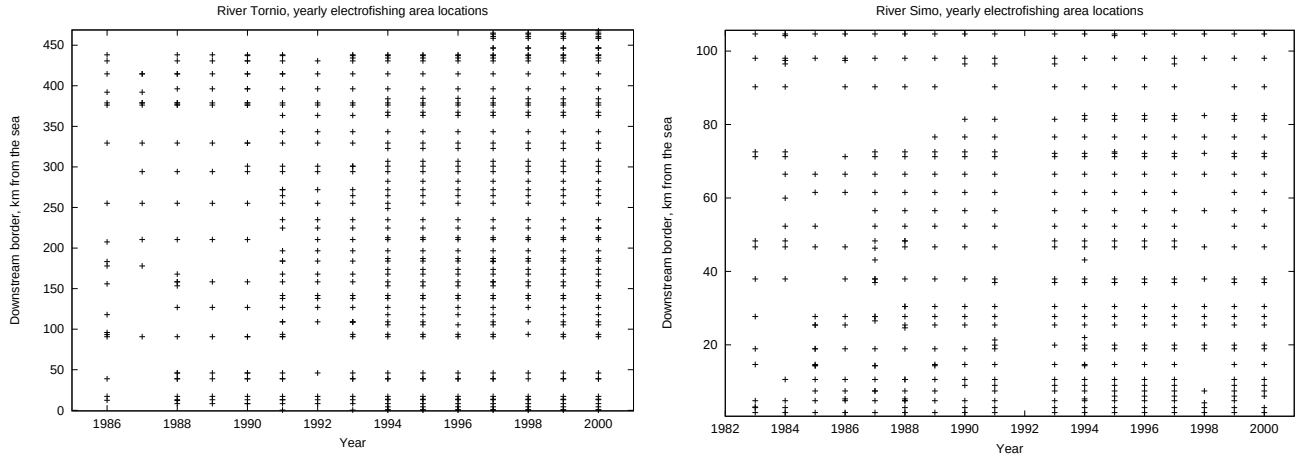
5.3.2. Obtaining abundances from density observations

At year y_i we have electrofishing data from S_{y_i}, a set of electrofishing sites. Each site $s \in S_{y_i}$ has been assigned a habitat type h_j. To describe the absolute observed densities at sites of type h_j during a year we use the average of the measured absolute densities for sites of type h_j. We might lose some information here, but we justify this by our assumptions: we do assume that density is strongly dependent on habitat type. If this basic assumption holds, the average should be a good estimate. We now have $\overline{d}_{ij}$, an estimate of the absolute density at year y_i given habitat type, for each habitat type h_j. In order to translate these into abundances, we equal-width discretize the range of $\overline{d}_{ij}$ to K_a bins. We then learn the empirical model from the data set consisting of these discretized values.

Remark 5.2. A more sophisticated abundance model could be introduced as well, adding component O, the observed densities, making our model $P(H)P(O|H)P(A|O,H)$. This would rid us of the average-taking process, but introduce the conditional distribution $P(A|O,H)$, the probability of abundance of a given age class given the observations of densities at a number of sites of the same habitat type. Without quite strong extra assumptions, this distribution cannot be learned from the *data*, however, (the data only has the observed densities, not the abundance) so we have not adopted this strategy here.

