PREDICTION OF HIGH WATER LEVELS IN THE BALTIC

A. Svensson and J. Holst
Department of Mathematical Statistics
Lund Institute of Technology, Box 118, SE-22100 Lund, Sweden
Phone: +46 46 2224679 (ASv); +46 46 2229538 (JH)
email: anderss@maths.lth.se (ASv)
         janh@maths.lth.se (JH)

Abstract

This paper presents a level crossing predictor for Gaussian ARMAX processes, which is optimal in the sense that it minimizes the number of false alarms for a given probability of detecting the level-crossings. It is applied to real data for predicting and warning for high water levels at the Danish coast in the Baltic Sea. The optimal alarm system is shown to work better than a simpler and more conventional alarm system. A method to optimally predict the crossings also when the external signals are not known is presented. In this particular case most of the variability of the predictions are due to system noise, so the performance of the system with predicted external signals are almost identical to the performance when the external signals are known. A smaller simulation study shows that the water level process is hard to predict and that the choice of model can be rather important.

KEY WORDS: level crossings, flooding alarm, catastrophe prediction, optimal alarm, ARMAX process.

1. Introduction

A flooding incident can be disastrous, especially if people are not warned. Hence, in many situations it is important to be able to give an alarm some time before the incident occurs. It is also important to give as few false alarms as possible, but still find a sufficient number of the flooding incidents.

In a more general setting, the problem is to predict level crossings, catastrophes, of a stochastic process a sufficient time in advance. This catastrophe prediction problem was treated by de Maré and Lindgren, and a definition of the optimal catastrophe predictor was given as the predictor that gives a minimum number of false alarms for a given
detection probability. This idea was further treated in Svensson, Holst, Lindquist & Lindgren, and leads to an explicit catastrophe predictor for Gaussian ARMA processes with constant catastrophe level. Since the construction of the optimal catastrophe predictor requires quite a large amount of calculations, two suboptimal predictors were also introduced. In Svensson & Holst

the technique was extended to cover both ARMAX and SETARMAX processes with a deterministic but changing catastrophe level. This made it possible to use the optimal catastrophe predictor on real data, describing water levels in the Baltic Sea, presented in this paper. Modelling of the water levels in the Baltic Sea is treated in Berntsen, Nielsen and Spliid & Nielsen. A complication with ARMAX processes is that the external signals might not be known in advance, which means that they have to be predicted too. An idea how this can be treated in the same framework as above is also included in this paper and applied to the data sets used.

2. The data set

The data sets used in this paper are from 1978, 1979 and 1980. They consist of the following measurements.

<table>
<thead>
<tr>
<th>Location</th>
<th>Water level</th>
<th>Head wind</th>
<th>Side wind</th>
<th>Air pressure</th>
<th>Temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korsør</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rødbyhavn</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gedser</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visby</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kadetrenden/Maribo (78)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Møn-Sydøst lightship (79,80)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Møn lighthouse:</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Christiansø lighthouse</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Hammer Odde lighthouse</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Only three of these signals are used in the final model describing the water level at Rødbyhavn. They are the water level at Rødbyhavn, the head wind at Christiansø lighthouse and the air pressure at Kadetrenden/Maribo (78) or Møn-Sydøst lightship (79,80). The original data sets contained measurements every hour, but since the process is oversampled, only one sample per 3 hours was used for modelling the water level. Before they have been used for modelling, the mean value using data from all three years has been subtracted. However, the catastrophe levels used later are related to the original data. In Figure /refwater78 the water level at Rødbyhavn is shown for the data set from 1978. The complete data sets with a short description can be found at the address: http://www.maths.lth.se/matstat/staff/anderss/data/data.html.
3. The models

The water level at Rødbyhavn has been modelled as ARMAX and SETARX processes, denoted \( X_t \), with two external signals, denoted \( u_{1,t} \) and \( u_{2,t} \). The structure of an ARMAX\( (p,q,r_1,r_2) \) process is

\[
X_t + a_1 X_{t-1} + \ldots + a_p X_{t-p} = b_{1,0} u_{1,t} + \ldots + b_{1,r_1} u_{1,t-r_1} + b_{2,0} u_{2,t} + \ldots + b_{2,r_2} u_{2,t-r_2} + c_0 e_t + \ldots + c_q e_{t-q},
\]

or shorter

\[
A(z^{-1})X_t = B_1(z^{-1})u_{1,t} + B_2(z^{-1})u_{2,t} + C(z^{-1})e_t,
\]

where \( \{e_t\}_{t=-\infty}^{\infty} \) is white noise and \( e_t \) is uncorrelated with \( X_s \), \( u_{1,s} \) and \( u_{2,s} \) for \( s < t \). It is furthermore assumed that \( e_t \in \mathcal{N}(0,1) \).

