Research Report
Department of Statistics
Göteborg University
Sweden

Monitoring a Freshwater Fishpopulation: Statistical Surveillance of Biodiversity

Magnus Pettersson

Research Report 1997:2
ISSN 0349-8034

Mailing address:	Fax	Phone	Home Page:
Department of Statistics	Nat: 031-773 12 74	Nat: 031-773 10 00	http://www.stat.gu.se
Göteborg University	Int: +46 31 773 12 74	Int: +46 31 773 10 00	
S-411 80 Göteborg			
Sweden			

Monitoring a Freshwater Fishpopulation: Statistical Surveillance of Biodiversity

Magnus Pettersson
Department of Statistics, Göteborg University
S-411 80 Göteborg, Sweden
E-mail: Magnus.Pettersson@statistics.gu.se

April 14, 1997

Abstract

Statistical surveillance are methods for repeated analysis of stochastic processes, aiming to detect a change in the underlying distribution. Such methods are widely used for industrial, medical, economical and other applications. By applying these general methods on data collected for environmetrical purposes, it might be possible to detect important changes fast and reliable.

We exemplify the use of statistical surveillance on a data set of fish catches in Lake Mälaren, Sweden, 1964-1993. A model for the in-control process of one species, vendace (*Coregonus albula*), is constructed and used for univariate monitoring. Further, we demonstrate the application of Hotelling's T^2 and the Shannon-Wiener index for monitoring biodiversity, where a set of five economically interesting species serve as bioindicators for the lake.

Keywords:

Vendace, Recursive Residuals, Shewhart test, AR process, Fourier series, Species correlation matrix, Shannon-Wiener index, Hotelling's T^2, Lake Mälaren, Catch data

1. Introduction

There is a growing interest for studying unwanted changes of mother earth's environment which is creating new opportunities for people dealing with environmental data. More often politicians, biologists and others try to find out if changes in our environment has occurred, by monitoring one or more variables of ecological interest over time. Hot topics include for example global warming, deterioration of water or soil quality, increasing incidence of cancer diseases caused by environmental factors and changes in biodiversity. The increasing awareness and interest for the status of the environment has given rise to a large amount of data collections. However, there is a risk that data is only collected, stored away and not followed in a systematic way. By using the theory for statistical surveillance it might be possible to design procedures for monitoring changes in the environment, calling an alarm for a change in the system, fast and accurate.

Sometimes time dependence is tested with an ordinary test of hypothesis on a fix set of data, *i.e.* the null hypothesis, that no change has occurred over the studied time interval versus an alternative where the period can be divided into before and after the change. However, when we are monitoring data to be able to detect a possible change at an unknown time and make repeated analysis, we can not use an ordinary hypothesis testing approach. Instead techniques for surveillance have to be used. Terms used in this context are for example monitoring, change point detection, statistical process control (SPC) and statistical quality control.

This paper will give an introduction to the use of statistical surveillance in environmental science. We will study an application, from a data set on fish catches in lake Mälaren in Sweden, where we will be able to evaluate the usefulness of statistical methods in monitoring the environment. We will both study the detection of change in the level of one species, by using univariate monitoring procedures, and extend the model for monitoring the correlation between the species. The emphasis in the paper is on identifying a useful model that can be used to transform the data into a form where standard SPC methods can be applied. We will only use Shewhart tests on the transformed data, not because it is the optimal method, but because it is easy to apply and have properties that makes it suitable for benchmarking.

In Section 2 we will give an overview of statistical surveillance methods. In Section 3 and 4 a data set from Lake Mälaren is studied unsing different models for the data to show the impact of model selection. A background to the material is given below in Section 1.1. Section 3 focuses on one species, vendace (*Coregonus albula*), while Section 4 discusses application of multivariate methods on five species at the same time. Finally, Section 5 discusses the conclusions and ideas for further study.

1.1. Monitoring fish in Lake Mälaren

Data from the catches of fish, made by professional fishermen around Lake Mälaren in central Sweden are collected since 1964 by the four regional authorities surrounding the lake. Of the species living in the lake, six have a major economical interest: Burbot (*Lota lota*), eel (*Anguilla anguilla*), perch (*Perca fluviatilis*), pike (*Esox lucius*), pike-perch (*Lucioperca lucioperca*) and vendace (*Coregonus albula*). We will use five of them as indicators of the biodiversity in the lake and evaluate the performance of different monitoring procedures. The eel have been excluded since its population is depending on artificial breeding, and is therefore increasing over time. We will humbly avoid the discussion of the relevance of these specific species as bioindicators and concentrate on the statistical aspects of the problem.

From 1987 the catch of vendace decreased. At first, this declination was considered part of an assumed 6-8 year cycle of all fish in the lake, but when the expected increase did not occur in 1990 further action was taken. As we will see below, other period lengths than the assumed might fit better to the data. No systematic analysis of departure from the in control pattern have been performed earlier. We will study the data material, kindly provided by the Fresh Water Laboratory in Drottningholm, from different views. The aim is to discuss the performances of different surveillance procedures asking the question: *If we had done this way, what would have happened?*

The analysis described in this paper is based on the landed catches of fish made by professional, mostly part-time, fishermen. Since we lack information about the effort it is not possible to reconstruct that information, we will use the raw data for analysis. Assuming that these figures are correlated with the amount of each fish species, they will suffice.

2. Statistical surveillance

Often we have data arriving to us one-by-one or in groups at discrete time steps. When the system producing the sequence of measurements behaves in some natural or wanted way we say that it is "in control". At a stochastic time, τ, the system leaves the in control state and goes "out of control". The aim of the surveillance procedure is to detect when the system goes out of control, under some performance criteria, *e.g.* fixed false alarm probability at a certain time. In many cases *ad hoc* methods are constructed or data are viewed by an expert, who decides about whether to take action or not. Using methods of statistical surveillance makes the monitoring more accurate since the performance of the methods can be evaluated and different methods can be compared with each other.

Statistical methods for detecting changes in the underlying distribution of a sequence of data is used in many other applications. Examples from the industry can be found in Wetherill and Brown (1991), intensive care medicine in Frisén

(1992), surveillance of rare events in Arnkelsdóttir (1995), postmarketing surveillance of new drugs in Sveréus (1995), surveillance of business cycles in Frisén (1994b) and even from a court case in Charnes and Gitlow (1995). Examples of environmental application can be found in Berthoux *et al.* (1978) (sewage treatment plants), Kjelle (1987) (background gamma radiation), Settergren Sørensen and la Cour Jansen (1991) (water quality) and Vaughan and Russell (1983) (point sources of pollution).

Frisén and de Maré (1991) defined a monitoring situation to be either *passive* or *active* depending on the action taken after an alarm. When the process either is interrupted or changed after an alarm we have an active situation, while if the process proceeds unchanged after the alarm the situation is passive. For example, monitoring sea levels and floods will be passive, since no action is taken to alter the process after an alarm. On the other hand, when monitoring water or air quality, it is likely that an alarm will cause actions attempting to re-establish the earlier quality, in which case we define the situation to be active.

We have a process of stochastic variables, $X(t), t = 1, 2, \ldots$, which can be univariate or multivariate, *i.e.* $X(t)$ is a vector of dimension $p \times 1$. Further, we define the cumulated process up to time s as $X_s = \{X(t), t = 1, 2, \ldots, s\}$. This convention will be used consistently: Time within parenthesis meaning a value at a particular time and time in subindex meaning the cumulated values up to and including the time. *N.b.* that we will always use *discrete* time even if the notation with time in parentheses is customary for continuous time.

2.1. Critical event

At each time step, s, we will formulate two possible states, that we want great power for distinguishing between: $D(s) \Leftrightarrow \tau > s$ (in control) and $C(s)$ (out of control). Given the data, X_s, we will evaluate the evidence of $C(s)$ versus $D(s)$. Note that even if we will use a formulation similar to hypothesis testing this is not the case. The out of control alternative, $C(s)$, will be formulated differently for different applications, however, examples of often interesting formulations are:

- System goes out of control at time s, $C(s) = \{\tau = s\}$.
- System is out of control, $C(s) = \{\tau \leq s\}$.
- System went out of control less than d time steps ago, $C(s) = \{s - d < \tau \leq s\}$.

Different parameters can be of interest, as will be seen below, therefore the interesting out of control event have to be modelled in different ways. Here we describe statistical surveillance in the case of a change in the mean. If we let $\mu(t) = E(X(t))$ and when the system is in control $\mu(t) = \mu^0$, for $t < \tau$, some examples of descriptions of the critical event is:

- Sudden shift to a new level, *i.e.*

$$\mu(t) = \mu^A = \mu^0 + \delta, \text{ for } t \geq \tau,\ \delta \neq 0.$$

This model is satisfactory for many situations and, thanks to its simplicity, also suitable for evaluation of different monitoring procedures.

- A generalization of the model above is to let the shift take a number, ν, time steps to reach the new level. If we restrict μ to monotone functions on $[\tau, \tau + \nu)$ we get

$$\mu(t) = \begin{cases} g(t) \text{ when } \tau \leq t < \tau + \nu \\ \mu^A \text{ when } t \geq \tau + \nu \end{cases},$$

where $\mu^0 < g(t) < \mu^A$, $a < b \Rightarrow g(a) < g(b)$, when $a, b, t \in [\tau, \tau + \nu)$.

- Further, setting $\nu = \infty$ in the model above the mean will increase towards infinity. An example of this model is the linear trend beginning at τ.