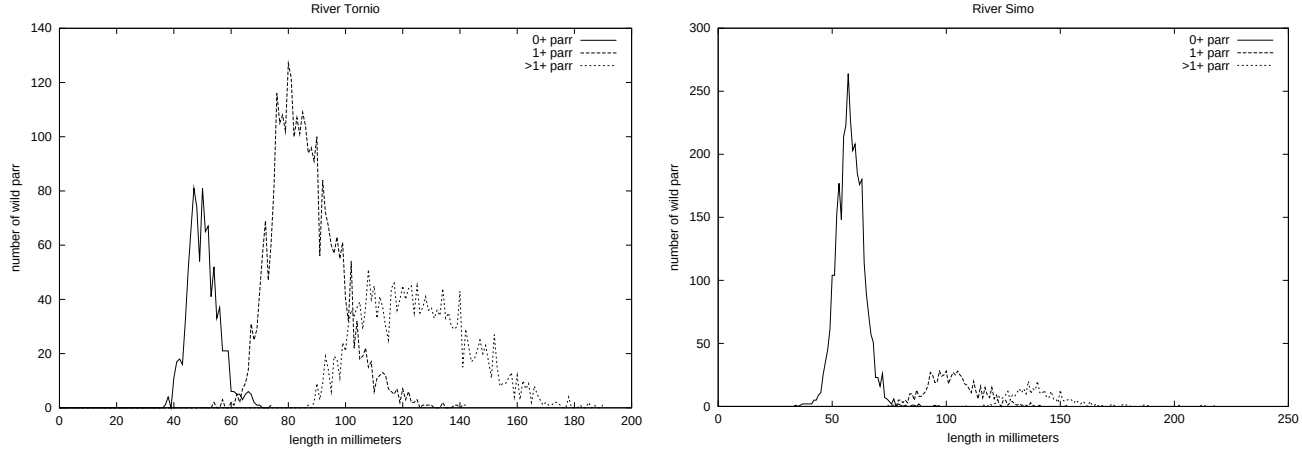
Example 5.3. Let H be the habitat type variable of Example 3.1. We have electrofishing data for a time series of 5 years from the electrofishing sites of Example 4.1. Our biological knowledge about the sites' habitats, $P(H)$, is as in Example 4.1,

Fig. 8. The yearly locations of electrofished sites. (a) River Tornio. (b) River Simo.



(a) (b)

Fig. 9. The empirical length distributions of fish, aged in the data. (a) River Tornio. (b) River Simo.



(a) (b)

and the habitat assignments of sites are also the same (see Table 2(a)). The observed densities are shown in Table 8(a).

We now calculate the average observed densities per habitat type for each year, proceeding to transform these to discretized relative densities per year. We choose $K_a = 2$ to match the two-valued abundance variable A of previous examples, obtaining time series data on abundances given a habitat type, shown in Table 8(b).

6. Empirical results

We will now describe empirical results obtained using real-world data from two Gulf of Bothnia salmon rivers: Simo and Tornio (the Finnish side). The techniques of comparing empirical models to biological knowledge have been described in Chapter 3 and illustrated in Chapter 4. Since we did not have available a joint distribution model $P(H, A)$ describing biological knowledge, we used the techniques shown in Chapter 4.3 to study the cumulative learning process of the joint distribution from observations. For the analysis of habitat sampling bias, we chose the global distribution of habitats in the river as our biological knowledge $P(H)$.

Our empirical models were built as described in Chapter 3.2. We categorized the data using the methods of Chapter 5.3, picking $K_h = 3$ and $K_a = 3$, i.e. three categories for abundance and each habitat variable. We aimed at as small a number of categories as possible, to avoid overfitting to our relatively small amount of data, while still allowing for meaningful qualitative analysis.

In addition to measuring the optimality of yearly choices, we compared against several different restrictions (in the order of increasing restrictiveness):

1. No restrictions on selection (except the number of sites).

Table 7. Example 5.1. (a) Three sites classified by experts with respect to V. (b) H using $K_h = 2$, i.e. V discretized using 2 bins.

Site	V					
	V_1			V_2		
	Depth: < 50 cm	Depth: 50 - 100 cm	Depth: > 100cm	Current: pool	Current: riffle	Current: rapid
s_i	20	35	45	20	80	0
s_j	5	65	30	15	75	0
s_k	40	45	15	10	85	5

(a)

Site	H					
	H_1			H_2		
	Depth: < 50 cm	Depth: 50 - 100 cm	Depth: > 100cm	Current: pool	Current: riffle	Current: rapid
s_i	0	0	0	0	1	0
s_j	0	1	0	0	1	0
s_k	0	0	0	0	1	0

(b)

2. Sites with a non-zero percentage of *Current: stillwater pool* are not eligible. This is an example of a real-world restriction.

3. Sites which have never been electrofished from during the entire time series are excluded. This is the maximal constant restriction imposable on the data.

If a constant restriction on the choice of sites actually exists for a river, we would expect it to lie somewhere between restrictions 2 and 3.

