HELSINKI
INSTITUTE FOR
INFORMATION
TECHNOLOGY

# Predicting the Wild Salmon Production Using Bayesian Networks

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

December 22, 2002

HIIT
TECHNICAL
REPORT
2002–7

# PREDICTING THE WILD SALMON PRODUCTION USING BAYESIAN NETWORKS

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

Helsinki Institute for Information Technology HIIT

Tammasaarenkatu 3, Helsinki, Finland

PO BOX 9800

FIN-02015 HUT, Finland

http://www.hiit.fi

HIIT Technical Reports 2002–7

ISSN 1458-9451

# Predicting the wild salmon production using Bayesian networks

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

**Abstract**: From the management point of view, the production of wild smolts is the most important indicator of the status of a river's salmon population. We present a methodology allowing the prediction of the number of wild smolts in a river in a consistent and well-defined fashion. Our framework is probabilistic and our approach Bayesian. Our models are Bayesian networks, which have a simple graphical representation allowing visualization of the obtained knowledge. Being the state-of-the-art classifier in many domains, they also possess predictive power. We emphasize empirical modeling, studying what can be learned from the existing real-world data for two Gulf of Bothnia rivers, Simo and Tornio (the Finnish side). To ensure that our models generalize well, we employ strict validation procedures, where care is taken to inhibit leakage of information from the validation set to the training set. Furthermore, with the needs of fisheries management in mind, we highlight the role of the loss function in modeling, evaluating our models also in a setting where it is a greater error to over- than underestimate the size of a population.

## 1. Introduction

The main goal of salmon fisheries management is to maximize the level of fishing, while maintaining a stock of sufficient size and genetic diversity. A method for assessing and predicting the status of a river's salmon population is needed to tackle this task. The aim of this paper is to exhibit such a methodology.

We limit ourselves to the nursery river phase in the life cycle of salmon. We divide this phase into three stages:

1. The reproduction stage. The output of this stage, eggs, depends on the abundance of ascending adults and their success in spawning, affected by environmental factors such as the M74 syndrome.

2. The parr stage. This is the period lasting from one up to as many as six years, during which the egg-emerged juvenile salmon stay in the river.

3. The smolt stage. Having undergone physiological changes, young salmon migrate downstream to the sea.

From the managerial point of view, the smolt stage is the most important one: the number of wild smolts is the yardstick of choice for determining the status of a river's wild salmon population. Hence, in this work we focus on it, and develop models for the prediction of wild smolt production, using real-world data from two Gulf of Bothnia rivers, Simo and Tornio.

**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
**Lars Karlsson and Ingemar Perä.**
Firstname.Lastname@fiskeriverket.se

This report is structured as follows. We first outline our approach to modeling in Chapter 2, proceeding to describe our real-world data sets in detail in Chapter 3. In Chapter 4 we define our methodology formally, giving examples of its application in fisheries. The results of our empirical work are described in Chapter 5. Finally, we summarize our conclusions in Chapter 6.

## 2. Modeling approach

To be able to handle uncertainty in a consistent and well-defined fashion, we adopt the probabilistic framework, and choose the Bayesian approach within it, with Bayesian networks [9] as our model family. Our goal is to learn a predictive model for wild smolt production based on the available data, using different criteria for model selection. Our emphasis is on empirical modeling: although our methodology allows the expression of biological knowledge, in this paper we obtain our models from the existing data alone. The results should thus be seen as a baseline to compare biological knowledge to, as well as an assessment of the amount of information in the available data from the point of view of smolt production prediction.

Our point of view is managerial: the resulting models should generalize well, and be capable of taking into account the needs of fisheries management. By this we mean that our goal is to find models that predict well in the future. The main problem in learning predictive models is to avoid overfitting, i.e. the situation where we fit our model too accurately to the available data, compromising our predictive performance for future data.

To test whether we have succeeded in generalizing, we validate our models using strict procedures, where pains are taken to ensure that the model learner is never allowed to gain information from the validation set. This goal of avoiding the fit of an unnecessarily complex model to the data is especially called upon in our empirical work because of the relatively short time series available. We also highlight the role of the loss function in the prediction scheme. That is, not only do we look for a model with a small amount of error in its predictions,

but keeping in mind the aims of fisheries management we also study the difference between a situation where it does not matter whether we over- or underestimate, and the more realistic situation where we prefer a pessimistic model, i.e. one whose errors tend to be underestimates rather than overestimates.

## 3. Real-world data sets

Our empirical studies use data from two Gulf of Bothnia rivers, Simo and Tornio. The data sets are limited to the river phase in the life cycle of salmon. The reproduction stage is covered by data on ascending adults and the M74 syndrome. The parr stage is described by electrofishing data, and the smolt stage by data on smoltification age and the number of seabound wild smolts. In this study we have studied wild salmon only. Table 1 provides an overview of our set of variables and their availability per river.

We will now elaborate on our variables in some detail, going through the various life stages in order.

### 3.1. Reproduction stage data

The earliest stage in the life cycle of salmon, the egg stage, depends both on the abundance of ascending adults and on spawning success. Neither of the two rivers contains a fish ladder, so the only available means of measuring the abundance of ascending adults is via catches of adults in the river. Both the sum of weights and the number of fish are available for each year. We will denote catches in numbers at year $i$ by $C_i^n$ and catches in kilos by $C_i^k$. Unfortunately we lack data on the fishing effort, which makes this data a somewhat uncertain indicator of abundance.

To take spawning success and environmental factors into account to some degree, we also use data on M74 mortality (in percentages) at year $i$, denoted by $M_i$.

As an attempt at a synthetic variable characterizing reproduction as a whole, we created "M74-affected" versions of the adult abundance variables, describing the estimated effect of M74 on reproduction. The values of these new variables $R_i^n$ and $R_i^k$ (reproduction in numbers and kilos, respectively) are catches multiplied by $(1 - M_i/100)$.

### 3.2. Parr stage data

For each year in our time series we have electrofishing data from a subset of the total set of electrofishing sites in the river. The yearly choices of sites and their number vary over time, especially in the beginning of our time series. See Fig. 1 for the variance in the yearly locations of electrofishing sites.

Our data set comprises electrofishing data at 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.

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

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.

Because our aim is to have a model transferable from Simo to Tornio and vice versa, we take all this site-specific data and summarize it for each year in terms of densities per age class. A model containing the sites themselves as random variables could naturally not be applied to a river with a different set of sites. We have adopted and compared two ways of obtaining age-class density estimates. The first one is based on estimation by domain experts using an electorfishing model, the second one on average observed densities per length class.

#### *3.2.1. Estimation by an electrofishing model*

The electrofishing data provides us with ready-made density estimates for age groups "$0+$" and "$> 0+$". Actually, for river Tornio we possess a finer-grained division to "$0+$", "$1+$" and"$> 1+$". Since we want to compare the results of using the data of either Simo or Tornio, or of both, we are forced to employ the coarser division. We will use $E_i^{0+}$ and $E_i^{>0+}$ to denote the expert-estimated density at year $i$ of age $0+$ and older than $0+$ parr respectively.

These estimates are derived by domain experts using an electrofishing model where the actual amount of fish at a site is estimated using measurements from a series of fishing runs. The main assumption is that the catchability of the fish stays constant across the series. It is also assumed that the age of the fish can be determined reliably (but actually this information is often missing).

#### *3.2.2. Average length-class densities*

To have a point of comparison, we decided to provide an alternative, more data-oriented way of estimating yearly density for each disjoint class of fish. An important point to note is that our assumptions are somewhat weaker than those adopted in the estimates of the domain experts' electrofishing model.

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 of the rivers. By taking the first fishing runs only, we weaken the assumption of constant catchability made in the domain expert estimates. We only assume the catchability of fish during the first fishing run to be the same as that of any first fishing run.

As observed above, we have ready-made ages for the fish in the data, but for part of the data the age is missing. As an

**Table 1.** An overview of the variables. "*" signifies availability.

| Stage | Variable group | Variable | Symbol | Simo | Tornio |
|---|---|---|---|---|---|
| Reproduction | Catch | Catch in kilos | $C_i^k$ | * | * |
| | | Catch in numbers | $C_i^n$ | * | * |
| | M74 | M74 mortality | $M_i$ | * | * |
| | M74-affected catch | Reproduction in kilos | $R_i^k$ | * | * |
| | | Reproduction in numbers | $R_i^n$ | * | * |
| Parr | Average length-class densities | Average density 0+ (det. by length) | $L_i^{0+}$ | * | * |
| | | Average density 1+ (det. by length) | | * | * |
| | | Average density 2+ (det. by length) | | * | * |
| | | Average density >0+ (det. by length) | $L_i^{>0+}$ | * | * |
| | Estimated densities | Estimated density 0+ | $E_i^{0+}$ | * | * |
| | | Estimated density >0+ | $E_i^{>0+}$ | * | * |
| | | Estimated density 1+ | | | * |
| | | Estimated density >1+ | | | * |
| Smolt | Smoltification age | % wild salmon smoltifying at age 1 | $A_i^1$ | * | * |
| | | % wild salmon smoltifying at age 2 | $A_i^2$ | * | * |
| | | % wild salmon smoltifying at age 3 | $A_i^3$ | * | * |
| | | % wild salmon smoltifying at age 4 | $A_i^4$ | * | * |
| | | % wild salmon smoltifying at age 5 | $A_i^5$ | * | * |
| | Smolt production | Estimated number of seabound wild smolts, based on mark-recapture data | $S_i$ | * | * |

alternative approach, we drop this assumption, and classify the fish in another disjoint and exhaustive way: by their length.

We use 7 and 11 cm as the split points, i.e. all fish smaller than 7 cm were considered to be $0+$, all fish longer than 11cm $2+$, and all others $1+$. These split points were determined by experts. To see how they correspond to the empirical length distributions of age-classified fish in our data sets, see Fig. 2. 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.

Given these observed densities from first fishing runs for fish of certain *length* class, we assume that the first fishing runs are comparable across the sites sampled during a year, and take the average of the observed densities as our estimate of the density for that length class during that year. We assume here that the bias in the selection of sites to electrofish stays constant across our time series. The veracity of this assumption in this data set was studied by us in [11], where it was seen to hold quite well.

We will use $L_i^{0+}$ and $L_i^{>0+}$ to denote the average density at year $i$ of length-class $0+$ and longer parr respectively.

#### *3.2.3. Comparison of estimation methods*

The biological knowledge -incorporating electrofishing model used by domain experts is more sophisticated than the length-class approach put forth here. The length-class method should be viewed as a data-based baseline: any system with stronger assumptions should at the least be able to beat it in the predictive sense.

Fig. 3 and Fig. 4 compare expert estimates with length-class estimates. It can be seen that for river Simo there is a plausible linear correlation between the two estimates, whereas for river Tornio only the plot for $> 0+$ parr exhibits such tendencies. It has to be kept in mind that we have no or very little data for much of the range — only the low end of the range is well covered.

### 3.3. Smolt stage data

The smolt stage is characterized in two ways in our data. The age distribution of smolts has been estimated by experts, based on samples from trapped seabound fish. Unfortunately, we only have this data in an age-classified form. It would have been interesting to compare age-classified data to length-classified ones, as we did in the case of parr density estimation. We denote by $A_i^j$ the percentage at year $i$ of wild salmon smoltifying at age $j$.

The values of $S_i$, the number of seabound smolts at year $i$, consist of domain expert estimates based on mark-recapture data.