In the alternatives described above the change at τ persists and the system will not return to the in control state. Below are three examples of alternatives where the system eventually returns to the in control state:

- Pulse lasting one time step,

$$\mu(t) = \begin{cases} \mu^0 \text{ when } t > \tau \\ \mu^A \text{ when } t = \tau \end{cases}.$$

- The model above can be extended to a pulse lasting ν steps.

- For some applications it is reasonable to say that the out of control event is a pulse that decays exponentially after τ, *i.e.*

$$\mu(t) = \mu^A \cdot e^{\beta(\tau - t)}, \text{ for } t \geq \tau.$$

The choice of critical event can have important effects on the performance of the surveillance procedure used (Sveréus (1995)).

2.2. Methods

Several methods for detecting the change in distribution have been designed. For univariate problems the first method was the Shewhart chart (Shewhart (1931)), followed by CUSUM (Page (1954)) and EWMA (Roberts (1959)). Bayesian approaches can be found in Zacks (1983). The likelihood ratio method (Shirayev

(1963) and Frisén and de Maré (1991)), which have maximum detection probability for a specific alternative and a false alarm probability. When the in control system contains unknown parameters that are updated, recursive residuals (Brown *et al.* (1975)) can be applied.

With the Shewhart method an alarm is triggered when the last observation exceeds a critical limit, *i.e.* when $|X(s)| > c$. The limit, c, is in traditional SPC literature set to $3.09 \cdot \sigma(s)$ or $3 \cdot \sigma(s)$, where $\sigma(t) = \sqrt{Var(X(t))}$ (Wetherill and Brown (1991)). CUSUM methods are based on the cumulative sums of observations, $S(t) = \sum_{i=1}^{t} X(i)$. An alarm criteria can be when the statistic $\max_{i=1,\ldots,t} |S(t) - S(i)|$ exceeds a limit $h + k \cdot (t - i)$, for some h and k. The EWMA, *Exponentially Weighted Moving Average*, uses a weighted sum of the observations as a test statistic, *i.e.* $U(t) = \sum_{i=0}^{t} \lambda^i X(t - i)$, for some $\lambda \in (0, 1)$, where a smaller value of λ gives more weight to more recent observations. For the likelihood ratio method the critical function, p, is defined as

$$p(X_s) = \frac{f_{X_s}(x_s \mid C(s))}{f_{X_s}(x_s \mid D(s))},$$

where both the $C(s)$ and $D(s)$ are fully specified states, $C(s)$ being the system is out of control and $D(s)$ in control, respectively.

For multivariate problems two natural strategies are either to monitor each process separately (an alarm is triggered at the first alarm of an individual process) or to transform the data into a univariate sequence. The likelihood ratio method can equally well be applied for a multivariate sequence as for a univariate one. A survey of methods for detecting changes in more than one variable can be found in Wessman (1996).

2.3. Evaluation criteria

Different measurements of performance have been suggested, see Frisén (1992) or Frisén (1994a) for a comprehensive description. Depending on the considered application we will have different demands on the performance. Widely used is the *average run length*, ARL^0, which is defined as the expected time from start to the first alarm given $\tau = \infty$. An other often used measure of run length is the ARL^1, the time from start to the alarm, given $\tau = 1$. Further a measure related to the run length is the *expected delay*, which is the expected time from τ to alarm.

In some cases it is necessary to detect the critical event within a certain time, d, or else it might be too late. An interesting measure for these cases is the *probability of successful detection*, $PSD(d)$. When we need to have control over the probability of having false alarms, we can use the *false alarm rate*, $\alpha(t)$, the probability of having a false alarm at t and the cumulated false alarm rate, α_t. When we have got an alarm, the probability of it being caused by a true change can be formulated in the *predictive value of an alarm*, $PV(t)$, defined as the

probability of a change given an alarm. Cost functions, *i.e.* measures that gives a cost to the event of a false alarm and a cost proportional to the delay of an alarm, $t - \tau$, can also be used.

2.4. Recursive residuals

To be able to detect a deviation out of control we have to specify what is considered to be in control. The in control model can also contain parameters with unknown values, that have to be estimated. Two natural ways for dealing with the unknowns are: Either we first run the system during a so called run-in period, estimate the parameters and thereafter treat them as known or we update the estimates in each time step using X_{s-1}. One example of the latter is to use recursive residuals (Brown *et al.* (1975)) where we compare the predicted value, $E\left[\widehat{X}(s) \mid X_{s-1}\right]$, against the real $X(s)$. The residuals are used for surveillance. The effect of using nearly the same data for both modelling and surveillance is interesting but will not be discussed further in this paper.

Instead of monitoring the process $\{X(t)\}$ we use the *residual process*

$$R(t) = X(t) - E(X(t) \mid X_{t-1}), \tag{2.1}$$

where the new process, $\{R(i)\}$, is monitored with some univariate method. An example is when the mean level is constant, but unknown, and is re-estimated at each time step, *i.e.*

$$R(t) = X(t) - \frac{1}{t-1}\sum_{i=1}^{t-1} X(i) = X(t) - \overline{X}_{t-1}.$$

Models of this kind have also been studied from a Bayesian approach by among others Zacks (1989). An other example is an $ARMA$ process where the forecast errors, *i.e.* the residuals between the real values and their forecasts, can be used for monitoring. For example, for an $AR(2)$ process, $X(t) = \phi_1 X(t-1) + \phi_2 X(t-2) + \varepsilon(t)$, we monitor

$$R(t) = X(t) - E(X(t) \mid X_{t-1}) = X(t) - \phi_1 X(t-1) - \phi_2 X(t-2),$$

with estimated values of ϕ_1 and ϕ_2. The one-step ahead forecast errors are *iid*, with the same distribution as $\varepsilon(t)$ (Wei (1990)).

3. Modelling and surveilling vendace

In this section we will study the data for vendace (*Coregonus albula*) from a univariate point of view. First we have to identify the in control state. We will study both the alternative where periodicity is concerned and when data is considered aperiodic. The periodicity will be estimated either using a periodic

mean model or a Fourier series. Further, we will also compare models where successive observations are independent and a model where we assume data to be an $ARMA(p,q)$ process. Ideally, the state, C, which is considered to be in control is given by knowledge of the biological process. Here we have to use data to determine the in control state. Model selection will be based on the data 1964—1988, since we find it possible that the system is out of control in 1990. The parameters will, however, sometimes be estimated using data from 1964 to an earlier year than 1988, to show the impact of re-estimation on the surveillance.

In Section 3.1 the modelling of the in control state for vendace is described, beginning with periodic models where successive observations are independent followed by $ARMA$ models. In Section 3.2 we apply the techniques for surveillance of the process and study the impact of different model assumptions.

3.1. Modelling

For the current purpose, to illustrate surveillance we find it sufficient to describe the alternative models by the residual mean squares, RMS. Suppose we estimate l parameters in the model using $\{X(1),\ldots,X(s-1)\}$, then denote the expected value of $X(i)$ by $\widehat{\mu}_{l,s-1}(X(i))$. We then define

$$RMS(X_s, s-1) = \frac{1}{s-l}\sum_{i=1}^{s}\left(X(i) - \widehat{\mu}_{l,s-1}(X(i))\right)^2.$$

We will not discuss the choice of measure for model selection further, since it is only used for describing purposes and since the emphasis in this paper is on the surveillance of the specific application. There are however, other - and sometimes better - measures to use.

In theoretical papers on surveillance, the in control state is usually assumed to be that the process have a constant mean, μ^0 and variance σ^2. In many applications, however, the in control state will be that the system has a periodic variation of length l. It is well known that many systems of species (Colinvaux (1986), ch. 13) have periodic variations (cyclic as defined below), where classical examples include the ten year cycle in trapping of lynx in northwestern Canada (Elton and Nicholson (1942)) and the four year cycle of lemmings.

We make a distinction between seasonality and cyclic variation, although they are both periodic variations and can be treated in the same way mathematically. By seasonality we define a periodicity where the length of the period is related to a "natural" period of time, *i.e.* annual or daily variation. By cyclic variation we define other periodicities, not explained by "natural" time periods, *e.g.* 20 days, 9 hours or 6 years.

Assume we have an additive process with independent and known mean and constant variance, *i.e.* $X(t) = \mu_l(t) + \varepsilon(t)$, where $\varepsilon(t)$ are *iid*, $N(0,\sigma)$. Then the transformed process, $X^c(t)$, defined by $X^c(t) = X(t) - \mu_l(t)$, becomes a white noise process that can be used for surveillance. When $\mu_l(t)$ is unknown we

will in the following use estimated values, $\widehat{\mu}_l(t)$, and when also l is unknown we estimate l first and then estimate $\mu_l(t)$ using $\widehat{l}$ and get $\widehat{\mu}_{\widehat{l}}(t)$.

Using a time domain approach (Box and Jenkins (1972)), the process can be identified and its parameters estimated. The in control state can then be defined such as the process is an $ARMA(p,q)$ process. Since the forecast errors of a properly identified process should be *iid*, we can use them for surveillance instead. However, the critical event can manifest itself in different ways in the original process and the forecast errors, a problem that will not be discussed further in this paper.

3.1.1. Frequency domain approach

Although the present data series only consist of 30 time steps, we see possible periodic patterns in more than one species. Using a frequency domain approach, we can estimate the periodically varying mean. To do so, we will have to know, or estimate, the cycle length, l.