6.1. River Tornio, Finnish side

There are 565 habitat-classified sites in river Tornio. Our electrofishing data contains measurements from a period of 15 years. Fig. 11(a) shows the habitat representativeness of each yearly selection of electrofished sites on the Finnish side of river Tornio. You can see that for river Tornio the representativeness has increased over time, although not essentially since 1994. Up to that year the increase in representativeness also correlates positively with the number of sampling sites, but after that the slight increase in the number of sampling sites does not really affect the representativeness any more. Measuring the different aspects of habitat separately we see as shown in Fig. 11(b) that *Current* has a trend towards higher representativeness up to 1991, but after that representativeness actually seems to have a slightly decreasing tendency. *Bottom* displays quite smooth and slow convergence to a steady bias. *Depth* seems to fluctuate the most in the beginning, never displaying much of a trend.

Fig. 11(c) shows the best-case representativeness for all possible yearly site sample sizes, compared to the actual choices committed in the data. It can be seen that if there were no restrictions, something like 25% of sites would have to be electrofished to have very good representativeness, but something like 10%, which is already a percentage occurring in the data, suffices for good representativeness. The actual choices made lie quite far from the non-pool restriction, however, so either sampling has been quite non-optimal or, what is more likely, there exists a constant stricter restriction constraining sampling.

Studying the convergence of the empirical distribution of electrofished habitat types in Fig. 11(d), it seems like habitat sampling has converged to a constant bias fairly well, excepting 1994 and 1997.

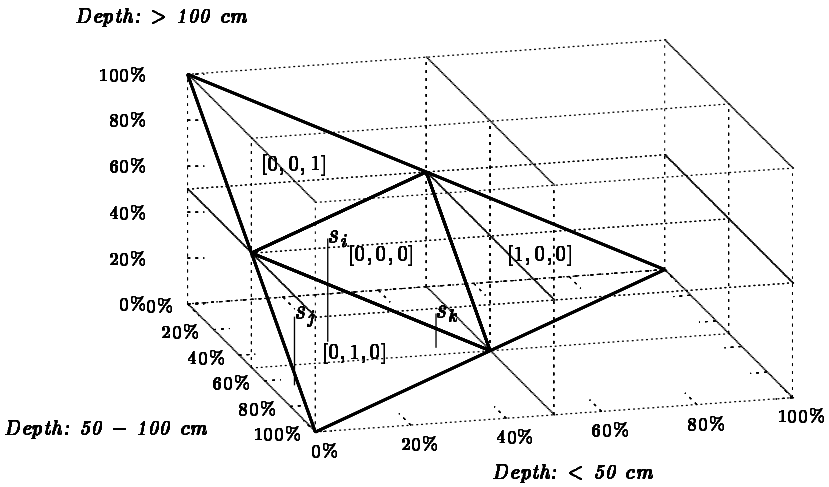
The convergence of the empirical joint distribution is shown in Fig. 12. The difference of 2+ fish from the other age classes is quite clear. The 0+ fish have a noticeable change in 1994, but by the end of the time series their distribution has converged the most. The 1+ fish had a change in 1996 and seem to have been experiencing some change from 1998 onwards. Finally, after changing in 1993, 2+ fish seem to have had a temporary stable period up to 1997, when another change enters.

6.2. River Simo

There are 377 habitat-classified sites in river Simo. Our electrofishing data contains measurements from 17 years (due to flooding, no measurements were possible in 1992). Fig. 13(a) shows the habitat representativeness of each yearly selection of electrofished sites in Simo. You can see that the representativeness has not really changed over time, regardless of the fluctuation in the sampling percentage, which in 1994-1997 doubled from that in 1983. The drop in percentage from the previous year in 1998 by almost a half affected representativeness only slightly. This is most likely the outcome of the coarse resolution of the original habitat classifications of river Simo (see Fig. 7): the type system is too crude, failing to differentiate the sites sufficiently.

Measuring the different aspects of habitat separately as in Fig. 13(b) we see that the representativeness with respect to granularity of bottom seems to have decreased from 1983 until 1990, after which all factors seem to have reached constant bias. Different current types also seem to have had varying representativeness up to 1990, whereas *Depth* has had a constant bias all along. An interesting observation is that when the

Fig. 10. H_1 of Example 5.1. The original values of V_1 lie on the plane whose outlines are shown in bold. The triangular subregions of this plane are labeled with the corresponding value of H_1.



number of sites sampled dropped sharply in 1998, the representativeness of bottom granularity and current type actually increased slightly.