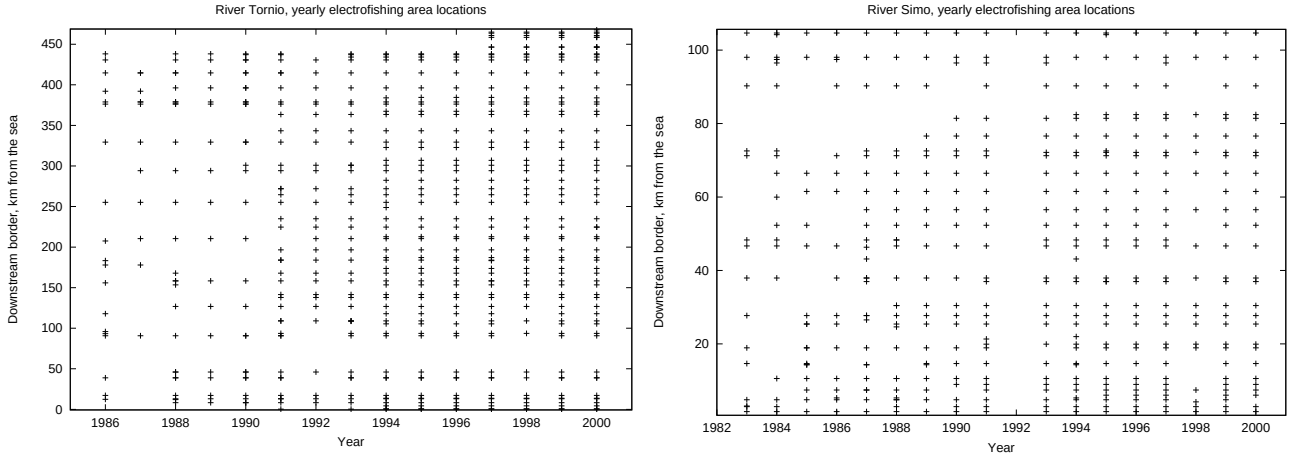
## 4. Methodology

Adopting the probabilistic framework, we assume our models to be probability distributions. Since we are in this work interested in finding a model that predicts well for a particular variable, our task is somewhat different from the general goal of modeling the joint distribution of all variables describing a river's salmon population.

### 4.1. The focused prediction problem

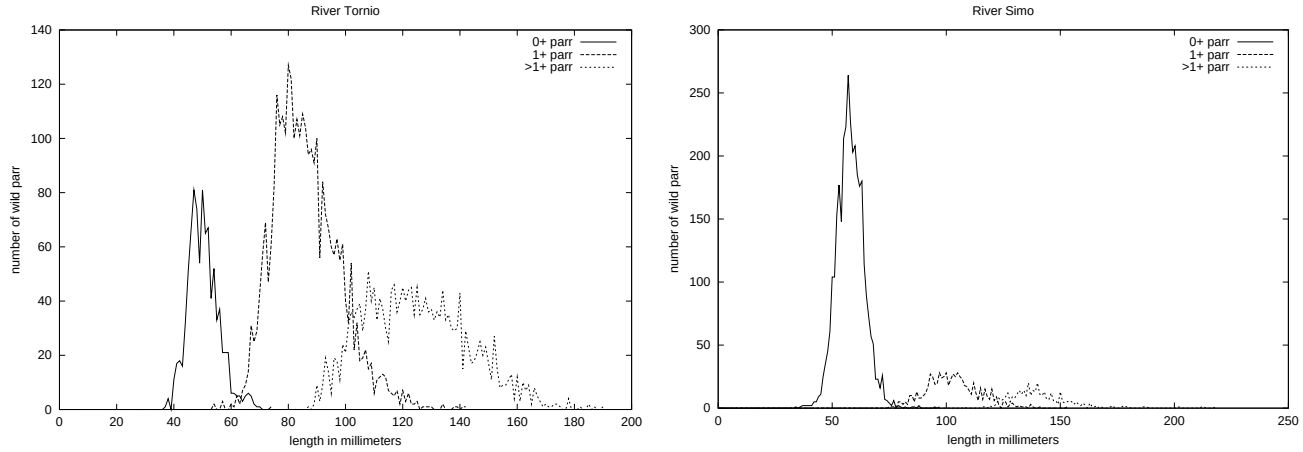
*Classification* means the task of predicting the value of a discrete class variable, given the values of other variables, called *predictors*. In classifier learning the goal is to build accurate classifiers given a sample of classified instances, i.e. vectors consisting of the values of the predictors together with the corresponding value for the class variable.

**Fig. 1.** The yearly locations of electrofished sites. (a) River Tornio (the Finnish side). (b) River Simo.




(a) (b)

**Fig. 2.** The empirical length distributions of fish aged in the data. (a) River Tornio (the Finnish side). (b) River Simo.



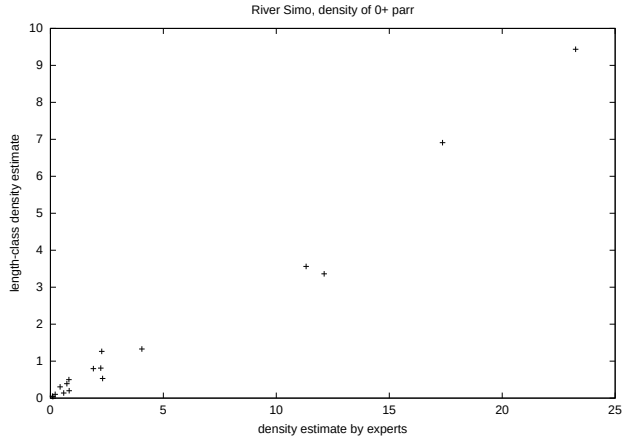

(a) (b)

In this work, our predicted variable is in fact not discrete, but continuous, and, properly speaking, we are doing *regression*. We transform the regression problem to a classification problem by discretizing the predicted variable and interpreting our posterior to be a continuous histogram distribution. Our point estimate will then be the expected value of this histogram. Hence, we can use *focused prediction* as a general term covering both cases.

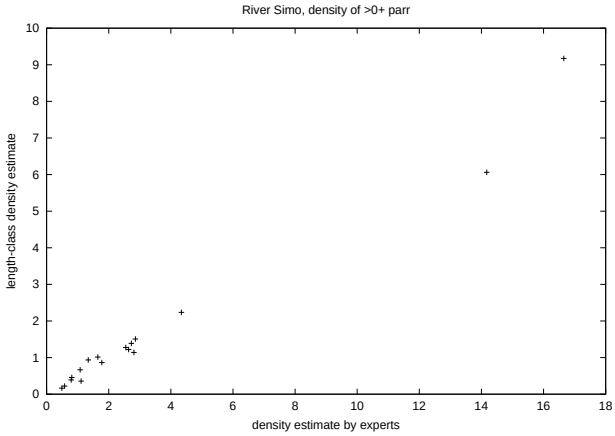
In learning a focused predictor, the goal is to build accurate predictors from a given *training data set* $\mathbf{D} = (\mathbf{x}^N, y^N)$, a matrix of $N$ vectors each consisting of values of $m$ predictor variables $X_1, \ldots, X_m$, together with a value for the predicted variable $Y$. Together, our variables form the *domain* $\mathcal{V} = \{X_1, \ldots, X_m, Y\}$. We will use notation $V_i$ to refer to any variable in our domain, whether it is the focus of prediction or not. In the interest of simplicity, from now on we assume the predictor variables $X_i$ to be discrete as well. We discretize the continuous variables in our data sets, so this assumption does not constrain us in any way. In general formal terms, our aim is to produce the predictive distribution $P(Y \mid X_1, \ldots, X_m)$.

Since our particular task in this work is to predict the wild salmon production in a river using all available information, our focus is on $S_i$, the number of smolts at year $i$. Furthermore, we aim at building a model that allows us to predict for a particular year $i$ given the past, that is, data from years preceding $i$, but not from year $i$ itself. As we are constrained to the nursery river phase, we cannot look more than five years back in time, since we assume that after six years all juvenile salmon have left the river. Putting all this together, the predic-

**Fig. 3.** Comparison of density estimates made by domain experts to average length-class estimates, river Simo. Each point is a pair of corresponding estimates for a year. (a) $0+$ parr. (b) Older parr.




(a)




(b)

tive distribution we are aiming at is

$$
\begin{aligned}
P(S_i \mid & C^n_{i-1}, C^k_{i-1}, M_{i-1}, R^n_{i-1}, R^k_{i-1}, E^{0+}_{i-1}, E^{>0+}_{i-1}, \\
& L^{0+}_{i-1}, L^{>0+}_{i-1}, A^1_{i-1}, A^2_{i-1}, A^3_{i-1}, A^4_{i-1}, A^5_{i-1}, \\
& S_{i-1}, \ldots, C^n_{i-5}, C^k_{i-5}, M_{i-5}, R^n_{i-5}, R^k_{i-5}, \\
& E^{0+}_{i-5}, E^{>0+}_{i-5}, L^{0+}_{i-5}, L^{>0+}_{i-5}, A^1_{i-5}, A^2_{i-5}, A^3_{i-5}, \\
& A^4_{i-5}, A^5_{i-5}, S_{i-5}).
\end{aligned} \tag{1}
$$

### 4.2. Bayesian networks

Taking the Bayesian approach within the probabilistic framework, we choose Bayesian networks[1] as our model family. Bayesian networks [9] define joint probability distributions via a set of independence assumptions $B_S$ that can be conveniently expressed as a directed acyclic graph (see Fig. 5(a)).

The nodes of the directed acyclic graph correspond to variables, while the arcs represent the independence assumptions. That is, whenever an arc is missing, we assume the two variables in question to be pairwise conditionally independent.

The model families $\mathcal{B}$ we consider thus consist of a finite number of probabilistic Bayesian network structures

$$\mathcal{B} = \{B_{S_1}, \ldots, B_{S_K}\}.$$

One of the key properties of Bayesian networks is that the joint probability distribution can be factorized as follows:

$$P(X_1, \ldots, X_m, Y) = \prod_{i=1}^{m+1} P(V_i \mid \mathbf{\Pi}_i), \tag{2}$$

where $\mathbf{\Pi}_i$ denotes the *parents* (immediate predecessors in the graph) of variable $V_i$. The parameters $B_\Theta$ of a Bayesian network model determine the local conditional probability distributions $P(V_i \mid \mathbf{\Pi}_i)$. This means that a Bayesian network structure $B_S$, together with $B_\Theta$, defines a joint probability distribution $P(X_1, \ldots, X_m, Y \mid B_S, B_\Theta)$ via (2).

[1] For an interactive tutorial on Bayesian networks and links to reference material, see site `http://b-course.hiit.fi`.

**Example 4.1.** *Let our domain be*

$$\mathcal{V} = \{S_i, C^n_{i-3}, M_{i-2}, E^{0+}_{i-2}, E^{>0+}_{i-1}\}$$

*i.e. each data vector consists of an estimate of the number of smolts at year $i$, catches in numbers three years earlier, M74 percentages two years earlier, domain expert estimates of the densities of $0+$ parr two years back and domain expert estimates of the densities of older parr in the previous year.*

*Let Fig. 5(b) present graphically a structure $B_{S_i} \in \mathcal{B}$ describing the domain. Given Fig. 5(b), our joint distribution can be written down as*

$$
\begin{aligned}
P(S_i, & C^n_{i-3}, M_{i-2}, E^{0+}_{i-2}, E^{>0+}_{i-1}) = \\
& P(C^n_{i-3}) P(M_{i-2}) \\
& \cdot P(E^{0+}_{i-2} \mid C^n_{i-3}, M_{i-2}) P(E^{>0+}_{i-1} \mid E^{0+}_{i-2}) \\
& \cdot P(S_i \mid E^{0+}_{i-2}, E^{>0+}_{i-1}).
\end{aligned}
$$

### 4.3. Model selection criteria

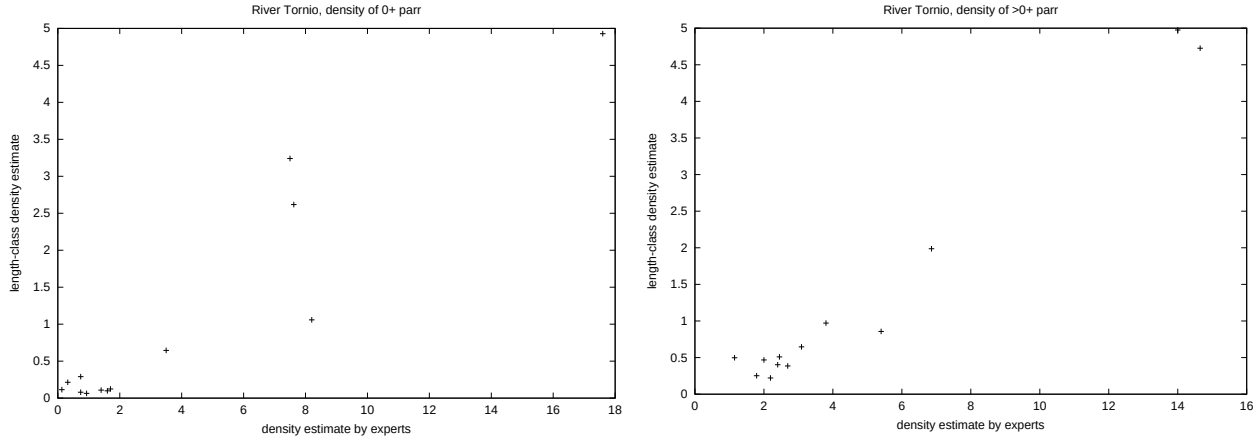
Given a data set $\mathbf{D}$ and a set of possible Bayesian network structures $\mathcal{B}$, we are faced with the task of selecting a model structure from our set of candidates. Our aim is to find the model (structure) that describes the domain the best, having seen a set of observations $\mathbf{D}$ from it. In this work we use two different selection criteria, one a purely Bayesian one, the other an empirical one with advantages which will become clearer in Chapter 4.6.