We will first study a model where the mean function is any function with period l, *i.e.* $\mu_l(t) = \mu_l(t+l), \forall t > 0$. Our model is that $X(t) = \mu_l(t) + \varepsilon(t)$, that the mean function, $\mu_l(t)$, is periodic with period l and that $\varepsilon(t) \sim N(0,\sigma^2)$ and independent. For a given l we estimate μ_l by using the time-subsets, $T_{l,s}(t) = \{t, t+l, t+2l, \ldots \leq s\}$, for $t = 1, \ldots, l$. They are all disjoint and their union becomes $\{1, \ldots, s\}$. The maximum likelihood estimate for $\mu_{l,s}(t)$, given l, becomes

$$\widehat{\mu}_{l,s}(t) = \frac{1}{\#T_{l,s}(t)} \sum_{i \in T_{l,s}(t)} X(i), \; t = 1, \ldots, l-1.$$

The estimated variance, σ^2, using X_s for the estimation becomes

$$\widehat{\sigma}_s^2 = \frac{1}{s-l-1} \sum_{i=1}^{s} \left(X(i) - \widehat{\mu}_{l,s}(X(i))\right)^2 = \frac{1}{s-l-1} \cdot RSS, \text{ for } s > l+1, \quad (3.1)$$

where RSS denotes the residual sum of squares.

The RMS for $l = 2, \ldots, 15$ are shown in Figure 1. It seems like we can get a better fit by using a periodic mean than a constant mean. However, the disadvantage of the periodic mean model is that we have to estimate $l+1$ parameters. There is a risk, when the sample size is small, that we overfit the model. If two models are having the same residual sum of square (RSS) but different number of parameters, the estimated variance will be unnecessary high for the model with more parameters, thereby reducing power to detect changes.

Instead of estimating a mean level for each part of the sample, we can fit a more parsimonious Fourier series of order 1, (see Tolstov (1962) or Churchill and Brown (1987) for example) *i.e.*

$$\mu_l(t) = \mu + \beta_1 \cos\left(2\pi \cdot \frac{t}{l}\right) + \beta_2 \sin\left(2\pi \cdot \frac{t}{l}\right).$$

This model needs three parameters to be estimated for the mean (*cf.* l above) independent of l. Parameters are estimated using linear regression and the variance is estimated analogously with (3.1). RMS for these models, $RMS_{Fo}(l)$ are also shown in Figure 1.

Using reasonably low values of l, we get local minima both for $RMS_{Fo}(l)$ and $RMS_{pm}(l)$ for $l = 9$ years. A comparison between the true data and the different models, with $l = 9$, are shown in Figure 2. Since the periodic mean model have more parameters than the Fourier model, RMS_{pm} is bigger than RMS_{Fo} for most values of l, except $l = 9$ where we seem to get a better fit with the parameter mean model even in spite of that penalty. Note also the big difference in RMS_{pm} when we move l away from 9, the Fourier series model seems to be more robust for the choice of l.

3.1.2. Time domain approach

In the time domain approach of the problem, we identify the process and estimate the parameters using the Box-Jenkins approach (Box and Jenkins (1974) or Wei (1990)). As usual we define the $ARMA(p,q)$ model with mean $\mu(t)$ as

$$\begin{aligned} X(t) = \; & \mu(t) + \varphi_1 X(t-1) + \ldots + \varphi_p X(t-p) \\ & -\theta_1 \varepsilon(t-1) - \ldots - \theta_q \varepsilon(t-q) + \varepsilon(t) \end{aligned},$$

where $\varepsilon(t)$ are *iid* and $N(0,\sigma^2)$. The one-step ahead forecast errors, will be *iid* with mean 0 and variance $\sigma^2 = Var(\varepsilon(t))$, and can therefore be used for surveillance procedures by for example the Shewhart method.

The identification of the process is made by using the sample autocorrelation function (*acf*), $\widehat{\rho}_k$, and the sample partial autocorrelation function (*pacf*) both shown in Figure 3. For the vendace data, we find that a suitable model would be the $AR(1)$, having three parameters shown in table 1.

Table 1. Estimates of the $AR(1)$ model

Parameter	μ	φ_1	σ
Estimated value	158	0.39	31.2

The sample mean and the Yule-Walker estimate are used for estimating μ and φ_1, respectively. The variance, σ^2, is estimated by $\widehat{\sigma}^2 = \widehat{Var}(X_t) \cdot (1 - \widehat{\varphi}_1 \cdot \widehat{\rho}_1)$. The forecast errors are plotted in Figure 4. The RMS for this model, $RMS_{AR1} = 922$. Diagnostic checking shows that the fit seems to be accurate enough, although we have an indication of a possible periodicity of 9 years.

3.1.3. Comparison between the models

We will now discuss the arguments for and against the models designed above: the $AR(1)$ model, the *iid* model, the periodic mean model and the Fourier series model. The best fits for the periodic models are for a period of 9 years and these

periods will be used in the comparisons. The 9 year cycle is also indicated when we use the $AR(1)$ model.

The results are summarized in table 2, sorted in descending order of RMS. The number of parameters include the estimated variance.

Table 2. Comparison between the RMS for the univariate models.

Model	#Parameters	RMS
iid	2	1077
$AR(1)$	3	922
Fourier series ($l = 9$)	4	708
Periodic mean ($l = 9$)	10	607

As expected, the *iid* have the maximum RMS. Adding only one extra parameter and using the $AR(1)$ model would give a notable improvement. Going further, would on one hand improve the fit even more, but it is obvious from Figure 1 that both the Fourier series model and the periodic mean model are sensitive to the choice of l. For $l \leq 8$ and $l \geq 11$ we get a better fit with the $AR(1)$ model. Also, after centering the process using the periodic models, the residual series is autocorrelated, thereby calling for a further step, with yet one more parameter.

We conclude that there are many different models that might fit the data, we therefore need better knowledge about the actual ecological model generating the data to be able to make a choice. If we consider a model with a periodic mean, a period of 9 years seems to be suitable. We also find that the models with a small number of parameters, the AR and the Fourier series models, are giving a fit that is sufficiently close, compared to the periodic mean model.

3.2. Monitoring vendace

We will now apply the models developed above for monitoring a change in distribution of the vendace population. We will only apply Shewhart test on the data, to make comparisons easier, even though other methods might give better performance. The mean and standard deviation are re-estimated at each time step, and the residuals are used for surveillance, using the formulation 2.1. We define $\widehat{\mu}_{s-1}(X(s))$ as the expected value of $X(s)$ estimated from X_{s-1} (page 8). Analogously, we define $\widehat{\sigma}_{s-1}(X(s))$, the standard deviation of the residual at s. The Shewhart test prescribes that an alarm is triggered at time s if $|X(s) - \widehat{\mu}_{s-1}(X(s))| > 3 \cdot \widehat{\sigma}_{s-1}(X(s))$.

Using the procedure described above, assuming that the process is *iid*, $X(i) \sim N(\mu, \sigma^2)$, we would get an alarm in 1990. If we instead apply the refined models described above, we will at least not get an alarm earlier than 1990. Table 3 shows the standardized deviation from the expected value, given the estimated mean and variance. The Shewhart method triggers an alarm when $S > 3$.

Table 3. The standardized deviation from the mean in 1990.

Model	S (1990)
$AR(1)$	-3.42
iid	-3.68
Fourier series ($l = 9$)	-5.42
Periodic mean ($l = 9$)	-6.59

The difference between the models are that the *iid* and the periodic ones on one side are comparing the current value with an expected value. The expected value of 1990 is higher than the 1989 one, in both the periodic models, and the drop in 1990 therefore sums with the expected increase.

The AR-model on the other hand, compares the current value with the preceding one. Since 1989 was lower than average, the 1990 drop was not extremely big. Autoregressive time series tend to behave in a way that they are strolling up and down around their means. The suspected cyclic variation could therefore possibly be explained as a part of this random walking.

With the surveillance procedures we have used, Shewhart with $3\hat{\sigma}$ limits, we would not get an alarm earlier than 1990. Decreasing the critical limit could make it possible to detect the change earlier if we accept the higher false alarm probability. Note however, that the Shewhart method is not always the optimal method to use, and an other choice of method might have detected a change earlier.

We conclude that different models can give different results. Since the data material, due to the long time steps (one year), will be small still for many years, ecological background or other prior information is needed to restrict attention to only a small number of possible models.

4. Monitoring five species simultaneously

Earlier we restricted the analysis to univariately monitor one species at a time; but we have five sequences of data that can be used in parallel. Monitoring the biological diversity have received more attention since the UN conference in Rio de Janerio 1992. In the Rio convention biodiversity is defined as "*the variability among living organisms... this includes diversity... between species and of ecosystems.*".

One way of combining the information from the multiple sources, and design a common system for monitoring all at a time, is by creating an index, that can be used for surveillance. Since the covariance matrix plays an important role we will also study it specially. We will see that it, for this example data set, has some interesting properties. Statistical methods for surveillance of multiple processes have been suggested by many authors, among which the most popular methods probably is Hotelling's T^2.