Fig. 13(c) shows the minimal bias for all possible yearly site sample sizes compared to the actual selections made in the data. Something like 5% would already suffice for good habitat sampling, well in the range of actual choices. As with river Tornio, the choices made over the time series are quite far from the non-pool restriction, hinting at a constant restriction close to the "selected at least once" restriction.

Studying the convergence of the empirical distribution of electrofished habitat types in Fig. 13(d), we see that habitat sampling seems to have stabilized fully by 1990.

Empirical joint distribution convergence is shown in Fig. 14. It seems like the habitat-abundance relationship had temporarily stabilized by 1991, but 1992 has no measurements and when we come to 1993, something has changed in the relationship. The 0+ fish seem to stabilize from 1996 on, but it seems like in 2000 things are changing again. For 1+ fish 1994 and 1998 are notable, and 2000 shows a remarkable change. The 2+ fish stabilize in 1994, but have an abrupt change in 1995, and again in 1999.

7. Conclusions and future work

We have studied the problem of measuring electrofishing bias and the interplay of habitat and abundance by means of an information-theoretic methodology put forth in Chapter 3. We have illustrated the benefits of our approach, explaining in Chapter 4 in detail several techniques pertaining to different goals and levels of biological knowledge, using examples and discussing the interpretation of results.

We have also tested our methodology using real-world data. Chapter 6 provides an analysis of rivers Simo and Tornio. Habitat sampling bias and the relationship of habitat and abundance were studied both together and separately. Our results indicated a consistent, yet non-optimal electrofishing bias for both rivers. The fact that electrofishing has not been even nearly bias-minimal in these rivers can be explained by a set of strict non-explicit restrictions on the eligibility of sites. The empirical habitat-abundance relationships in these rivers were seen to be still in the process of changing periodically. The shortness of the available time series might be a factor here, however.

As described in Chapter 2, the traditional way of studying bias is via a simulator mimicking nature. This simulator can be seen in our approach as a case where the simulator defines $P(H, A)$. We can then study $D(P_{y_1,\ldots,y_i}(H, A)) \parallel P(H, A))$ as explained in this work. The simulator approach is thus simply a situation where the "true" joint distribution is known and encoded in the guise of a simulator.

A most important difference is that in our approach no artificial data set is needed, because our measure works on models directly. It studies differences in *distributions*, not differences in data sets. After all, $P(H, A)$ is all the information the simulator approach contains: the artificial data generated from the simulator reflect $P(H, A)$ in the limit of unrealistic data set sizes (real-world data have less than 20 years of data).

Our methodology only requires that the variables define a probabilistic sample space, and the models be probabilistic mass functions. Even though we chose to categorize our variables in this work, the methodology can be defined analogously for continuous variables.

Many of the implementations of our approach in this work could be refined and/or extended. The basic structural assumption could be made more complex. The empirical models could be constructed in different ways: criteria such as the MDL principle, or a predictive score, could be tried out instead of our simple approach in this work [5, 2]. Different habitat type systems and ways of deriving abundances from observational density data could be studied and tested as well. The study of the interplay of habitat and abundance would benefit from expert-given biological knowledge $P(H, A)$ in stead of our purely empirical modeling in this study.

In this work our methodology has been used for analysis of existing data sets. A tool aiding planning in addition to providing analyses could be built as well. This tool would give a fishery scientist a chance to try out different models encoding biological knowledge, seeing how they interact with the data. Also, the biological knowledge could be fixed, and the diver-

Table 8. Example 5.3. (a) The time series of observed densities. (b) Discretized relative densities per habitat type.