#### *4.3.1. The marginal likelihood criterion*

Given a training set $\mathbf{D}$ it is possible, with certain technical assumptions (see [5]), to compute the predictive distribution for a single Bayesian network structure $B_S$ via

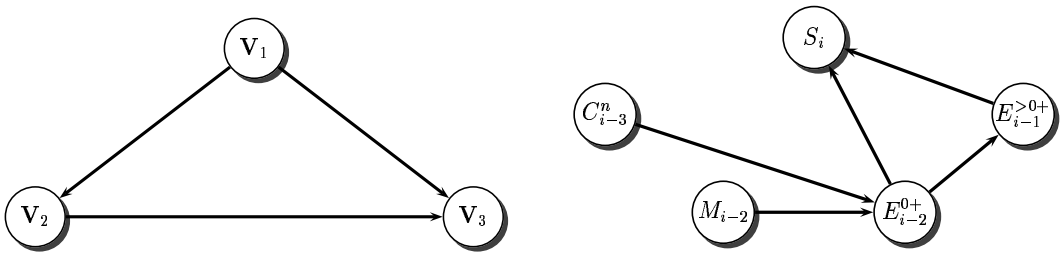
$$
\begin{aligned}
P(X_1, & \ldots, X_m, Y \mid B_S, \mathbf{D}) = \\
& \int P(X_1, \ldots, X_m, Y \mid B_S, B_\Theta, \mathbf{D}) \\
& \quad \cdot P(B_\Theta \mid B_S, \mathbf{D}) dB_\Theta.
\end{aligned} \tag{3}
$$

**Fig. 4.** Comparison of density estimates made by domain experts to average length-class estimates, river Tornio (Finnish side). Each point is a pair of corresponding estimates for a year. (a) $0+$ parr. (b) Older parr.




(a) (b)

**Fig. 5.** (a) An example of a Bayesian network representing the joint distribution $P(V_1, V_2, V_3)$ as $P(V_1)P(V_2|V_1)P(V_3|V_1, V_2)$. (b) The structure $B_{S_i}$ of Example 4.1.




(a) (b)

If we now, instead of using only a single model structure, average over all Bayesian network structures $B_S \in \mathcal{B}$ in our model family, we get

$$P(X_1, \ldots, X_m, Y \mid \mathbf{D}, \mathcal{B}) = \sum_{B_S \in \mathcal{B}} P(X_1, \ldots, X_m, Y \mid B_S, \mathbf{D}) P(B_S \mid \mathbf{D}, \mathcal{B}), \tag{4}$$

where the first term was given in (3). The second term is the posterior probability of $B_S$ after seeing the data $\mathbf{D}$. Intuitively, if one wants to choose a single model from $\mathcal{B}$, it makes sense to select the model maximizing this posterior since that particular model has the highest overall weight in sum (4). Assuming the prior $P(B_S \mid \mathcal{B})$ to be uniform, this is equivalent to choosing the model with the highest *marginal likelihood* $P(\mathbf{D} \mid B_S, \mathcal{B})$, since

$$P(B_S \mid \mathbf{D}, \mathcal{B}) \propto P(\mathbf{D} \mid B_S, \mathcal{B}) P(B_S \mid \mathcal{B}). \tag{5}$$

With certain technical assumptions [5], the marginal likelihood can be calculated in closed form:

$$P(\mathbf{D} \mid B_S, \mathcal{B}) = \prod_{i=1}^{m+1} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})}, \tag{6}$$

where $\Gamma$ denotes the gamma function, $q_i$ is the number of value combinations for parents of variable $V_i$, $r_i$ is the number of values variable $V_i$ has, $N_{ijk}$ are the sufficient statistics (the number of cases in the data where variable $i$'s parents' values are in configuration (value combination) $j$ when the variable itself has value $k$), $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ and $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$.

The constants $N'_{ijk}$ are the *hyperparameters* determining the parameter prior distribution $P(B_\Theta \mid B_S, \mathcal{B})$. Following the suggestion in [1], in our empirical work we have picked the prior

$$N'_{ijk} = \frac{N'}{r_i \cdot q_i} \tag{7}$$

as our parameter prior, with the setting $N' = 1$. Intuitively put this means that we deem that all states of the conditional

distribution of a variable given its parents are a priori equally likely. This prior is also overridden by data relatively fast ($N'$ is relatively small). Our reasons are twofold. Firstly, since our domain comprises 65 predictors, and we seek among different discretizations, it would be a formidable task for experts to assess and specify the parameter priors for all possible structures and discretizations. Secondly, the amount of data is relatively small from the viewpoint of empirical modeling: data sets of 15 and 17 vectors per river. Any strong prior is prone to override the data, whereas our aim in this paper is to see what can be learned from the existing data alone.

**Example 4.2.** *Let our domain be $\mathcal{V} = \{S_i, R^n_{i-5}\}$, i.e. each data vector consists of the estimated number of smolts at year $i$ and an index of reproduction in numbers five years earlier.*

*Let us consider all possible Bayesian network structures in this domain, i.e. $\mathcal{B} = \{B_{S_1}, B_{S_2}, B_{S_3}\}$, where $B_{S_1}$ corresponds to the assumption that $S_i$ is independent of $R^n_{i-5}$ and v.v., and $B_{S_2}$ and $B_{S_3}$ are models where they are dependent. Fig. 6 shows the set of structures $\mathcal{B}$.*

*For simplicity of exposition, let us further assume that both the number of smolts and the reproduction index have been discretized to only two categories:* few *and* many.

*Let $\mathbf{D}_1$ consist of 20 years of data. The sufficient statistics of $\mathbf{D}_1$ are shown in Table 2(a). You can see that regardless of the value of $R^n_{i-5}$, the relationship of events "$S_i$= few" and "$S_i$= many" stays more or less the same.*

*We can now calculate the marginal likelihood of all structures $B_S \in \mathcal{B}$ using (6) and the prior (7). For example,*

$$P(\mathbf{D}_1 \mid B_{S_2}, \mathcal{B}) = \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{1}{2}+12)}\left(\frac{\Gamma(\frac{1}{4}+10)}{\Gamma(\frac{1}{4})}\cdot\frac{\Gamma(\frac{1}{4}+2)}{\Gamma(\frac{1}{4})}\right) \cdot \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{1}{2}+8)}\left(\frac{\Gamma(\frac{1}{4}+8)}{\Gamma(\frac{1}{4})}\cdot\frac{\Gamma(\frac{1}{4}+0)}{\Gamma(\frac{1}{4})}\right).$$

*The marginal likelihood of each structure is shown in Table 2(b).*

*It can be seen that the structure with no arc (dependency) is slightly preferred, being 1.25 times more likely than the structures with an arc, and that the direction of the arc doesn't matter in this case: $B_{S_2}$ and $B_{S_3}$ are equivalent with respect to our criterion, given our prior and the data set $\mathbf{D}_1$ (to see why this is so, see [5]).*

*Let $\mathbf{D}_2$ now be a similar data set of 20 years with the same variables, but with different sufficient statistics as shown in Table 3(a). In other words, the sufficient statistics of the event "$R^n_{i-5}$ = many" have been exchanged, making the sufficient statistics of "$S_i$= few" and "$S_i$= many" radically different depending on the value of $R^n_{i-5}$. Calculating the marginal likelihoods, we arrive at the results of Table 3(b), which show that $\mathbf{D}_2$ provides evidence for a dependency between $S_i$ and $R^n_{i-5}$, indicating a 263 times higher likelihood than the structure with no dependency.*

#### *4.3.2. Empirical criteria*

Another, non-Bayesian, way of scoring model structures is by using an empirical criterion, i.e. by comparing the predictive performance of structure candidates in a test set. The parameters $B_\Theta$ for a candidate structure $B_S$ are first learned from a training data set. The predictive performance of the resulting model is then measured in the test set in terms of the loss function (see Chapter 4.6) adopted. In Chapter 4.4.1 we discuss the use of empirical criteria for structure search in more detail.

### 4.4. Searching for the best structure

Even using the criteria of the previous chapter for comparing structures, searching among all possible Bayesian network structures is computationally too hard for practical purposes, especially in our domain where there are 65 predictors: the problem is NP-hard if a node can have more than one parent. Therefore, a natural approach is to limit $\mathcal{B}$ to a subset of all possible structures.

As discussed above, a Bayesian network model represents the joint distribution $P(X_1, \ldots, X_m, Y)$. From this joint distribution we aim to extract the predictive distribution $P(Y \mid X_1, \ldots, X_m)$. We can distinguish two different approaches to estimating the predictive distribution [2]: in the *diagnostic* paradigm one tries to estimate the distribution directly, while in the *sampling* paradigm one estimates the distributions $P(X_1, \ldots, X_m \mid Y)$ and $P(Y)$, from which the desired predictive distribution can be computed by using the Bayes rule, which implies

$$P(Y \mid X_1, \ldots, X_m) \propto P(X_1, \ldots, X_m \mid Y)P(Y). \quad (8)$$

In visual terms, in the sampling paradigm all of the arcs connected to the focus node are leaving arcs, in the diagnostic paradigm arriving arcs.

While our approach here is general, biological knowledge could be taken into consideration when choosing the set of candidate structures. We will now describe some means of searching for good structures from within a subset of all possible structures in both paradigms.

#### *4.4.1. The sampling paradigm*

An example of a sampling-type Bayesian network is the Naive Bayes model, a Bayesian network with one arc from the predicted node to each of the predictor nodes (see Fig. 7). This graph structure represents the assumption that the predictors are independent of each other, given the value of the predicted variable. This assumption might sound naive, but the Naive Bayes classifier is in fact in many real-world cases the state-of-the-art classifier, as, for example, its success in prediction competitions like the KDD Cup and the CoIL competition illustrate. Naturally, often this independence assumption is more or less false. We can try to counter this deficiency by several means.

One strategy is *variable selection*. In variable selection only variables which have a sufficient dependency from the focus of prediction are modeled as dependent on it. In graphical terms, we seek for a subset of arcs from the predicted variable to the predictors. To do this, we use either one of the criteria of Chapter 4.3 to assess whether to draw an arc from the focus of prediction to a predictor.

**Example 4.3.** *Let our domain be*

$$\mathcal{V} = \{S_i, E^{0+}_{i-1}, E^{>0+}_{i-1}, E^{0+}_{i-2}, E^{>0+}_{i-2}\},$$

*i.e. in addition to the focus of prediction, $S_i$, we have density estimates from the two previous years. As in earlier examples,*

**Fig. 6.** $\mathcal{B}$ of Example 4.2.

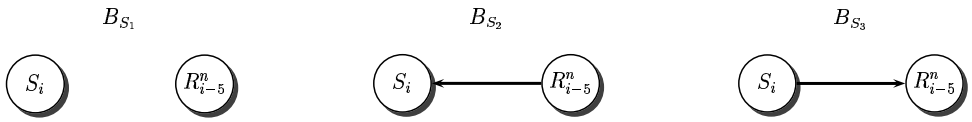



**Table 2.** Example 4.2. (a) $N_{ijk}$ of $\mathbf{D}_1$, i.e. the numbers of cases where the variables have a particular value combination in the data. (b) The marginal likelihoods of the structures using data $\mathbf{D}_1$.

| | $R^n_{i-5}$= few | $R^n_{i-5}$= many |
|---|---|---|
| $S_i$= few | 10 | 8 |
| $S_i$= many | 2 | 0 |

(a)

| | marginal likelihood |
|---|---|
| $B_{S_1}$ | $6.55 \cdot 10^{-11}$ |
| $B_{S_2}$ | $5.23 \cdot 10^{-11}$ |
| $B_{S_3}$ | $5.23 \cdot 10^{-11}$ |

(b)

*let all data be discretized to two categories,* few *and* many. *The sufficient statistics of a data set* $\mathbf{D}$ *consisting of 20 years of data are shown in Table 4(a).*

*Let us pick marginal likelihood as our structure search criterion. As in Example 4.2, let us use Buntine's prior for our parameters. Because we have restricted our set of structures* $\mathcal{B}$ *to those in the Naive Bayes class, all arcs are of the* $S_i \longrightarrow X$ *type, where* $X$ *is any predictor. The marginal likelihood of each possible focus of prediction–predictor substructure, i.e. the score of not having vs. adding an arc is shown in Table 4(b).*

*We can see that it is 330 times more likely that there is a dependency between* $S_i$ *and* $E^{0+}_{i-2}$*, than that there is not,* $E^{>0+}_{i-2}$ *is slightly on the independent side,* $E^{0+}_{i-1}$ *more so, and* $E^{>0+}_{i-1}$ *is shown to be 1.26 times more likely to be dependent on* $S_i$ *than not.*

An important practical feature of the marginal likelihood criterion is that if any node has at most one parent, the criterion decomposes to subscores for each arc, i.e. we can evaluate the gain or loss of adding an arc regardless of the rest of the structure. This naturally makes searching in this restricted structure space very efficient.