4.1. Diversity indices

Several indices of biodiversity, with different statistical and demographic properties, have been suggested. Overwiews can be found in for example Colinvaux (1986), Magurran (1988), Noss (1990) or Pielou (1975). However, no universally acceptable index exists. Least complicated is probably the species richness index, defined as the number of species. A widely used statistic for biodiversity is the Shannon-Wiener index (Shannon and Weaver (1949)), H', originally designed for measuring information content. It is defined as

$$H' = -\sum p_i \log(p_i),$$

where p_i is the proportion of individuals of species i, measured by some suitable unit. Given a number, N, of species, H' is maximized for $p_i = 1/N$ giving $H'_{\max} = \log(N)$. The Q index and the Pielou J' statistic, for example, are smaller alterations of H', giving higher weight to the richness (N), but when N is constant.

When we measure diversity any definition of "amount" can be used: Number of individuals, mass, area or anything else that has a relevant meaning for the studied species. We use the landed mass of each species. Standardized values of H', based on estimated mean and variance of H', *i.e.*

$$H''(t) = \frac{H'(t) - \widehat{E}_{1988}(H')}{\widehat{\sigma}_{1988}(H')},$$

have been plotted in Figure 5. Assuming that H' is approximately normal, we see from that Figure that a Shewhart $3\widehat{\sigma}$ limit will give no alarm at all.

4.2. Estimation of the covariance matrix

If we assume that the data comes from a multinormal distribution, $X(t) \sim MN_5(\mu(t), \Sigma)$, where the sequence is *iid.* The dimension of $X(t)$ and $\mu(t)$ is 1×5 and the dimension of Σ is 5×5. The mean is estimated for each process separately, by using X_s, is denoted $\widehat{\mu}_s(i)$. Let p be the number of parameters used in $\widehat{\mu}_s(i)$, for constant means within each process $p = 5$. We estimate Σ successively over time, $\widehat{\Sigma}_s$ estimated using X_s, by

$$\widehat{\Sigma}_s = \frac{1}{s-p}\sum_{i=1}^{s}(X(i) - \widehat{\mu}_s(i))^T(X(i) - \widehat{\mu}_s(i)) \text{ when } s > p.$$

When we do so and plot $\widehat{\Sigma}_s$ versus s we see an interesting pattern, Figure 6. The correlations between the species can be grouped into three groups:

1. Correlation between pike-perch and other.
2. Correlation between vendace and all but pike-perch.

3. Correlation within burbot, pike and perch.

Using this grouping we can reduce the number of parameters from 15 to 8. Below, we have replaced the individual covariances with the arithmetic means of the groups. The estimated covariance matrix becomes

$$\hat{\hat{\Sigma}}_{1988} = \begin{pmatrix} & \text{Vendace} & \text{Pike} & \text{Pike-perch} & \text{Perch} & \text{Burbot} \\ \text{Vendace} & 835.5 & 85.21 & 39.78 & 85.21 & 85.21 \\ \text{Pike} & . & 70.33 & 39.78 & 22.03 & 22.03 \\ \text{Pike-perch} & . & . & 743.0 & 39.78 & 39.78 \\ \text{Perch} & . & . & . & 14.97 & 22.03 \\ \text{Burbot} & . & . & . & . & 26.06 \end{pmatrix}. \quad (4.1)$$

A plot of $\hat{\hat{\Sigma}}_t$ against t is found in Figure 7. We have used all the data from 1964 up to t at each time step. One disadvantage that will not be discussed in detail here, is that since we accumulate more and more data for the estimation, the power for detecting a change recently in time, *e.g.* detecting the event $\{\tau = t\}$, will decrease. An alternative approach to avoid this risk would be to use a window of length w for estimation, *i.e.* $\hat{\Sigma}_t$ is estimated using $X(t-w+1), \ldots, X(t)$, this approach will however not be studied in this paper.

4.3. Hotelling's T^2

Hotelling's T^2-statistic (Hotelling (1947)) is defined as

$$T^2(t) = (X(t) - \mu)^T \Sigma^{-1} (X(t) - \mu),$$

where μ and Σ are the mean vector and covariance matrix, respectively. A more general definition would be to allow the mean vector $\mu(t)$ and covariance matrix $\Sigma(t)$ to vary over time. We will, however, in the following restrict attention to the case where the in control variance matrix and mean vector are constant, $\Sigma(t) \equiv \Sigma^0$ and $\mu(t) \equiv \mu^0$. The T^2-statistic can detect deviations from both mean and variance, but is most sensitive for changes in mean, especially when all the means are changing at the same time and direction. Since Hotelling's T^2 is a well known and often used statistic we will give it a go on our data.

Using the estimated covariance matrix for 1988 (4.1), $\hat{\hat{\Sigma}}_{1988}$, and a constant mean vector, $\hat{\mu}_{1988} = (159.7, 36.76, 145.5, 11.98, 12.50)$, the T^2-statistic based on the estimated values is plotted in Figure 8. The estimated mean vector have a multinormal distribution, $\hat{\mu}_t \sim MN_q(\mu, \Sigma/t)$, and the estimated variance matrix have a Wishart distribution, $\hat{\Sigma}_t \sim W_q(t-1, \Sigma)$, which is a multivariate extension of the χ^2-distribution (Crowder and Hand (1993)). Approximating the distribution of $\hat{\Sigma}_t$ by a $W_q(t-1, \Sigma)$ we get

$$\frac{q}{(t+1)(t-q+1)} T^2 \overset{approx}{\sim} F_{q,t-q+1}.$$

Using μ and Σ estimated with data up to 1988 ($t = 25$), we would yield an alarm already in 1989 with the critical limit -4.55.

5. Discussion and conclusions

We see from the univariate analysis of the vendace (*Coregonus albula*) data above, that the choice of model is of great importance. When different cyclic patterns or autocorrelations are present, data have to be modified to take this into account. We find that goodness of fit can be improved by estimating cyclic patterns or autocorrelation. By using either of the four models studied in this paper, the Shewhart procedure would have detected a change by 1990. The weakest reactions come from the $AR(1)$ and *iid*-models. The $AR(1)$ includes a natural wandering around the mean of the system that partly explains the drop in 1990. Due to this, the $AR(1)$-model stops calling the alarm in 1991. The periodic models both expected an increasing catch by 1990. Because of this the difference between the expected and actual catch in 1990 was magnified and an alarm was triggered.

As expected, the species of fish are all positively correlated with each other. This leads to the conclusion that the drop in vendace will either be explained by the other species behaving in the same way or else lead to a decreasing correlation between vendace and other species. Both scenarios are interesting as the ecological causes possibly have to be sought in different places. As with the univariate problem above, the description of the critical event is crucial and is also affected by the in control system.

Neither the Shannon-Wiener index nor the variants of it studied in this paper manages to call any alarm at all. Hotelling's T^2, however, detects a change already in 1989, one year earlier than the univariate procedures. This is likely to depend on the fact the change in the vendace population is a small one in the proportions between the species while the change for vendace it self is quite big. Since Pielou J' is sensitive to changes in proportions between the species it fails while Hotelling's T^2 can detect big changes in one species at a time.

Using more than one species can make it possible to detect changes, when the critical event is being magnified by the interaction between the species. Also, using more than one species will make it possible to monitor biodiversity.

There is a great potential and usefulness for statistical surveillance on environmental data. Over the world enormous amounts of data are collected about the environment and stored for analysis. Quality control and statistical process control is used, as was mentioned earlier, in many places in society, and now the turn must have come for environmental data.

Acknowledgment

This paper is a part of a research project on statistical surveillance at the Göteborg University, which is supported by the Swedish National Council for Research in the Humanities and Social Sciences. I would like to thank the supervisor of the project, professor Marianne Frisén, for her helpful advice, ideas and encouragement. I am also specially grateful to Kasra Afsarinejad and all my friends and colleagues at the department.

Data material for the fish example was provided by Olof Filipsson at the Fresh Water Laboratory, Drottningholm.

This paper have been presented at a poster session at the conference on Environmental Statistics and Earth Science in Brno, Czech republic, 1996. The participation was financed by the Victor Rydberg, memory fund, at the faculty of Arts and Science, Göteborg University.

References

[1] Arnkelsdóttir, H. (1995). *Surveillance of rare events. On evaluations of the sets method.* Research report 1995:1, Departement of Statistics, Göteborg University.

[2] Berthoux, P. M., W. G. Hunter and L. Pallesen (1978). "Monitoring Sewage Treatment Plants: Some Quality Control Aspects". *Journal of Quality Technology* 10, 139–149.

[3] Box, G. E. P. and G. M. Jenkins (1976). *Time series analysis forecasting and control.* Holden-Day.

[4] Brown R. L., J. Durbin and J. M. Evans (1975). "Techniques for Testing the Constancy of Regression Relationships over Time". *Journal of the Royal Statistical Society - B* 37, 49–163.

[5] Charnes, J. M. and H. S. Gitlow (1995). "Using Control Charts to Corroborate Bribery in Jai Alai". *The American Statistician* 49(4), 386–389.

[6] Churchill, R. V. and J. W. Brown (1987). *Fourier Series and Boundary Value Problems.* McGraw-Hill.

[7] Colinvaux, P. (1986). *Ecology.* John Wiley and Sons.

[8] Crowder, M. J. and D. J. Hand (1993). *Analysis of Repeated Measures.* Chapman and Hall.

[9] Elton and Nicholson (1942). "The ten-year cycle in numbers of lynx in Canada". *Journal of Animal Ecology* 11, 215–244.

[10] Frisén, M. (1992). "Evaluations of Methods for Statistical Surveillance". *Statistics in Medicine* 11, 1489–1502.

[11] Frisén, M. (1994a). *Characterization of methods for surveillance by optimality.* Research report 1994:2, Department of Statistics, Göteborg University.