Year		s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
y_1	Age class 0	1.1	3.5	0.4	3.0	0.9					
	Age class 1	0.3	1.8	0.1	2.1	0.4					
	Age class 2	0.1	0.7	0	1.1	0.1					
y_2	Age class 0	1.8	6.2		5.0						
	Age class 1	0.4	3.2		2.4						
	Age class 2	0.1	0.3		0.5						
y_3	Age class 0	0.9	6.1		3.1					4.3	
	Age class 1	0.8	2.9		3.5					1.5	
	Age class 2	1.1	0.5		0.1					0.3	
y_4	Age class 0	4.2	3.2		3.4					3.6	
	Age class 1	0.4	2.3		2.2					1.2	
	Age class 2	0.6	0.3		0.1					0.1	
y_5	Age class 0	1.3	6.0		4.5					3.9	
	Age class 1	0.5	2.4		2.5					2.5	
	Age class 2	0.5	0.4		0.3					0.4	

(a)

Year		H = poor	H = good
y_1	Age class 0	scarce	abundant
	Age class 1	scarce	abundant
	Age class 2	scarce	abundant
y_2	Age class 0	scarce	abundant
	Age class 1	scarce	abundant
	Age class 2	scarce	abundant
y_3	Age class 0	scarce	abundant
	Age class 1	scarce	abundant
	Age class 2	abundant	scarce
y_4	Age class 0	abundant	scarce
	Age class 1	scarce	abundant
	Age class 2	abundant	scarce
y_5	Age class 0	scarce	abundant
	Age class 1	scarce	abundant
	Age class 2	abundant	scarce

(b)

gence of different empirical models from it measured. The optimality of suggested selections of sites could be measured by the tool as well, and bias-minimal selections suggested.

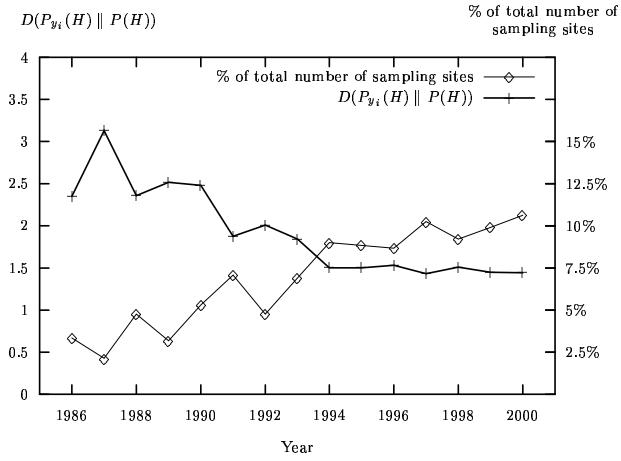
Acknowledgements

This paper has been funded by EU project nr 99/064 ”Probabilistic modelling of Baltic salmon stocks”.

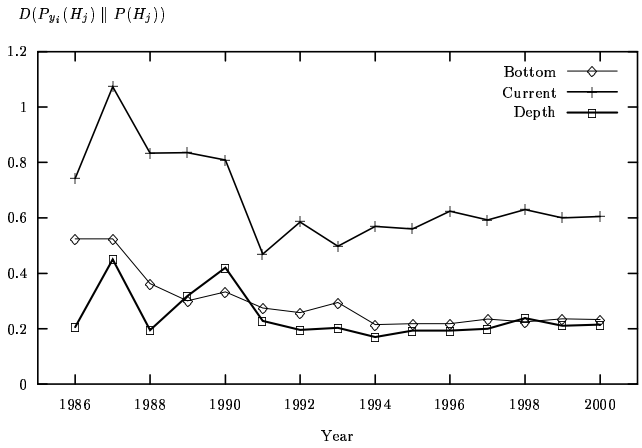
References

1. T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, NY, 1991.
2. P. Grünwald. *The Minimum Description Length Principle and Reasoning under Uncertainty*. PhD thesis, CWI, ILLC Dissertation Series 1998-03, 1998.
3. S. Kullback. *Information Theory and Statistics*. John Wiley & Sons, New York, NY, 1959.
4. J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company, New Jersey, 1989.
5. J. Rissanen. Hypothesis selection and testing by the MDL principle. *Computer Journal*, 42(4):260–269, 1999.