**Example 4.4.** *Since no node has more than one parent in the model class of example 4.3, we can express the varying evidences for dependency between variables graphically by letting the thickness of an arc indicate the amount of evidence (marginal likelihood) for that particular arc. Because the range of values for the likelihood ratio can vary from 1 to thousands in practice, we take its logarithm to keep the result visually pleasing. Furthermore, if there is no arc from the focus of prediction* $S_i$ *to a predictor in the set of structures under consideration, that predictor has no effect on the predicted variable. Thus, we can leave out such unconnected nodes from our graph, arriving at Fig. 8 in our case.*

In the case of an empirical criterion, the criterion is not similarly decomposable, so a search algorithm is needed. Since the number of possible structures can be huge, a randomized search is a natural choice. In our empirical work we have used stochastic greedy search: we pick randomly an arc operation to be performed, and evaluate empirically whether it is likely to enhance predictive performance in the validation set. Note that the model must not see any of the validation set prior to the actual validation. Otherwise the empirical criterion will overfit to the validation set, providing misleadingly positive results. For this reason, the predictive value of an arc operation has to be assessed by splitting randomly the training data to a "second-order" training set and a test set. The two structures, prior and after the arc operation, are then both taught on the "second-order" training set, and their performance assessed by predicting for the test set. To avoid good or bad luck in the choice of a split, this splitting is done a number of times, and the performance measured by a loss function (see Chapter 4.6).

We can also relax the independence assumption by other means, e.g. by connecting highly relevant predictors via a fully connected subnetwork. If we connect subsets of predictors fully, we speak of a *partitioning* network (see Fig. 9(a)). A problem with this approach is that the number of parameters grows exponentially with the sizes of the subsets. In our task, where we have a small amount of data, this is a major concern. See [7] for more on partitioning networks.

### *4.4.2. The diagnostic paradigm*

A major problem in the learning of diagnostic structures from data is the number of parameters: the conditional distribution of the focus variable given the predictors has a number of parameters growing exponentially with the number of predictors.

To see this, let us look at the case of using marginal likelihood as our search criterion. First of all, in addition to the previous definition of marginal likelihood, we can also define the *supervised (conditional) marginal likelihood* as

$$P(y^N \mid \mathbf{x}^N, B_S, \mathcal{B}) = \int P(y^N \mid \mathbf{x}^N, B_S, B_\Theta, \mathcal{B}) \cdot P(B_\Theta \mid \mathbf{x}^N, B_S, \mathcal{B})\, dB_\Theta, \tag{9}$$

Our motivation for this definition is that the unsupervised marginal likelihood criterion tends to favor models that model well both

**Table 3.** Example 4.2. (a) $N_{ijk}$ of $\mathbf{D}_2$, i.e. the numbers of cases where the variables have a particular value combination in the data. (b) The marginal likelihoods of the structures using data $\mathbf{D}_2$.

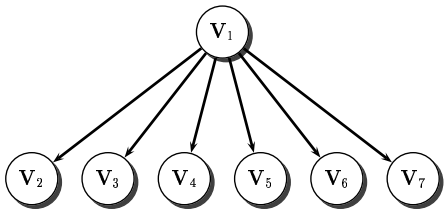
| | $R^n_{i-5}$= few | $R^n_{i-5}$= many |
|---|---|---|
| $S_i$= few | 10 | 0 |
| $S_i$= many | 2 | 8 |

(a)

| | marginal likelihood |
|---|---|
| $B_{S_1}$ | $0.02 \cdot 10^{-11}$ |
| $B_{S_2}$ | $5.23 \cdot 10^{-11}$ |
| $B_{S_3}$ | $5.23 \cdot 10^{-11}$ |

(b)

**Fig. 7.** The Naive Bayes structure, i.e. the focus of prediction is $V_1$ and $P(V_1, V_2, V_3, V_4, V_5, V_6) = P(V_1)P(V_2 \mid V_1)P(V_3 \mid V_1)P(V_4 \mid V_1)P(V_5 \mid V_1)P(V_6 \mid V_1)P(V_7 \mid V_1)$.




the predictors and the focus of prediction, which is clearly nonoptimal with respect to the focused prediction task. (See [3, 6, 8]).

The supervised marginal likelihood (9) can be computed in closed form similarly to (6):

$$P(y^N \mid \mathbf{x}^N, B_S, \mathcal{B}) = \prod_{j=1}^{q_y} \frac{\Gamma(N'_j)}{\Gamma(N'_j + N_j)} \prod_{k=1}^{r_y} \frac{\Gamma(N'_{jk} + N_{jk})}{\Gamma(N'_{jk})}. \quad (10)$$

where $q_y$ is the number of value configurations for the predictors $X_1, \ldots, X_m$, $r_y$ is the number of values $Y$ (the focus of prediction) has, $N_{jk}$ are the sufficient statistics (the number of cases in the data where the predictor values are in configuration $j$ and $Y$ has value $k$), and $N_j = \sum_{k=1}^{r_y} N_{jk}$. $N'_{jk}$ is our parameter prior as earlier.

Using (10) we can in principle calculate the supervised marginal likelihood of any diagnostic structure. The impracticality of the procedure in a domain like ours (with 65 predictors) is evident, however: $q_y$ grows rapidly with the number of predictors connected to the class variable.

We can bypass this obstacle by constructing mixtures of diagnostic networks, where each individual network has only a small number of arcs from the predictors to the predicted variable. The relevant predictor sets of each network can be overlapping or non-overlapping. For more on diagnostic structures, see [7]. Fig. 9(b) shows an example of two diagnostic structures of this type. Consequently, the result is a finite mixture of several diagnostic Bayesian network classifiers, where the individual predictions made by the models $B_S \in \mathcal{B}$ are weighted by the supervised marginal likelihood (9).

### 4.5. Discretization

Since most of our variables are continuous, we need to discretize them in order to be able to compute the marginal likelihood in closed form.

Formally, we define discretization as the process of finding a mapping $d : \mathcal{R}_i \to \mathcal{D}_i$, where $\mathcal{R}_i$ is the range of variable $V_i$, and $\mathcal{D}_i = \{0, 1, 2, \ldots, K-1\}$ is the set of $K$ discrete values we map the original values of variable $V_i$ to. The process of discretization consists of finding a set of $K-1$ threshold values $\mathcal{T}_i = (t_{i,1}, \ldots, t_{i,K-1})$, $t_{i,j} \leq t_{i,j+1}$.

An original value $v_{i,j}$ of variable $V_i$ is then mapped to $\mathcal{D}_i$ as follows:

$$d(v_{i,j}) = \begin{cases} 0 & \text{if } v_{i,j} \leq t_{i,1}. \\ i & \text{if } t_{i,k-1} < v_{i,j} \leq t_{i,k}. \\ K-1 & \text{if } v_{i,j} > t_{i,K-1}. \end{cases}$$

Whereas in our empirical work we have taken a fully data-oriented approach, biological knowledge could be employed in the determination of the threshold values, since the qualitative categories of say $0+$ parr density should be assessable by experts. In our empirical studies, we have employed two distinct types of discretization: *context-independent* and *context-dependent* mappings. The difference between these two types of methods is that in context-dependent methods we take values of other variables into account when searching for the threshold values.

#### *4.5.1. Context-independent mappings*

In context-independent discretization we study in isolation the values of the continuous variable to be discretized. We have used two different kinds of context-independent mapping in the empirical studies reported here: *equal-width* and *K-means* discretization.

Equal-width discretization is a simple method, making little or no use of the data itself. We simply split the range of the attribute into $K$ parts of identical size.

**Example 4.5.** *Let $L^{>0+}_{i-1}$, the average length-class density of $> 0+$ parr in the previous year, have values in the range*

**Table 4.** Example 4.3. (a) $N_{ijk}$, i.e. the numbers of cases where the variables have a particular value combination in the data. (b) The marginal likelihoods of not having vs. adding an arc between all focus of prediction - predictor pairs.

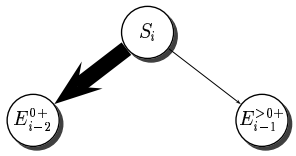
| | $S_i$= few | $S_i$= many |
|---|---|---|
| $E_{i-2}^{0+}$= few | 9 | 1 |
| $E_{i-2}^{0+}$= many | 1 | 9 |
| $E_{i-2}^{>0+}$= few | 8 | 4 |
| $E_{i-2}^{>0+}$= many | 2 | 6 |
| $E_{i-1}^{0+}$= few | 5 | 5 |
| $E_{i-1}^{0+}$= many | 5 | 5 |
| $E_{i-1}^{>0+}$= few | 5 | 1 |
| $E_{i-1}^{>0+}$= many | 5 | 9 |

(a)

| | marginal likelihood |
|---|---|
| $S_i \quad E_{i-2}^{0+}$ | $2.82 \cdot 10^{-14}$ |
| $S_i \longrightarrow E_{i-2}^{0+}$ | $933.31 \cdot 10^{-14}$ |
| $S_i \quad E_{i-2}^{>0+}$ | $4.22 \cdot 10^{-14}$ |
| $S_i \longrightarrow E_{i-2}^{>0+}$ | $3.46 \cdot 10^{-14}$ |
| $S_i \quad E_{i-1}^{0+}$ | $2.82 \cdot 10^{-14}$ |
| $S_i \longrightarrow E_{i-1}^{0+}$ | $0.37 \cdot 10^{-14}$ |
| $S_i \quad E_{i-1}^{>0+}$ | $14.62 \cdot 10^{-14}$ |
| $S_i \longrightarrow E_{i-1}^{>0+}$ | $18.47 \cdot 10^{-14}$ |

(b)

**Fig. 8.** Example 4.4. The structure discovered in Example 4.3 using marginal likelihood as the search criterion. The thickness of the arcs corresponds logarithmically to the evidence for that particular dependency. Only nodes connected to the focus of prediction (i.e. with higher evidence for an arc than for its absence) shown.




$\mathcal{R}_i = [0..9]$ *in the data. If we use equal-width discretization and* $K$ *is 3,* $\mathcal{T}_i = \{3, 6\}$*. Let our observed data be as shown in Table 5(a). Our discretization maps it as shown in the lower part of Table 5(a).*

Equal-width discretization thus depends only on the range of the variable. The range can be supplied as part of biological knowledge or determined from the data.

K-means discretization finds the threshold values by an iterative process. $K - 1$ division points are first placed within $\mathcal{R}_i$, defining subsets of values. The following procedure is then repeated $n$ times:

1. The means of each subset are calculated.
2. The new division points are put at the exact midpoints between successive means.
3. The values are dealt out into the new subsets defined by the new division points, in order.

The initial division points can be set in various ways. The procedure we used in our empirical studies was the following: the values of $V_i$ that occurred in the data were first ordered, and then split, in order, into $K$ disjoint subsets of equal size (i.e. each subset has $\lfloor r_i/K \rfloor$ elements). If $r_i$ is not exactly divisible by $K$, we make the first $r_i - K$ subsets one member larger. The end result of this iterative process is our set of threshold values $\mathcal{T}_i$. We have used $n = 5$ in our empirical studies.

**Example 4.6.** *Let our observed data and* $K$ *be as in example 4.5. Initially our disjoint subsets are* $\{0, 0.1\}$*,* $\{0.3\}$ *and* $\{8\}$*. The means of these subsets are* $[0.05, 0.3, 8]$*. Our first division points are* $[0.175, 4.15]$*. The value subsets defined by these split points equal our initial subsets, so the process has converged already. Our discretization of the original values is shown in the lower part of Table 5(a).*

To compare these two techniques, note that whereas equal-width discretization only depends on the range of $V_i$, K-means takes into account the distribution of the values of $V_i$ in the data: regions more densely packed with values get more densely packed threshold values.