[12] Frisén, M. (1994b). *Statistical surveillance of business cycles.* Research report 1994:1, Departement of Statistics, Göteborg University.

[13] Frisén, M. and J. de Maré (1991). "Optimal surveillance". *Biometrika* 78, 271–280.

[14] Hotelling, H. (1947). "Multivariate Quality Control. Illustrated by the Air Testing of Bombsights". In *Techniques of Statistical Analysis*, eds. Eisenharts and Hastay, McGraw-Hill, 111–184.

[15] Kjelle, P.-E. (1987). *Alarm critera for the fixed gamma radiation monitoring stations.* Research report 1987:7, National Institute of Radiation Protection.

[16] Magurran, A. E. (1988). *Ecological diversity and its Measurements.* Princeton university press.

[17] Noss, R. F. (1990). "Indicators for monitoring biodiversity: A hierarchial approach". *Conservation Biology* 4, 355–364.

[18] Page, E. S. (1954). "Continous inspection schemes". *Biometrika* 41, 100–114.

[19] Pielou, E. C. (1975). *Ecological diversity.* John Wiley and Sons.

[20] Roberts, S. W. (1959). "Control chart tests based on geometric moving averages". *Technometrics* 1, 239–250.

[21] Settergren Sørensen, P. and J. la Cour Jansen (1991). "Statistical Control of Hygienic Quality of Bathing Water". *Environmental monitoring and assessment* 17, 217–226.

[22] Shannon, C. E. and W. Weaver (1959). *The mathematical theory of communication.* The University of Illinois press: Urbana.

[23] Shewhart, W. A. (1931). *Economic Control of Quality Control.* Reinhold Company, Princeton, NJ.

[24] Shiryaev, A. N. (1963). "On optimum methods for quickest detection problems". *Theory probab. appl.* 8, 22–46.

[25] Sveréus, A (1995). *Detection of Gradual Changes. Statistical Methods in Post Marketing Surveillance.* Research report 1995:2, Department of Statistics, Göteborg University.

[26] Tolstov, G. P. (1962). *Fourier Series.* Translated by R. A. Silverman. Dover.

[27] Vaughan, W. J. and C. S. Russell (1983). "Montoring Point Sources of Pollution: Answers and More Questions from Statistical Quality Control". *The American Statistician* 37, 476–487.

[28] Wei, W. W.-S. (1990). *Time Series Analysis.* Addison-Wesley.

[29] Wessman, P. (1996). *Multivariate Surveillance.* Research report 1996:4, Department of Statistics, Göteborg University.

[30] Wetherill, G. B. and D. W. Brown (1991). *Statistical Process Control - Theory and practice.* Chapman and Hall.

[31] Zacks, S. (1983). "Survey of classical and Bayesian approaches to the change-point problem: Fixed sample and sequential procedures of testing and estimation". *Recent advances in statistics*, 245–269.

[32] Zacks, S. (1989). "Detection and Change-Point Problems". In *Handbook of Sequential Analysis.* Marcel Dekker.

Legend to Figures

1. Residual mean square (RMS) for different models of the vendace data: Constant mean model (dashed line), the periodic mean model (rings), the Fourier series model (cross) and the $AR(1)$ model (dotted line). The constant mean and the $AR(1)$ model are independent of period length.

2. The real data series (stars) compared with its expected value of the fitted periodic mean (dashed line) and Fourier series model (dotted line), all estimations are made on data 1964—1988. The shape of the three sequences coincides at least until 1988.

3. Sample autocorrelation function (*acf*) and sample partial autocorrelation function (*pacf*) for the catches of vendace, estimated using data from 1964—1988. Estimation is made on data from which the sample mean have been subtracted. A reasonable identification seems to be the $AR(1)$-model.

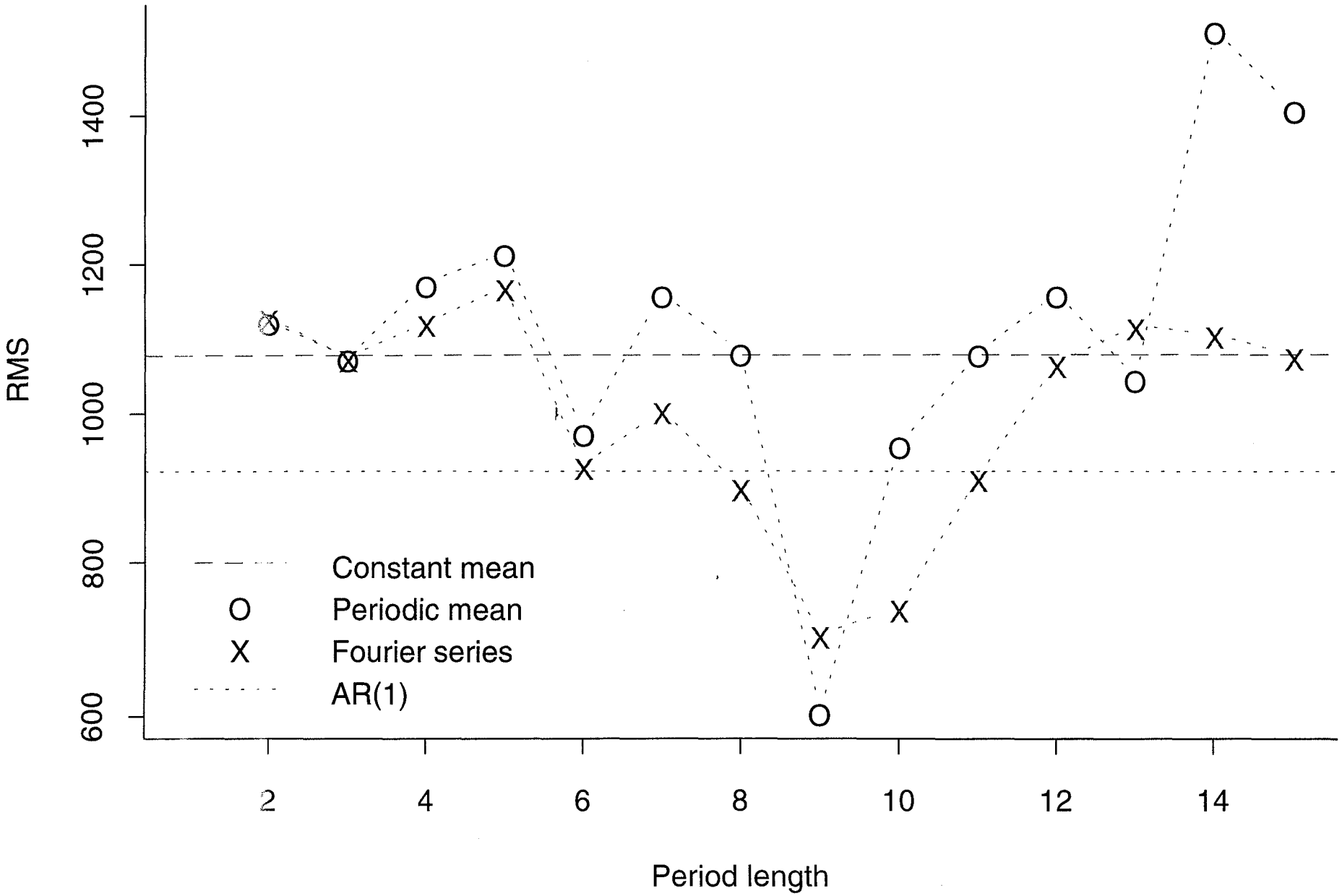
4. One step ahead forecast errors for the catches of vendace. Forecasts are based on the $AR(1)$-model. The alarm limit is the $3 \cdot \sigma_{t-1}(t)$ Shewhart limit with $\sigma_{t-1}(t)$ estimated from the data X_{t-1}.