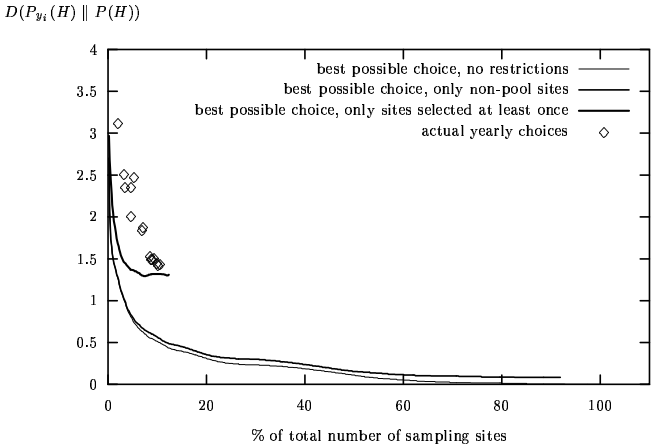
Fig. 11. Habitat sampling, river Tornio (Finnish side). (a) Yearly habitat sampling bias. (b) Yearly habitat sampling bias, different aspects of habitat studied separately. (c) Actual yearly site choices vs. bias-minimal choices for a site percentage and a given restriction on choices. (d) Momentary changes in habitat sampling bias, i.e. the convergence of the empirical distribution of electrofished habitat types.



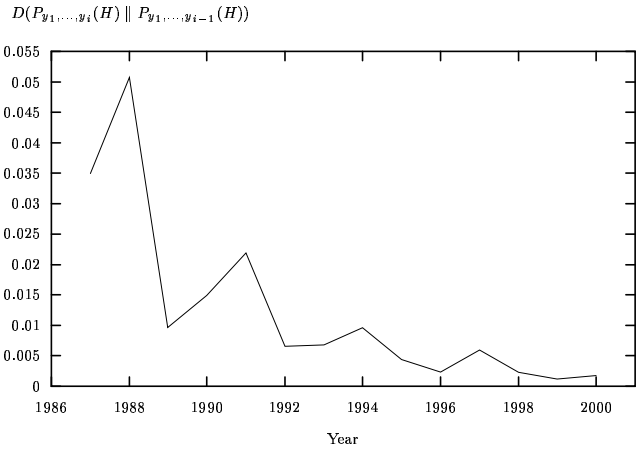
(a)



(b)

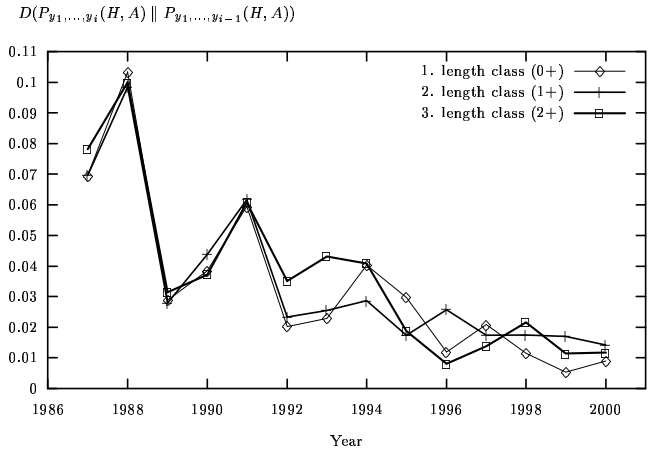


(c)

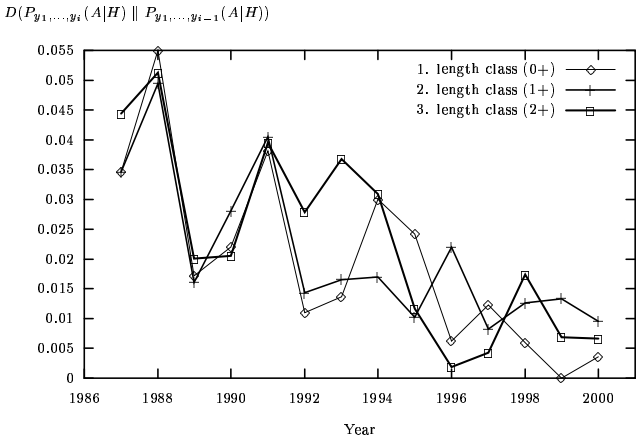


(d)

Fig. 12. Convergence of the empirical joint distribution, river Tornio, Finnish side. (a) Momentary changes in the empirical joint distribution. (b) Momentary changes in the abundance - habitat relationship.