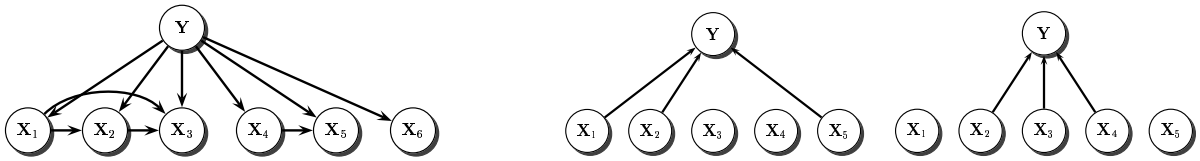
In our empirical studies, we performed both kinds of discretization based on the training data alone, again in order to avoid information leakage from the validation data to the model learner.

### *4.5.2. Context-dependent mappings*

In context-dependent discretizations the values of other variables co-occurring in data vectors are taken into account. Since our goal is focused prediction of variable $Y$, we take the values of $Y$ in each data vector as our context when discretizing $X_i$: i.e. we seek for such a set of threshold values that the values within each discrete category have as similar a context as possible.

When searching for $\mathcal{T}_i$ we need a metric $\mathcal{M}(\mathcal{T}_i, \mathbf{D})$ to tell us when to insert a threshold value $t_{i,j}$ between two data values that occurred for variable $V_i$. If our metric satisfies a set of

**Fig. 9.** Examples of Bayesian network structures. (a) A sampling type Bayesian network structure, with partitioning of predictor space. (b) Two examples of diagnostic Bayesian networks with overlapping relevant predictor subsets of size 3.




(a) (b)

**Table 5.** Examples 4.5, 4.6 and 4.7. (a) The observed data and the resulting discretizations, $\mathcal{R}_i = [0, 9]$. (b) The scores of different discretizations for $K \in \{2, 3\}$ using $\mathcal{M}_{2pc}(\mathcal{T}_i, \mathbf{D})$.

| $L_{i-1}^{>0+}$ data | 0 | 0.1 | 0.3 | 8 |
|---|---|---|---|---|
| The corresponding value of $S_i$ | few | few | many | many |
| Equal-width, $K = 3$ | 0 | 0 | 0 | 2 |
| K-means, $K = 3$ | 0 | 0 | 1 | 2 |
| $\mathcal{M}_{2pc}(\mathcal{T}_i, \mathbf{D})$, $K = 2$ | 0 | 0 | 1 | 1 |

(a)

| K | $\mathcal{T}_i$ | $\mathcal{M}_{2pc}(\mathcal{T}_i, \mathbf{D})$ |
|---|---|---|
| 2 | $[0.05]$ | 5.66296 |
| | $[0.2]$ | 4.68213 |
| | $[4.15]$ | 5.66296 |
| 3 | $[0.05, 0.2]$ | 4.96981 |
| | $[0.05, 4.15]$ | 5.66296 |
| | $[0.2, 4.15]$ | 4.96981 |

(b)

technical requirements, especially the requirement of decomposability, we can find the metric-optimal $\mathcal{T}_i$ by dynamic programming. Several metrics that meet these criteria have been proposed in the literature. (See [10] for an overview).

In our empirical studies we have chosen to use an information-theoretic metric $\mathcal{M}_{2pc}(\mathcal{T}_i, \mathbf{D})$ (the DL evaluation function of [10])

$$\begin{aligned} &\mathcal{M}_{2pc}(\mathcal{T}_i, \mathbf{D}) = \\ &\log |V_i| + \log \binom{|V_i| - 1}{K - 1} \\ &- \sum_{j=1}^{K} \log \frac{\Pi_{k=1}^{r_c}(\gamma \cdot (\gamma + 1) \cdot \cdots \cdot (\gamma + n(c_k, j) - 1)}{(r_c\gamma) \cdot (r_c\gamma + 1) \cdot \cdots \cdot (r_c\gamma + n_j - 1)}, \end{aligned} \tag{11}$$

where $|V_i|$ is the number of different values $V_i$ has, $r_c$ is the number of values our discretized focus of prediction has, $j$ goes over all $K$ categories of our discretization, $n_j$ is the number of original values assigned to category $j$, and $n(c_k, j)$ is the number of times context $c_k$ (predicted variable's value) occurs within that category, i.e. the number of times the original values assigned to discrete category $j$ occur in context $c_k$ in $\mathbf{D}$. The predicted variable needs to be discretized first, if it is continuous. We did this by K-means discretization in our empirical studies. The parameter $\gamma$ is a prior on the occurrences of contexts within each category. We have used $\gamma = 1$ at all times, i.e. we pretend to having observed one occurrence of each context prior to looking at $\mathbf{D}$, the actual observations.

In information-theoretic terms, this metric calculates the cost of using a *two-part code* as an encoding of the discretization (see e.g. [4] for more on two-part codes). Intuitively speaking, we first encode the number of different original values $V_i$ has, which can be done with cost $\log |V_i|$. Then the positions of $K - 1$ split points from among the $|V_i| - 1$ possible candidates are added to the code. Finally, we encode the distributions of contexts within each category, using sampling with replacement as our model. It should be kept in mind that our aim is to *minimize* $\mathcal{M}_{2pc}(\mathcal{T}_i, \mathbf{D})$.

Note especially that this discretization method allows automatic determination of a metric-optimal $K$. In our empirical studies, we let the method determine the optimal $K$ within range $[2, 10]$.

**Example 4.7.** *Let our observed data and $K$ be as in example 4.5. We search over all possible discretizations, letting the number of categories $K$ be either 2 or 3. Table 5(b) shows the scores. You can see that the context makes the metric prefer (i.e. give smaller scores to) discretizations which make the resulting categories internally as homogeneous with respect to context (value of $S_i$) as possible, i.e. discretizations with a split point between original values 0.1 and 0.3 are preferred. Of these, the one with only two categories ($K = 2$) is preferred over the three-category case, since the splitting of the second category is superfluous according to the metric, adding unnecessary complexity. Our discretization of the original values is shown in Table 5(a).*

### 4.6. The loss function

Since, from the point of view of decision making, an important goal in the analysis of wild salmon populations is the maintaining of biodiversity, we should be conservative in our

predictions: to err on the positive side (predicted value is greater than the correct value) is more serious than erring on the negative side (predicted value is smaller than the correct value). In other words, our *loss function* should be asymmetric.

We have used loss functions of the following type:

$$\mathcal{L}(y_p, y_c) = |y_p - y_c|^{\alpha}, \tag{12}$$

where $y_p$ is our prediction, $y_c$ is the correct value, and $\alpha$ controls the steepness of our penalization for error. For the goal of biodiversity maintenance, we used an asymmetric loss function

$$\mathcal{L}_{asymm}(y_p, y_c) = \begin{cases} |y_p - y_c|^{\alpha_1} & \text{if } y_p > y_c, \\ |y_p - y_c|^{\alpha_2} & \text{otherwise.} \end{cases} \tag{13}$$

That is, we employ a different error exponent for the cases where our model is optimistic ($\alpha_1$) vs. pessimistic ($\alpha_2$). And since we want to be conservative, $\alpha_1 \geq \alpha_2$ always holds. To summarize the loss of a series of predictions made in a validation set, we take the average of the losses of individual predictions, i.e. a loss incurred by an erroneous prediction is treated equally independently of the moment in time it occurs at.

**Example 4.8.** *Let our focus of prediction be $S_i$, and our loss function be symmetrical absolute difference, i.e. $\alpha_1 = 1$ and $\alpha_2 = 1$. We are searching for the best predictive Bayesian network structure using an empirical criterion. We have a set of "second-order" training data, and a set of test data to assess the predictive performance of the models. At the moment we have two structures to consider, $B_{S_1}$ and $B_{S_2}$.*

*If a series of correct values for $S_i$ is $[3000, 5000, 10000]$ and model $B_{S_1}$ is optimistic and predicts $[7000, 6000, 10000]$, our loss is on average $(4000+1000+0)/3 = 1670$. Model $B_{S_2}$ on the other hand is pessimistic and predicts $[1000, 2500, 8500]$, incurring average loss $(2000 + 2500 + 1500)/3 = 2000$. Hence, the optimistic model is preferred. Note that whereas $B_{S_2}$ always errs, its errors seem to be bounded. On the other hand $B_{S_1}$ only makes one serious error, but that error results in gross overestimation.*

*Let now $\alpha_1 = 2$, i.e. we penalize for optimistic predictions. With the same correct values and predictions our average loss is now $(4000^2+1000^2+0)/3 = 5666666.67$ for $B_{S_1}$, whereas for $B_{S_2}$ the loss is $(2000 + 2500 + 1500)/3 = 2000$ as earlier, i.e. using the optimistic model incurs nearly $3000$ times as much loss as using the pessimistic model now.*

*We can visualize the different nature of predictive models as shown in Fig. 10, where the $x$ axis shows the correct values and the $y$ axis the predictions of the models in question.*

In the case of the marginal likelihood criterion described earlier, the loss function minimized is

$$\mathcal{L}_{log}(P, y_c) = -\log P(y_c), \tag{14}$$

where $P(Y)$ is the predictive distribution. Intuitively speaking, marginal likelihood seeks for the model whose predictive distribution is the "closest" one to a "correct" one. The advantage of an empirical search criterion is that an arbitrary loss function different from $\mathcal{L}_{log}(P, y_c)$ can be used already in the model selection phase, although a consequence is that the model selection procedure does not lie within the Bayesian framework any more.

## 5. Empirical results

In the following we describe the results of predicting the wild smolt production using the data available for rivers Simo and Tornio (the Finnish side). Our data consist of time series of 17 and 15 years of data, respectively. At the highest level we split our results into three classes:

- The results of learning our model from the data for the Finnish side of river Tornio, validating the model by predicting for river Simo.
- The results of learning our model from the data for river Simo, validating the model by predicting for the Finnish side of river Tornio.
- The results of learning our model from the data for both rivers, and validating the models by splitting the data randomly to 80% of training data and 20% of validation data. The splitting was done 50 times. The particular split percentages chosen were based on the smallness of the data set: a 50-50 split would provide too little training data, whereas if the choice were skewed, say 95-5, the one or two vectors left to the test set might differ from the training set too much.

The first two types of result validate our models by using one of the rivers as the training set and the other one as the validation set. This enables us to assess to some extent whether our models are transferable across rivers, i.e. we check against overfitting to a particular river. The last type shows the results of assuming the data for different rivers is compatible.

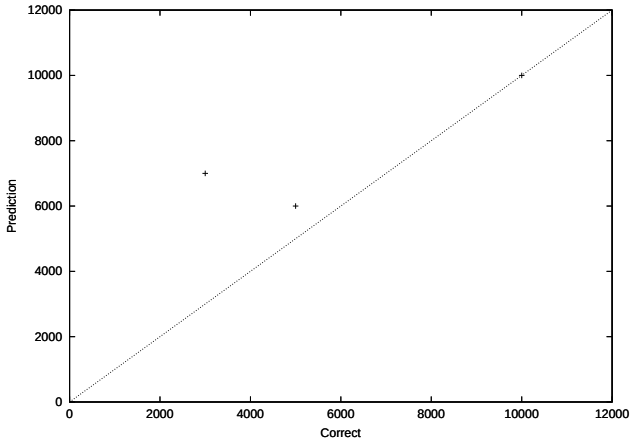
For each of these lines of study, we let the length of history available to us extend to 5 years. Given a training data set and a validation data set as described above, we proceeded in the following fashion:

1. We split the domain to subdomains as follows:
   (a) Only densities as predictors.
   (b) Only densities and smoltification age data as predictors.
   (c) Only densities and catch data as predictors.
   (d) Only densities and reproduction indices as predictors.
   (e) All of the available data as predictors.

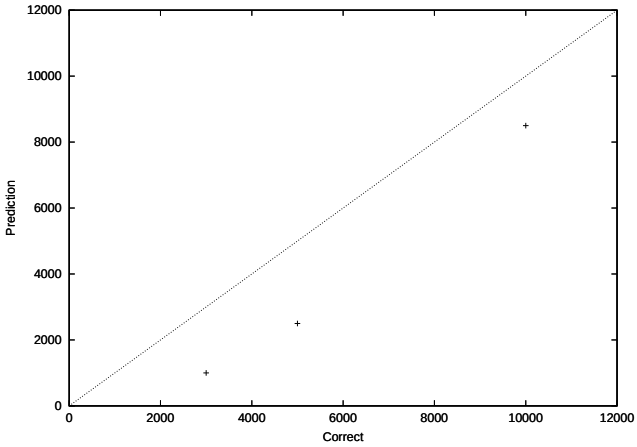
   Furthermore, we compared using domain expert-given density estimates to using average length-class densities.