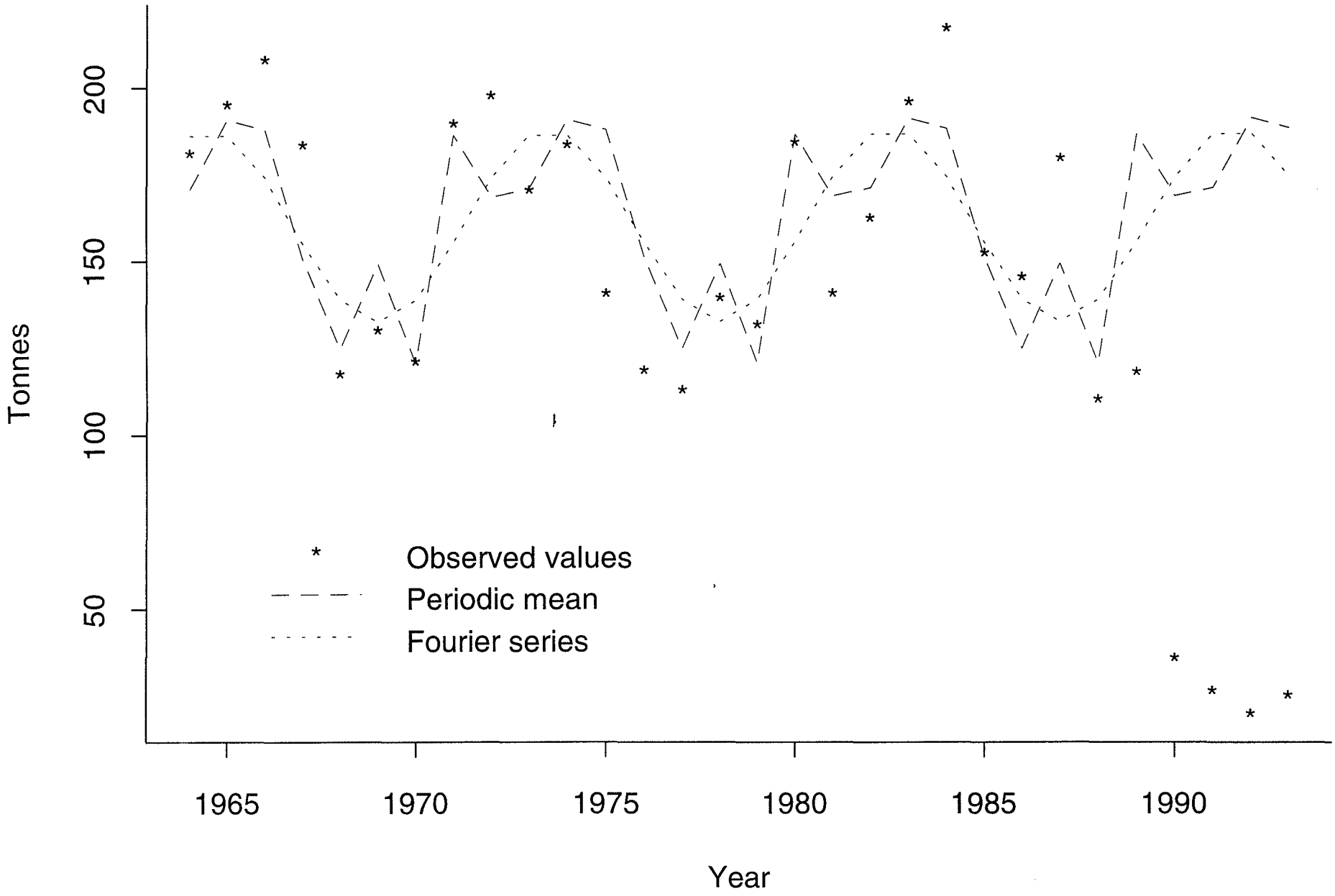
5. Standardized Shannon-Wiener index. The standardisation is made on the sample mean and standard deviation, *i.e.* H'_{t-1}. A value exceeding 3 would trigger an alarm with the Shewhart method.

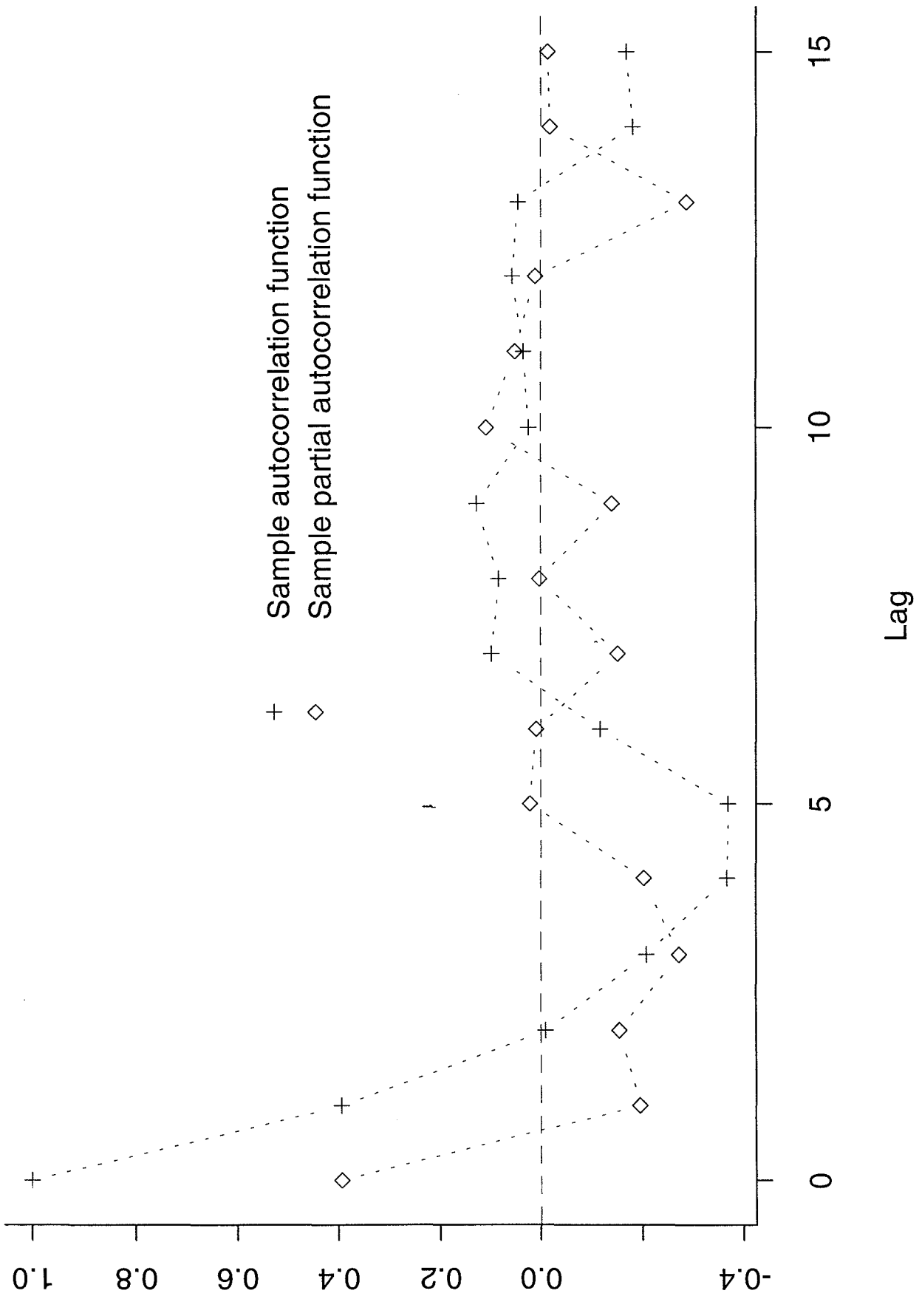
6. Pairwise correlation coefficients between species, estimated between 1964 and the current year. Correlations have been grouped into three groups where the correlations are almost equal to each other: The "pike-perch"-group (correlations between pike-perch, *Lucioperca lucioperca*, and any other specie), "vendace"-group (correlations between vendace, *Coregonus albula*, and any other specie except pike-perch) and "other"-group (correlations between burbot, *Lota lota*, perch, *Perca fluviatilis* and pike, *Esox lucius*).

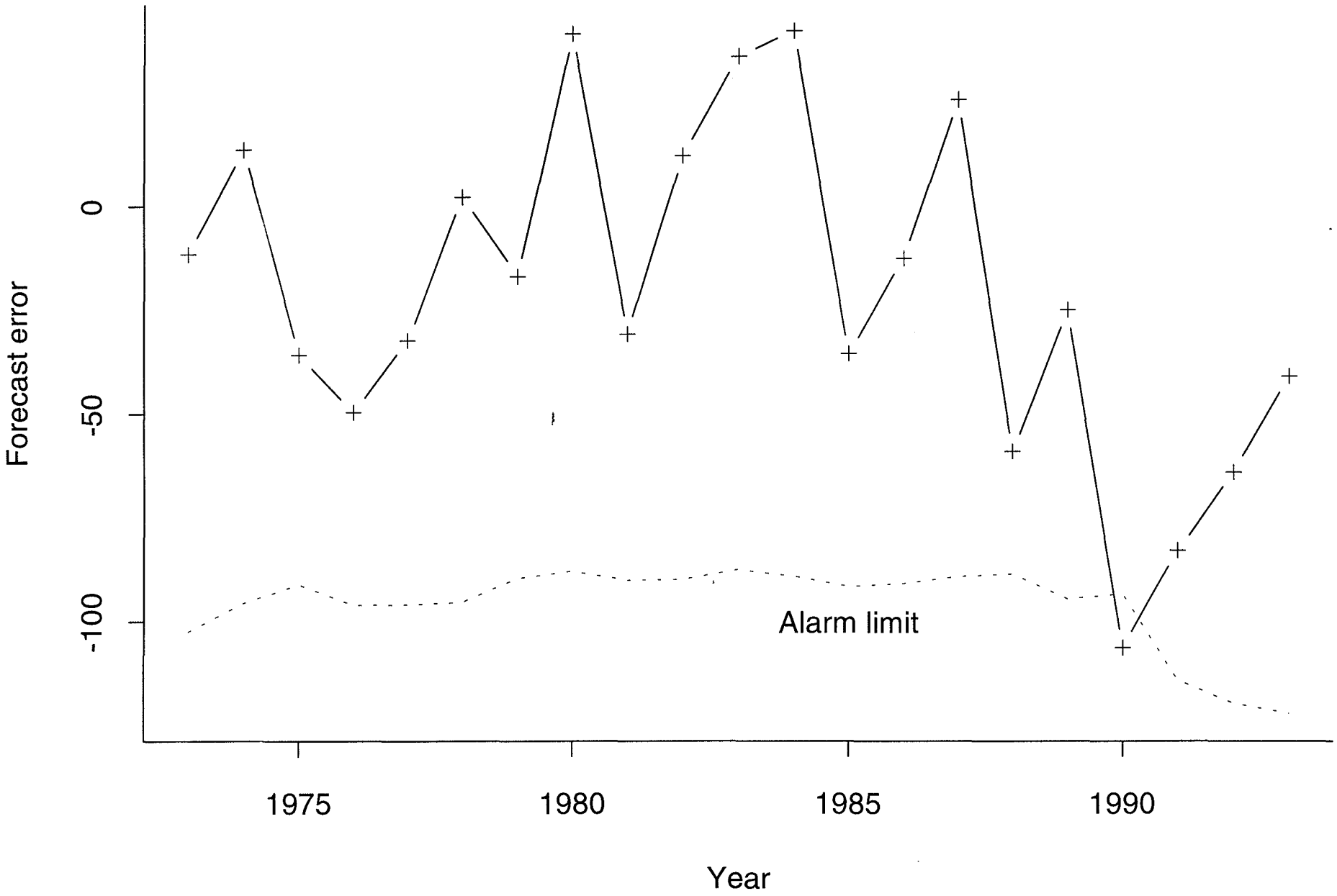
7. Aritmetic mean of the pairwise correlation coefficients taken over the correlation groups. Vendace contains the correlation between vendace and all other species except pike-perch. Other contain the correlations between burbot, perch and pike.