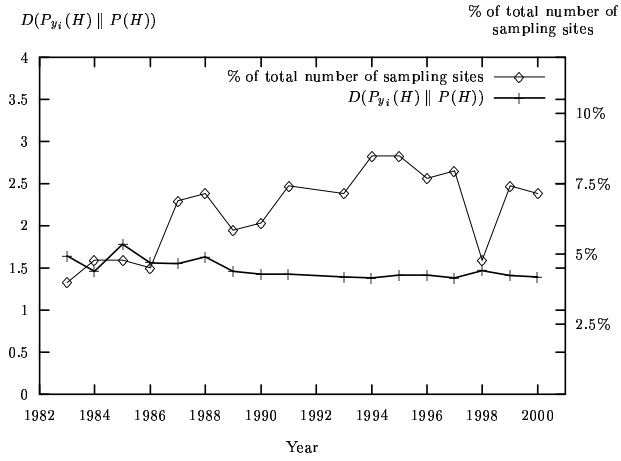


(a)

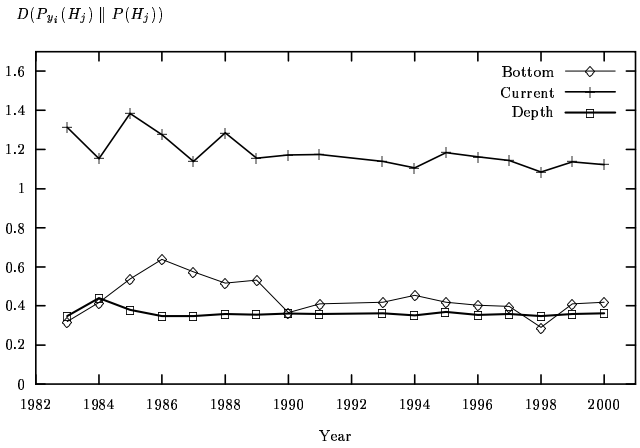


(b)

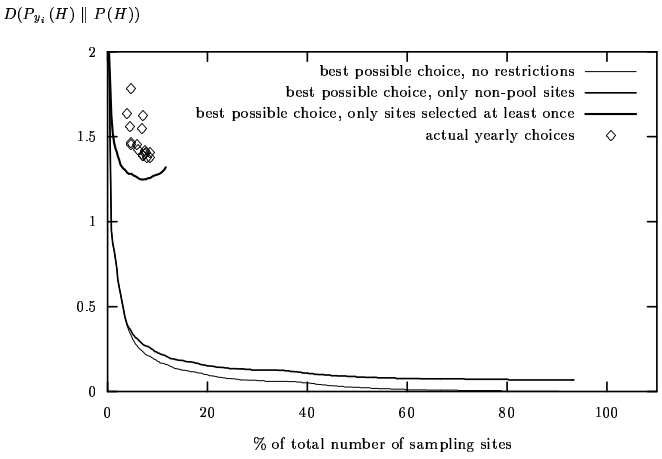
Fig. 13. Habitat sampling, river Simo. (a) Yearly habitat sampling bias. (b) Yearly habitat sampling bias, different aspects of habitat studied separately. (c) Actual yearly site choices vs. bias-minimal choices for a site percentage and a given restriction on choices. (d) Momentary changes in habitat sampling bias, i.e. the convergence of the empirical distribution of electrofished habitat types.



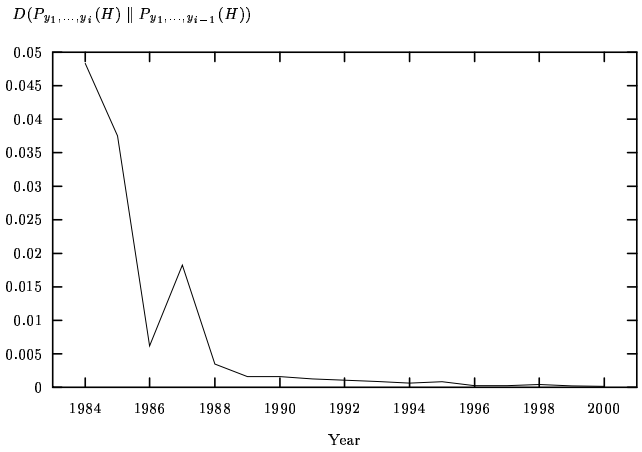
(a)



(b)