   This domain splitting was done to study the dependency of predictive performance on the amount and quality of knowledge about the domain.
2. We tried each candidate from a set of model classes, i.e. subsets of possible Bayesian network structures, together with a set of discretization schemes. The structure types we tried were:
   (a) Sampling-type structures, no variable selection.
   (b) Sampling-type structures, with variable selection.
   (c) Sampling-type structures, partitioning structure.

**Fig. 10.** Example 4.8. The points are predictions made by the model for a given correct value. The line depicts a perfect predictor, i.e. predictions which are equal to the correct value. (a) $B_{S_1}$, the optimistic model. (b) $B_{S_2}$, the pessimistic model.




(a)




(b)

(d) Mixtures of diagnostic structures with 1 to 3 arcs per component.
(e) Diagnostic structures of 1 to 3 arcs, with variable selection.

Our arsenal of discretization methods was:

(a) Equal-width discretization with 2 to 5 categories.
(b) K-means discretization with 2 to 5 categories.
(c) $\mathcal{M}_{2pc}(\mathcal{T}_i, \mathbf{D})$, letting the method choose the optimal number of categories from within the range $[2, 10]$.

3. Given a set of structures and a discretization scheme, we sought for the structure describing the domain the best via two criteria:
   (a) The marginal likelihood criterion.
   (b) An empirical criterion, where the search was started from an empty graph, proceeding as in Chapter 4.4.1. Repeatedly, a random arc operation was picked, either an addition or a removal. The training data was then split randomly to 80% of "second-order" training data and 20% of test data. The parameters of the model were learned both prior and after the operation from the second-order training data, proceeding to measure the predictive performance of the pre- and post-operation models on the test data (the 20% part). To ensure the representativeness of the test set, the random splitting to a second order training set and a test set was performed 50 times. If the performance was better after the arc operation, the operation was performed, and a new one picked randomly. The number of arc operations tried for a particular second-order training data–test data pair was 1000.
4. Given the resulting model, we measured its predictive performance using $\mathcal{L}_{asymm}(y_p, y_c)$, with the symmetrical case $\alpha_1 = 1$, and the optimism-penalizing one ($\alpha_1 = 2$). $\alpha_2 = 1$ always.

### 5.1. Normalization of data

Our first approach was to use the data as it is, but we soon realized that the two rivers have quite different magnitudes, making transfer of knowledge using absolute values impossible. A model trained on Simo would never have seen numbers of the magnitude of Tornio, and thus would always shoot low and v.v.

Hence we decided to normalize our data using

$$\mathcal{N}_i = \frac{v_{i,j} - \mu_i}{\sigma_i} \tag{15}$$

where $v_{i,j}$ is an original value of variable $V_i$, $\mu_i$ is the empirical mean of $V_i$ and $\sigma_i$ is the empirical standard deviation of $V_i$. Note that to stay absolutely honest, $\mu_i$ and $\sigma_i$ have to be calculated from the training data alone. Otherwise they will provide quite a lot of information about the validation set, given our scanty data. Also, if an empirical criterion is used, the normalization parameters have to be determined from the second-order training data as well.

Normalization produces a model that only speaks of things in relative terms, but note that we can always translate the predictions of our model back to absolute values, provided we obtain $\mu_i$ and $\sigma_i$ somehow. They need not be determined empirically: given a river with no data on smolt production a biologist or fishery scientist can hypothesize about the mean and standard deviation of smolt production in the river, plug the values in and see the absolute values. Most importantly, from the managerial point of view, relative values suffice for the qualitative analysis of changes in the population over time.

### 5.2. Presentation of results

Since our results indicated that in the sampling paradigm variable selection usually paid off, and on the other hand the partitioning networks as well as diagnostic structures of more than one arc performed poorly (most likely due to the small size of the data set, leading to drastic overparameterization when using these model classes), and one-arc diagnostic structures were too impoverished to possess predictive potential, we present only the results obtained in the sampling paradigm using variable selection.

We have reported the results of variable selection in the sampling paradigm by two means: using marginal likelihood as the criterion and $\alpha_1 = 1$ as the loss function, and using an empirical criterion with two loss functions, one symmetric ($\alpha_1 = 1$), the other penalizing for optimistic predictions ($\alpha_1 = 2$). Since the choice of a set of structure candidates and discretizations would ideally be done by a biologist or a fishery scientist, we show as a baseline indicator of performance the model (from among our fixed set of discretization schemes) with the best predictive performance measured by a particular loss function.

### 5.3. Learning a model from river Tornio, validating by river Simo

Despite the fact that river Tornio is a much larger river compared to river Simo, our normalization allowed our models to predict for the smaller river reasonably well when taught on the data for the bigger one.

Figures Fig. 11(a) - Fig. 11(d) show the results of seeking via marginal likelihood. All variables were eligible. It can be seen that both types of density estimates for year $i-5$ are relevant, albeit feebly so. With respect to domain expert-estimated densities, the density of $> 0+$ parr in the previous year is quite relevant, whereas for average length-class densities their density three years earlier is weakly relevant. For all of the other variables the models agree, emphasizing data for year $i-3$. The reproduction indices of both kinds three years back are relevant, whereas the raw catch numbers are only weakly so. Somewhat surprisingly, M74 percentages in the previous year are considered helpful, as are smoltification age estimates for smoltification at 3 and 4 years three years earlier.

Figures Fig. 11(e) - Fig. 11(h) show the results of seeking for a model using our empirical criterion. It can be seen that the results are poorer than with marginal likelihood. Possibly, despite our efforts at avoiding overfitting, the empirical models have learned patterns which are more suited to river Tornio.

Fig. 12 shows an example of restricting the available variables to a subdomain. In this case we study the use of a reproduction index in numbers together with density estimates. The estimates of both expert-estimated and average length-class densities are very similar, no doubt due to the low relevance of the average length-class variables selected. In this case the best discretization scheme favoured the index of reproduction in numbers four years back in time. The number of smolts five years earlier was selected as well: studying the behaviour of our models we saw that broadly speaking there exists a negative correlation between $S_i$ and $S_{i-5}$, due to the periodic nature of smolt production in our data.

### 5.4. Training on Simo, validating by Tornio

We also reversed roles, learning the model from the normalized data for the smaller river, trying to predict for the larger river. The results were slightly poorer than the other way around, and a noticeable feature in all subdomains and with both criteria was that the models tended to be optimistic if domain experts' estimates of densities were used, except in the case of having been learned using an empirical criterion favouring pessimism.

Figures Fig. 13(a) - Fig. 13(d) show the results of seeking via marginal likelihood. All variables were eligible. This time the best discretizations produced quite different models for the different types of density estimates. It is noticeable how average length-class densities did not get picked at all. In all subdomains, only the average length-class densities of older parr at year $i-4$ were considered relevant. Reproduction indices and smolt production in the past got picked again, as well as data on M74.

Note how the predictions made using length-class estimates (which actually were irrelevant) are pessimistic compared to the optimistic estimates made using expert estimates.

Figures Fig. 13(e) - Fig. 13(h) show the results of seeking for a model using our empirical criterion. While the use of a loss function penalizing for optimism seems to work using both types density estimates, it seems like the procedure penalized for optimism too steeply: the models tend to be overwary.

Fig. 14 shows an example of restricting the available variables to a subdomain. In this case we study the use of a reproduction index in kilos together with density estimates. It can be seen that using a symmetrical loss function when selecting a model produces an optimistic model, whereas making use of pessimism-gratifying scoring for models clearly works, although the resulting model might be too cautious. Note the drastic difference in discovered structures: seeking for pessimistic models cuts down the number of predictors from 16 to 3. Also, while the model learned using a symmetric loss function made use of all $R_i^k$ variables bar the one for year $i-4$, the pessimistic model drops the reproduction indices altogether.

### 5.5. Learning a model from the combined data

Finally, we put the normalized data of both rivers together.

In the scheme where we learn the model from the data for one river and validate by predicting for another, the fact that the data sets come from separate biological systems provides a strict validation procedure. When we use the combined data, we have to do validation by repeated artificial separations to 80% of training and 20% of validation data. Due to this less stringent validation scheme, we also show in Fig. 15 the structure with the highest marginal likelihood over all discretization schemes, regardless of the predictive performance.

It can be noted that reproduction stage variables dominate in relevance. Smoltification age variables drop to a marginal role, and M74 is altogether missing. Smolt production in the past is considered quite relevant as well. Of the density estimates, only expert estimates for older parr five years back are picked, the length-class estimates not at all.

The results of doing repeated 80-20 validation are shown in Fig. 16 for both criteria. When using marginal likelihood, the best models for both types of density estimates dropped the density variables altogether, arriving at the same model, which

we show in Fig. 16(a) and Fig. 16(b). The reproduction stage data is highly relevant again, while smoltification age data is slightly more relevant than in the maximum marginal likelihood structure. The predictive performance of the model is quite good, barring the outlying correct value at the high end of the range.

Figures Fig. 16(c) to Fig. 16(f) show how using an optimism-avoiding model selection criterion affects the results drastically, pushing most of the predictions to the pessimistic side of the line depicting perfection.

## 6. Conclusions

We have defined and demonstrated a methodology for the prediction of smolt production in the probabilistic framework. Our goals were managerial, aiming at generalizing models capable of adjusting to the needs of fisheries management, while maintaining good predictive performance. Our empirical results illustrated the performance of our methodology on real-world data. In the interest of unbiased evaluation, our validation schemes were as strict as possible.

When learning the model from the data for one river and predicting for the other, it was seen that the results were somewhat different. When learning models from the data for river Tornio, a large river, the models performed reasonably well when predicting for river Simo. The other way around the models tended to be overoptimistic, however, unless penalized steeply for optimism in the model selection phase. Possibly there is more information (with respect to the prediction of smolt production) in the data set for river Tornio.

When we combined the data, and then validated by repeated splitting to 80% of training and 20% of validation data, the results were quite good, although the combination of density estimates from the two rivers did not seem to work: in marked contrast to the river vs. river scheme, no density estimates of either kind were included in the best models. Possibly the density estimates of the two rivers differ in some essential way in addition to magnitude, which we handled by normalization.

Overall, comparing the two kinds of density estimates, domain expert-given and average length-class densities, it could be seen that the expert estimates worked much better, the length-class estimates being completely irrelevant in some cases. Possibly the inclusion of biological knowledge in the expert estimates helps. However, there remains the fact that when the data was combined, the expert estimates for the two rivers, when combined, did not seem be relevant. There seems to be something river-specific about the link between expert-made density estimates and estimates of smolt production.

Comparing different model selection criteria, it was seen that marginal likelihood produced nicer, nearly linear, correlations between correct and predicted values, but sometimes tended to produce overoptimistic models. When learning a model using an empirical criterion penalizing for optimism, the resulting models were seen to avoid optimism in the validation data as desired, although they tended to exaggerate pessimism at the cost of the correct value–predicted value correlation.

Looking at the predictive performance of the models in general, the results were reasonably good, except for the outlying correct values at the high end of the range. That our models failed to predict correctly for these values which are several orders of magnitude bigger than the rest of the data, is really an indication of the nature of the data. Even in the river vs. river scheme, having seen only one example of such values in the data, a generalizing model should not really be expected to be able to predict correctly for the one similarly outlying case in the validation set, keeping in mind that our rivers are separate biological systems, and the results and the data indicate that they are somewhat different as well.

In our empirical results we studied how informative the real-world data sets we used are. If desired, biological knowledge could also be made use of, e.g. in the choice of possible model structures, in the determination of parameter priors or in the choice of threshold values in the discretization. Another direction for future work would be to utilize more complex loss functions than the relatively simple asymmetrical loss function used here. For example, the steepness of the penalization for error could depend on the correct value, e.g. if the population is actually “large”, errors are less serious than when the population is on the verge of extinction.

*Acknowledgements*

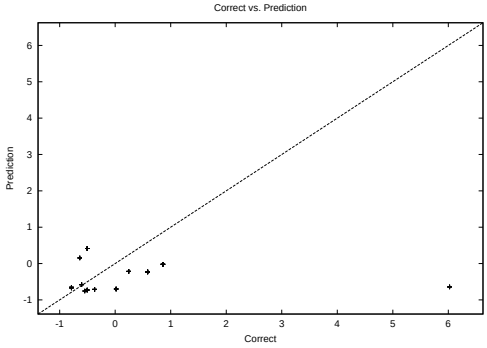
This paper has been funded by EU project nr 99/064 ”Probabilistic modelling of Baltic salmon stocks”.