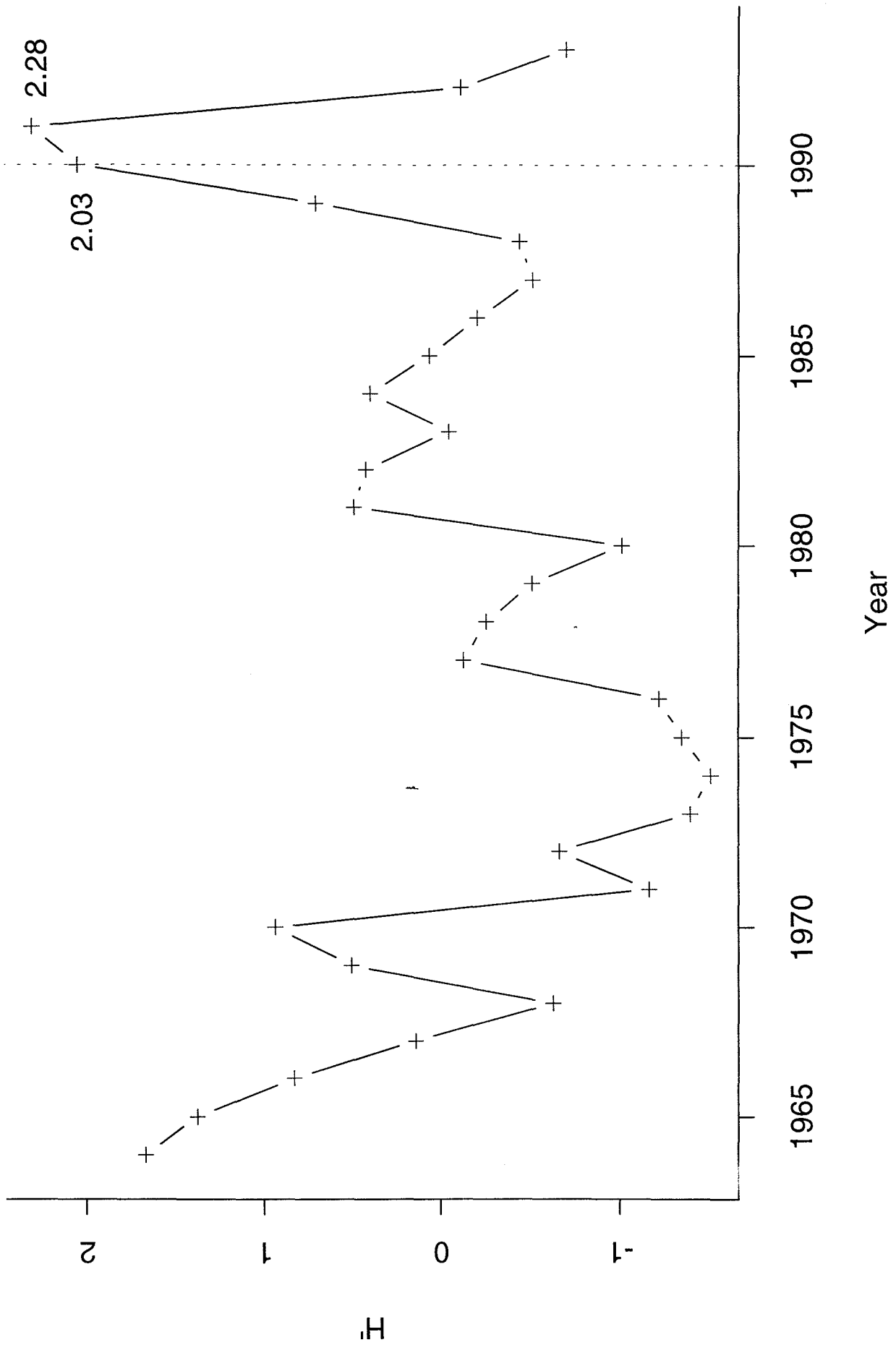
8. Hotelling's T^2-statistic based on mean and variance estimated up to 1988. The variance matrix have been replaced by the reduced variance matrix, $\widetilde{\Sigma}$, where the covariances have been replaced by their arithmetic means within the correlation groups. The critical limit, -4.55, have been estimated using an approximated using the F-distribution.

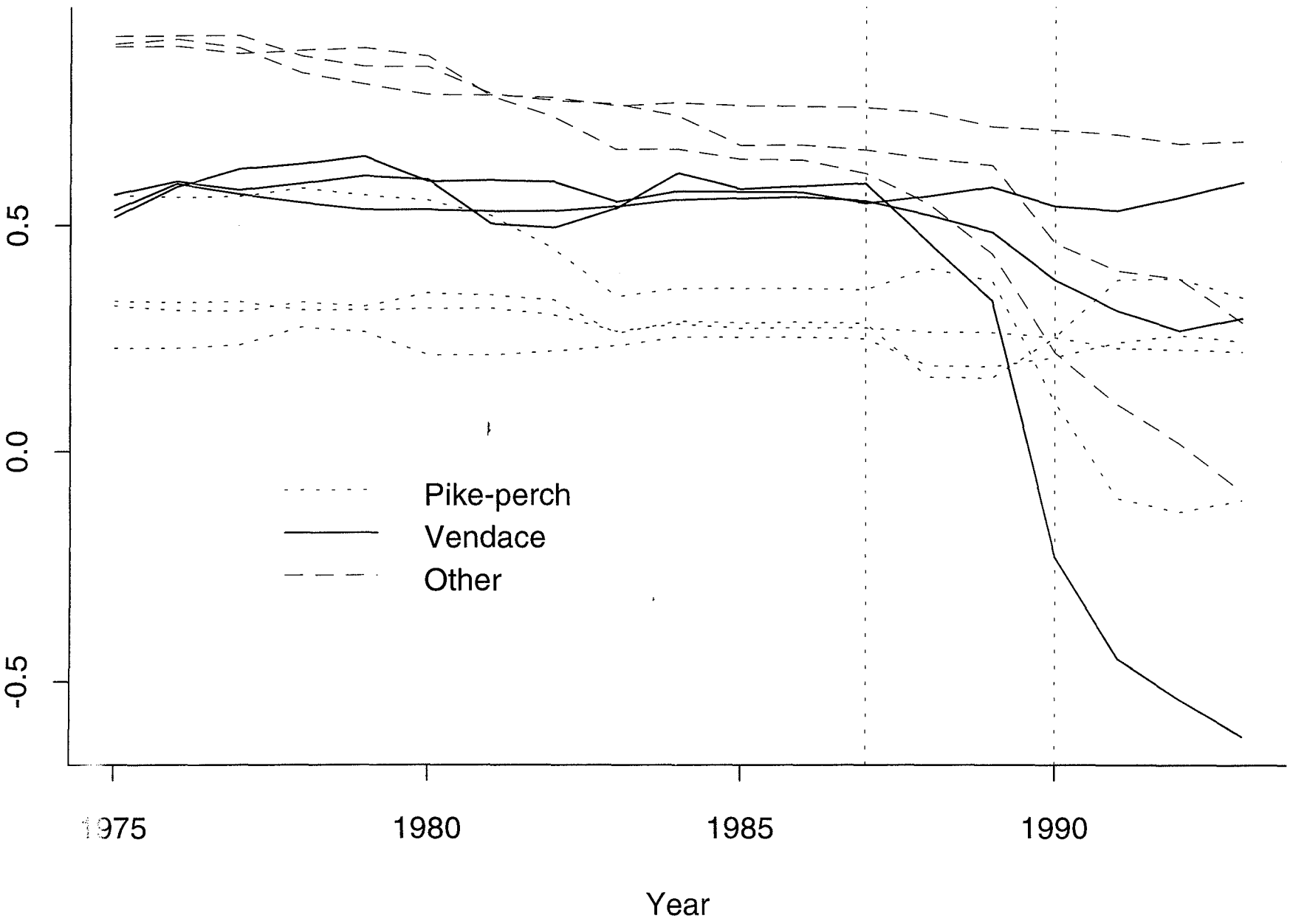

RMS
1400
1200
1000
800
600
2
4
6
8
10
12
14
Period length
Constant mean
Periodic mean
Fourier series
AR(1)