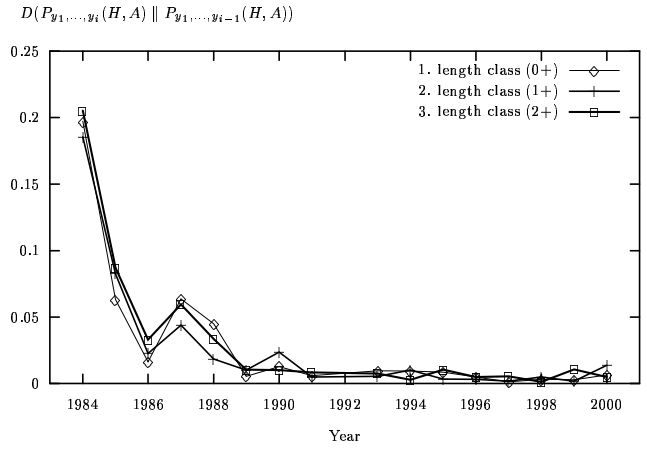


(c)

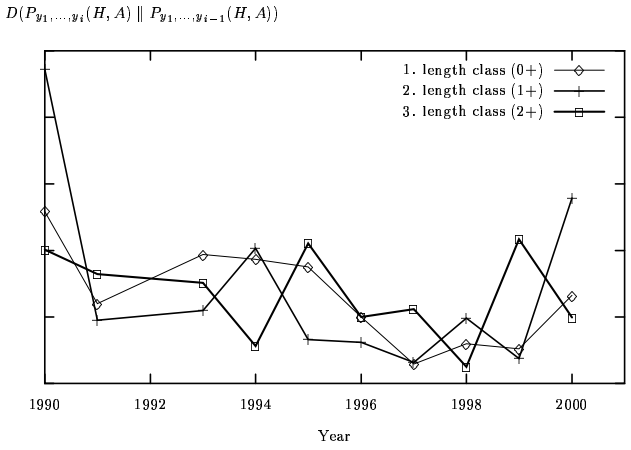


(d)

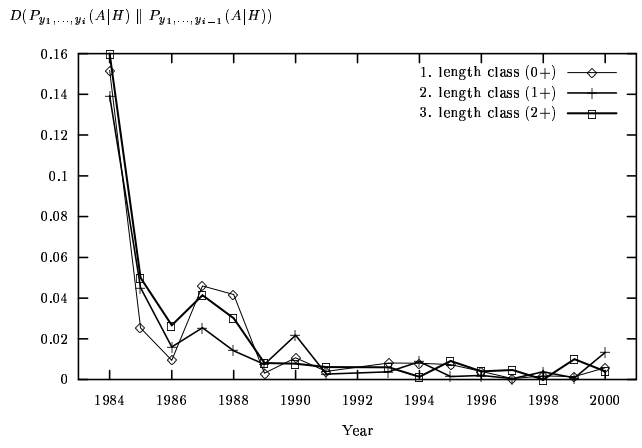
Fig. 14. Convergence of the empirical joint distribution, river Simo. (a) Momentary changes in the empirical joint distribution, the whole time series. (b) Momentary changes in the empirical joint distribution, a closer look at the last ten years. (c) Momentary changes in the abundance - habitat relationship, the whole time series. (d) Momentary changes in the abundance - habitat relationship, a closer look at the last ten years.



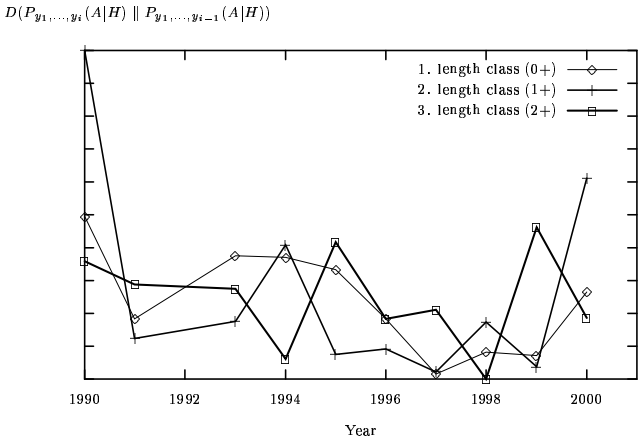
(a)



(b)



(c)



(d)