## References

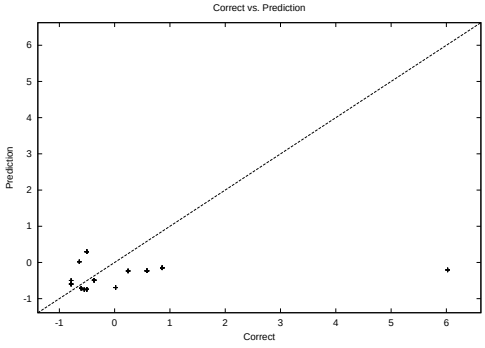
1. W. Buntine. Theory refinement on Bayesian networks. In B. D’Ambrosio, P. Smets, and P. Bonissone, editors, *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pages 52–60. Morgan Kaufmann Publishers, 1991.
2. A.P. Dawid. Properties of diagnostic data distributions. *Biometrics*, 32:647–658, 1976.
3. N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.
4. P. Grünwald. *The Minimum Description Length Principle and Reasoning under Uncertainty*. PhD thesis, CWI, ILLC Dissertation Series 1998-03, 1998.
5. D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, September 1995.
6. P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. On supervised selection of Bayesian networks. In K. Laskey and H. Prade, editors, *Proceedings of the 15th International Conference on Uncertainty in Artificial Intelligence (UAI’99)*, pages 334–342. Morgan Kaufmann Publishers, 1999.
7. P. Kontkanen, P. Myllymäki, and H. Tirri. Classifier learning with supervised marginal likelihood. In J. Breese and D. Koller, editors, *Proceedings of the 17th International Conference on Uncertainty in Artificial Intelligence (UAI’01)*. Morgan Kaufmann Publishers, 2001.
8. P. Kontkanen, P. Myllymäki, and H. Tirri. Comparing prequential model selection criteria in supervised learning of mixture models. In T. Jaakkola and T. Richardson, editors, *Proceedings of the Eighth International Conference on Artificial Intelligence and Statistics*, pages 233–238. Morgan Kaufmann Publishers, 2001.
9. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA, 1988.

10. J. Rousu. *Efficient Range Partitioning in Classification Learning*. PhD thesis, Report A-2001-1, Department of Computer Science, University of Helsinki, January 2001.
11. K. Valtonen, T. Mononen, P. Myllymäki, H. Tirri, J. Erkinaro, E. Jokikokko, S. Kuikka, A. Romakkaniemi, L. Karlsson, and I. Perä. A study of electrofishing bias in terms of habitat and abundance using information-theoretic tools. Unpublished manuscript., 2002.

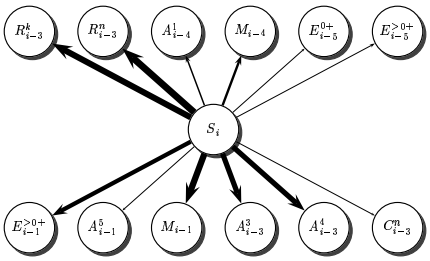
**Fig. 11.** Learning a model from the data for the Finnish side of river Tornio, predicting for river Simo. Sampling-type structure with variable selection. Densities and all other data as predictors, history of five years. (a) - (d): Structure search by marginal likelihood over different discretizations. Structure with the best predictive performance using $\alpha_1 = 1$ shown. (a) and (c) Estimated densities. (b) and (d) Average length-class densities. (e) - (h): Structure search by an empirical criterion over different discretizations. Models with the best predictive performance for $\alpha_1 = 1$ and $\alpha_1 = 2$ shown. (e) Estimated densities, $\alpha_1$ = 1. (f) Estimated densities, $\alpha_1$ = 2. (g) Average length-class densities, $\alpha_1$ = 1. (h) Average length-class densities, $\alpha_1$ = 2.



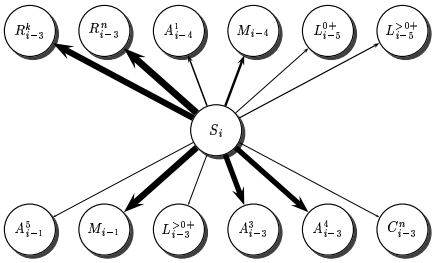

(a)



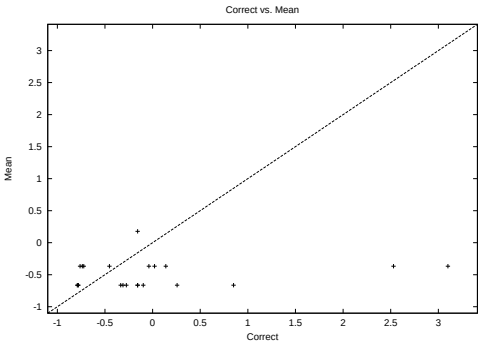

(b)



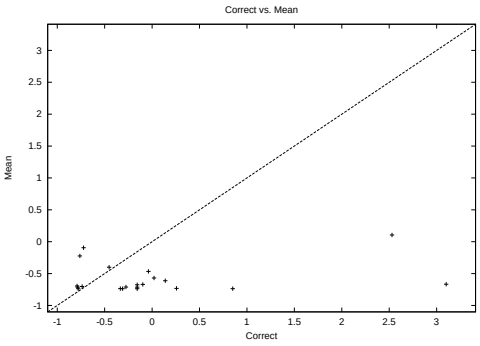

(c)



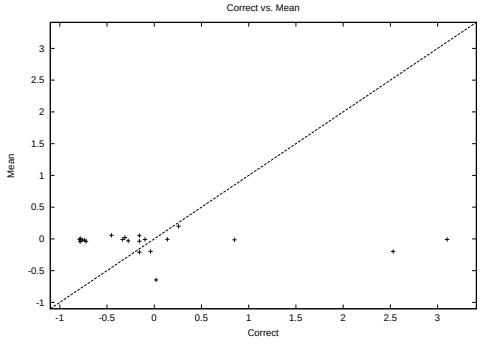

(d)



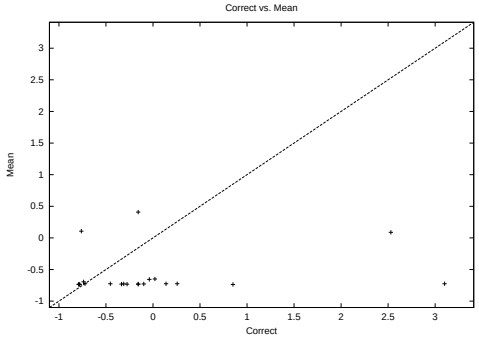

(e)




(f)

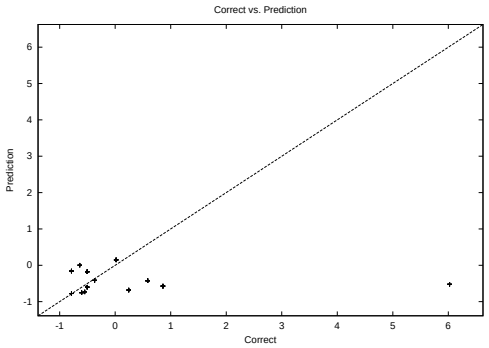



(g)

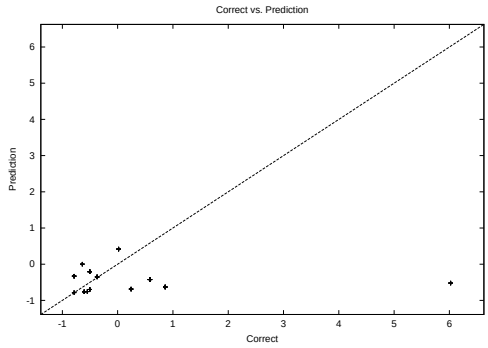



(h)

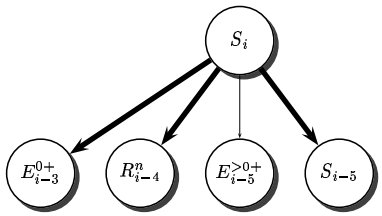
**Fig. 12.** Learning a model from the data for the Finnish side of river Tornio, predicting for river Simo. Sampling-type structure with variable selection, structure search by marginal likelihood over different discretizations. Structure with the best predictive performance using $\alpha_1 = 1$ shown. Densities and $R_i^n$ as predictors, history of five years. (a) and (c) Estimated densities. (b) and (d) Average length-class densities.



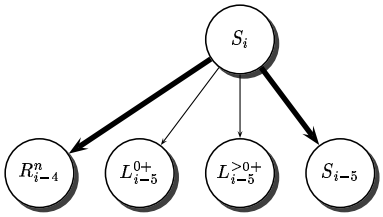

(a)




(b)

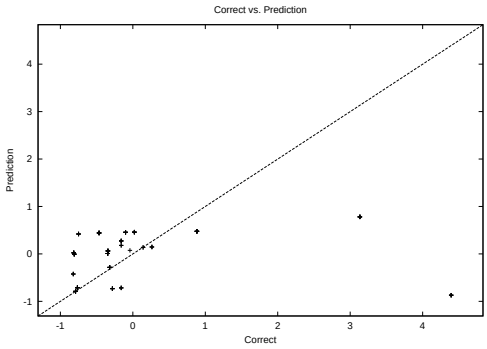



(c)

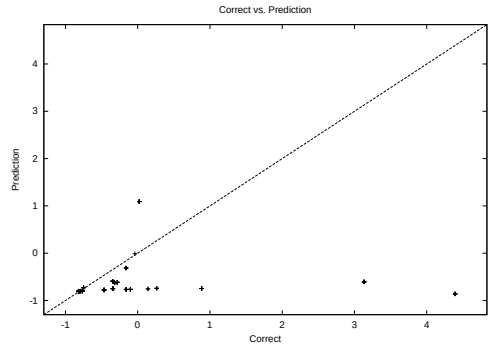



(d)

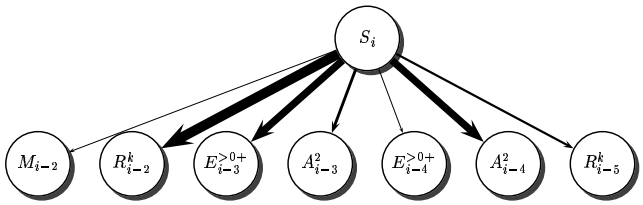
**Fig. 13.** Learning a model from the data for river Simo, predicting for the Finnish side of river Tornio. Sampling-type structure with variable selection. Densities and all other data as predictors, history of five years. (a) - (d): Structure search by marginal likelihood over different discretizations. Structure with the best predictive performance using $\alpha_1 = 1$ shown. (a) and (c) Estimated densities. (b) and (d) Average length-class densities. (e) - (h): Structure search by an empirical criterion over different discretizations. Models with the best predictive performance for $\alpha_1 = 1$ and $\alpha_1 = 2$ shown. (e) Estimated densities, $\alpha_1$ = 1. (f) Estimated densities, $\alpha_1$ = 2. (g) Average length-class densities, $\alpha_1$ = 1. (h) Average length-class densities, $\alpha_1$ = 2.