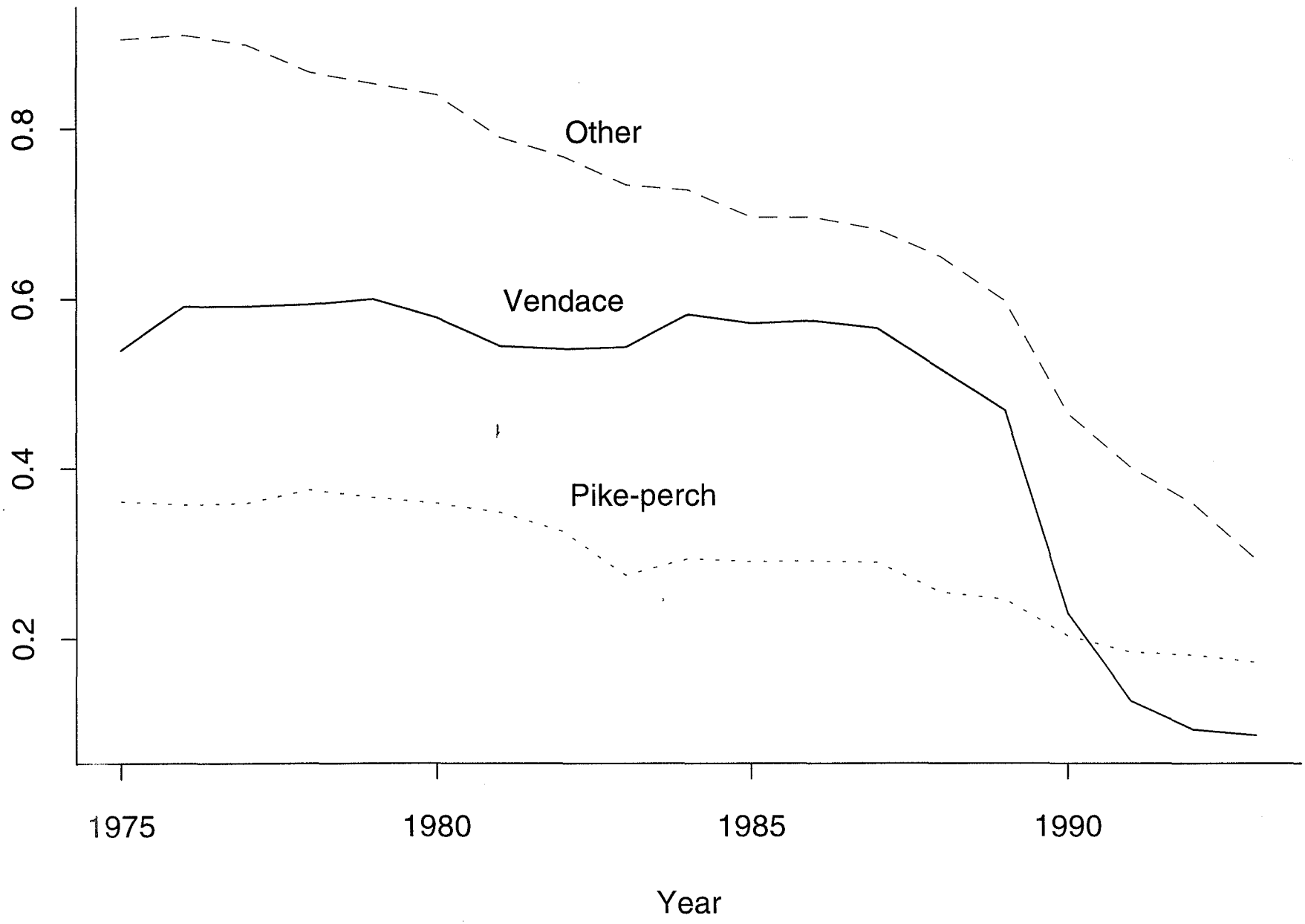

Tonnes
50
100
150
200
1965
1970
1975
1980
1985
1990
Year
Observed values
Periodic mean
Fourier series

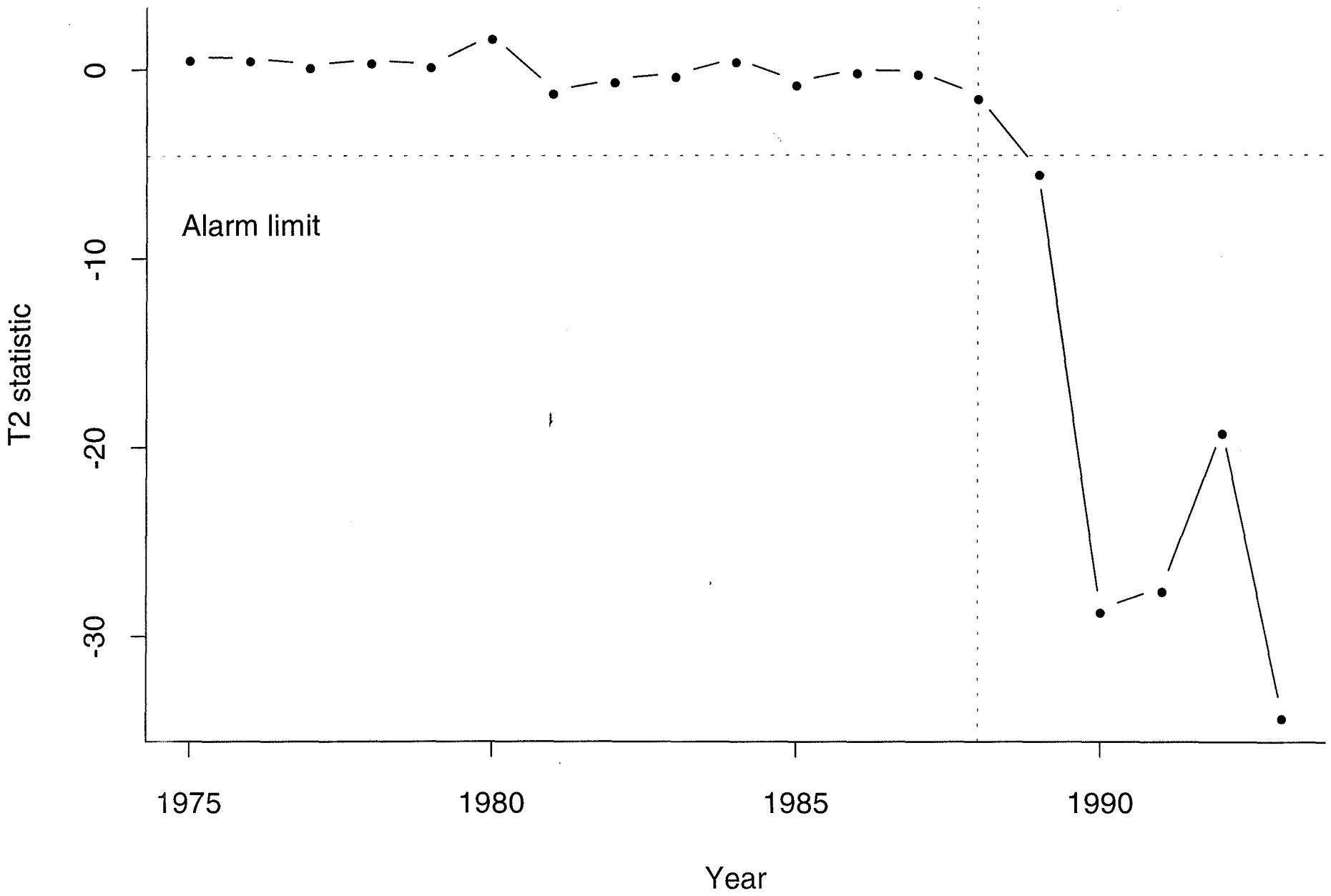

Sample autocorrelation function
Sample partial autocorrelation function
1.0
0.8
0.6
0.4
0.2
0.0
-0.4
0
5
10
15
Lag


Forecast error
0
-50
-100
Alarm limit
1975
1980
1985
1990
Year


2.28
2.03
H'
2
1
0
-1
1965
1970
1975
1980
1985
1990
Year


0.5
0.0
-0.5
Pike-perch
Vendace
Other
1975
1980
1985
1990
Year


0.8
0.6
0.4
0.2
Other
Vendace
Pike-perch
1975
1980
1985
1990
Year


T2 statistic
0
-10
-20
-30
Alarm limit
1975
1980
1985
1990
Year

Research Report

1995:1	Arnkelsdottir, H	Surveillance of rare events. On evaluations of the sets method.
1995:2	Sveréus, A	Detection of gradual changes. Statistical methods in post marketing surveillance.
1995:3	Ekman, C	On second order surfaces estimation and rotatability.
1996:1	Ekman, A	Sequential analysis of simple hypotheses when using play-the-winner allocation.
1996:2	Wessman, P	Some principles for surveillance adopted for multivariate processes with a common change point.
1996:3	Frisén, M. & Wessman, P	Evaluations of likelihood ratio methods for surveillance.
1996:4	Wessman, P. & Frisén, M	Evaluation of univariate surveillance procedures for some multivariate problems.
1996:5	Särkkä, A	Outlying observations and their influence on maximum pseudo-likelihood estimates of Gibbs point processes.
1997:1	Ekman, A	Sequential probability ratio tests when using randomized play-the-winner allocation.