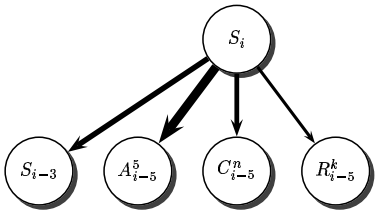

(a)



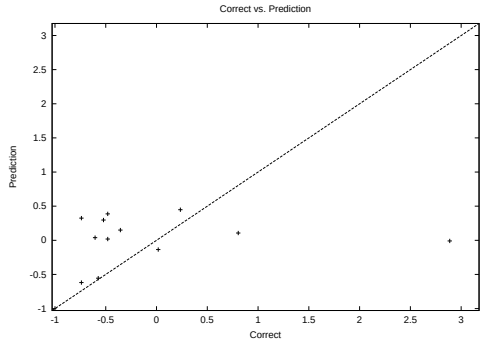

(b)



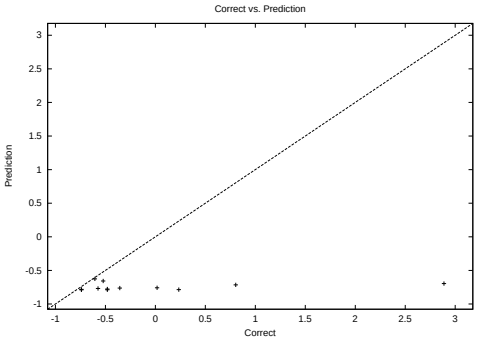

(c)



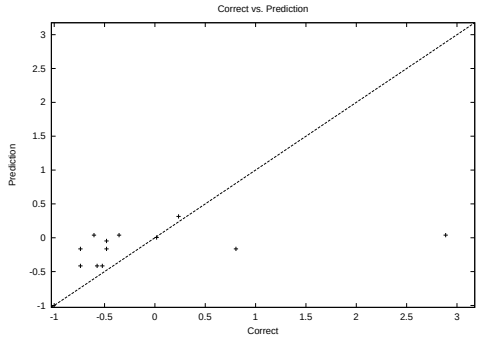

(d)



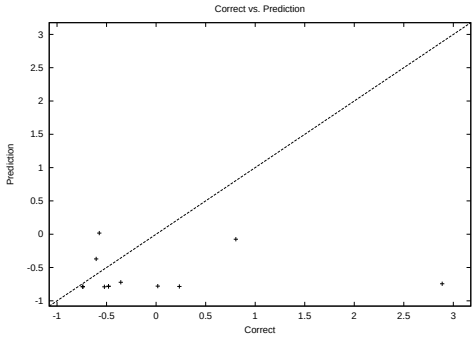

(e)




(f)

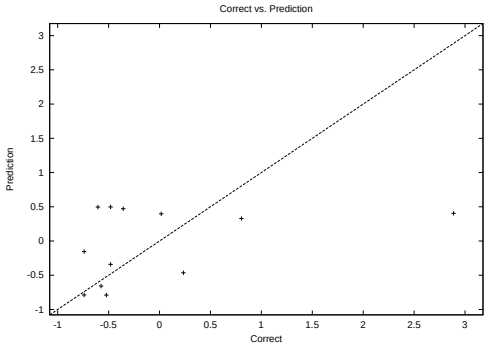



(g)

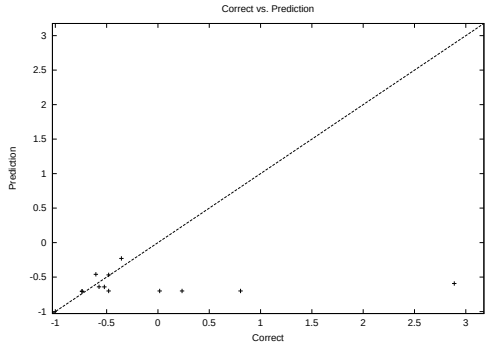



(h)

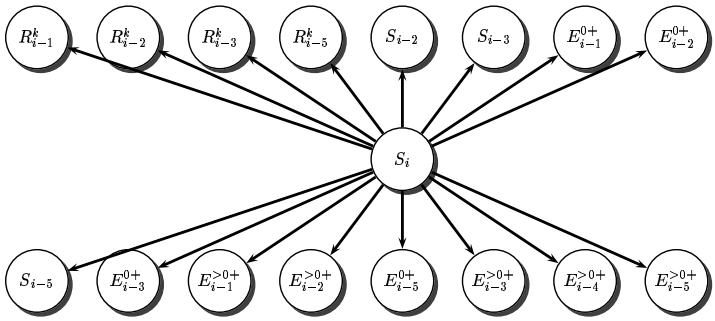
**Fig. 14.** Learning a model from the data for river Simo, predicting for the Finnish side of river Tornio. Sampling-type structure with variable selection, structure search by an empirical criterion over different discretizations. Structures with the best predictive performance for $\alpha_1 = 1$ and $\alpha_1 = 2$ shown. Densities and $R_i^k$, history of five years, predictive performance. (a) and (c) Estimated densities, $\alpha_1$ = 1. (b) and (d) Estimated densities, $\alpha_1$ = 2.



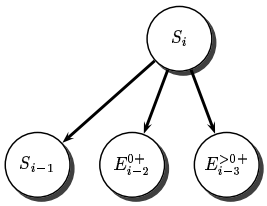

(a)




(b)




(c)




(d)

**Fig. 15.** Learning a model from the combined data for the two rivers. Sampling-type structure with variable selection, structure search by marginal likelihood. All of the available data as predictors, history of five years. Structure with the highest marginal likelihood over all discretizations.

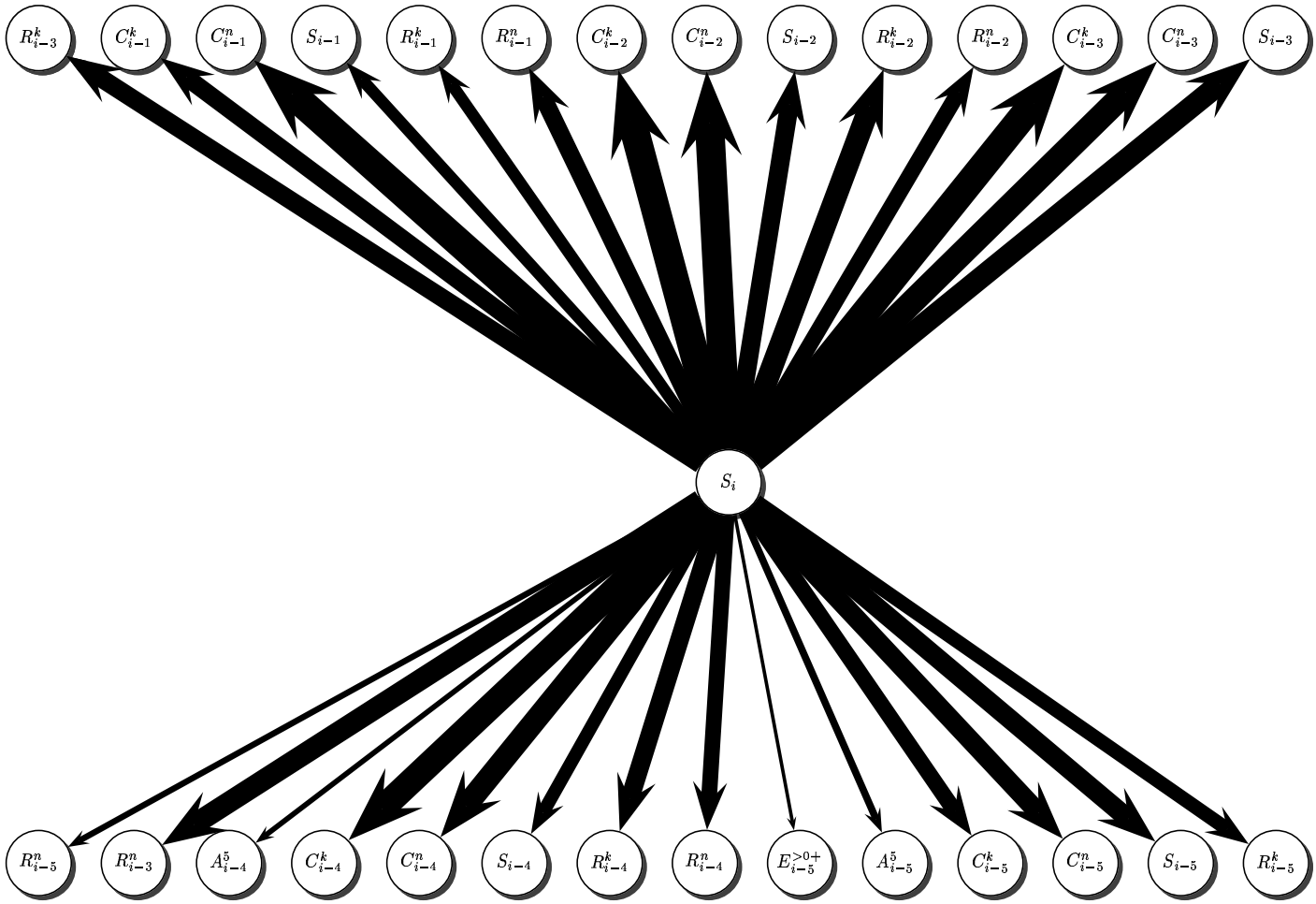
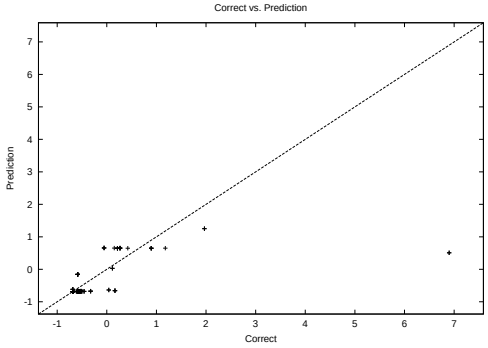
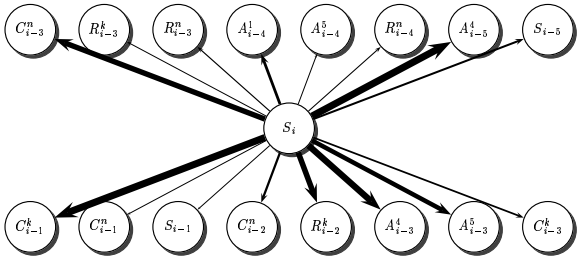


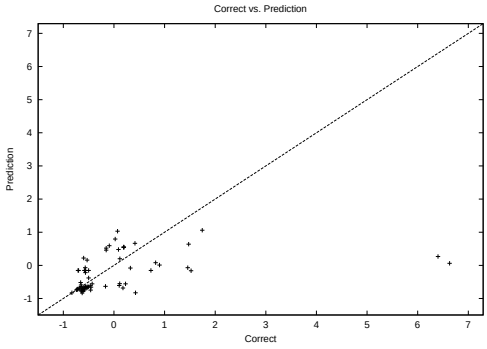
**Fig. 16.** Learning a model from the combined data for both rivers, Validating by 80-20 splits to training and validation data. Sampling-type structure with variable selection. Densities and all other data as predictors, history of five years. (a) - (b): Structure search by marginal likelihood over different discretizations. Model with the best predictive performance using $\alpha_1 = 1$ shown. (a) and (b) All data. (c) - (f): Structure search by an empirical criterion over different discretizations. Models with the best predictive performance for $\alpha_1 = 1$ and $\alpha_1 = 2$ shown. (c) Estimated densities, $\alpha_1$ = 1. (d) Estimated densities, $\alpha_1$ = 2. (e) Average length-class densities, $\alpha_1$ = 1. (f) Average length-class densities, $\alpha_1$ = 2.



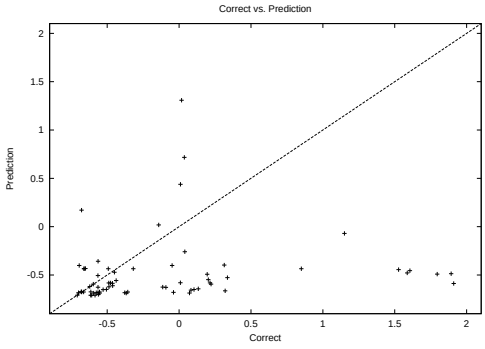

(a)



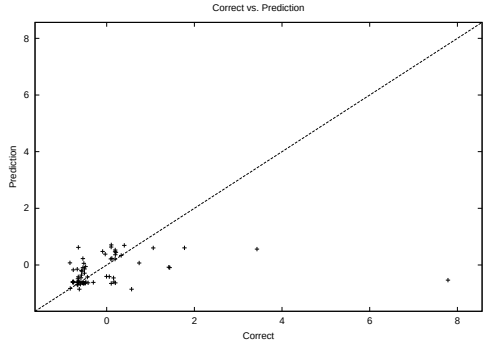

(b)



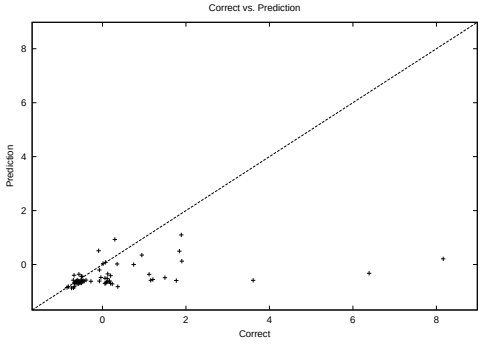

(c)




(d)




(e)




(f)