After trying a number of different models three were chosen and estimated on the data from 1978, and optimal alarm systems were calculated. The models are ARMAX\( (2,1,1,1) \), ARMAX\( (4,2,1,1) \) and SETARX\( (2;2,2;1,1) \). The noise is assumed to be independent and Gaussian with variance 1.

The ARMAX\( (2,1,1,1) \)-model is

\[
\begin{align*}
A(z^{-1}) &= 1.0000 - 1.2794z^{-1} + 0.3786z^{-2} \\
C(z^{-1}) &= 5.5678 + 3.1836z^{-1} \\
B_1(z^{-1}) &= -0.0072z^{-1} \\
B_2(z^{-1}) &= 0.0232z^{-1}
\end{align*}
\]

The empirical density functions for the one and two-step prediction error for the data set from 1978 are shown in Figure , together with the normal density function, and normal probability plots. It can be seen that the residuals have slightly heavier tails than in the normal distribution. However, in spite of these deviations the normal distribution has been used for modelling and calculation of the alarm systems. It seems to work rather well.

The ARMAX\( (4,2,1,1) \)-model is

\[
\begin{align*}
A(z^{-1}) &= 1.0000 - 1.7227z^{-1} + 1.7602z^{-2} - 1.5950z^{-3} + 0.6652z^{-4} \\
C(z^{-1}) &= 5.1942 + 0.4259z^{-1} + 3.7263z^{-2} \\
B_1(z^{-1}) &= -0.0074z^{-1} \\
B_2(z^{-1}) &= 0.0225z^{-1}
\end{align*}
\]

The SETARX-model is composed of two ARX-models where

\[
\begin{align*}
A(z^{-1}) &= 1.0000 - 1.5280z^{-1} + 0.6137z^{-2} \\
C(z^{-1}) &= 6.0661 \\
B_1(z^{-1}) &= -0.0049z^{-1} \\
B_2(z^{-1}) &= 0.0155z^{-1}
\end{align*}
\]
A. SVENSON, J. HOLST

is used when the process value $X_{t-2} < 30$ and

\[
\begin{align*}
A(z^{-1}) & = 1.0000 - 1.4226z^{-1} + 0.7434z^{-2} \\
C(z^{-1}) & = 6.0661 \\
B_1(z^{-1}) & = 0.0202z^{-1} \\
B_2(z^{-1}) & = 0.1312z^{-1}
\end{align*}
\]

when the process value $X_{t-2} \geq 30$.

In cases when models for the external signals are needed these signals have been modelled as AR processes. The model for the head wind at Christiansø lighthouse is an AR(3) process with the parameters

\[
\begin{align*}
A_1(z^{-1}) & = 1.0000 - 1.7042z^{-1} + 0.5411z^{-2} + 0.1713z^{-3} \\
C_1(z^{-1}) & = 7.7460
\end{align*}
\]

and the model for the air pressure at Kadetrenden/Maribo (78) or Møn-Sydøst lightship (79,80) is an AR(1) process with the parameters

\[
\begin{align*}
A_2(z^{-1}) & = 1.0000 - 0.9657z^{-1} \\
C_2(z^{-1}) & = 27.8675.
\end{align*}
\]

It could be considered using one model for the air pressure at Kadetrenden/Maribo (78) and another model for Møn-Sydøst lightship (79,80), but since the locations are rather close to each other, the same model has been used. This also requires fewer calculations.

4. The optimal alarm system

The optimal alarm systems used in this paper are optimal in the sense that they minimize the probability of false alarms for a given probability of detecting the catastrophes. Optimality is reached by the alarm system defined through the likelihood ratio,

\[
\frac{dP_{Y(t-k)}(y|C_t^*)}{dP_{Y(t-k)}(y|C_t)} \leq \text{constant},
\]

where $Y(t)$ denotes the available information at time $t$, $C_t$ is the event that a catastrophe occurs at time $t$ and $C_t^*$ is the complementary event that no catastrophe occurs at time $t$. This condition can be simplified, so that the alarm system can be based on only the predictor $(\hat{x}_{t-1}, \hat{x}_t)$ of the process $X_t$ at times $t-1$ and $t$, instead of all the available information $Y(t)$. The result is

\[
P(C_t|\hat{x}_{t-1}, \hat{x}_t) > P_b,
\]

which was shown in Svensson et al., to be the optimal alarm system for ARMA processes with a constant catastrophe level. It is then possible to calculate the alarm region in advance, which makes the alarm system rather fast. A typical alarm region
in the \((\hat{x}_{t-1}, \hat{x}_t)\)-plane is shown in figure. The model is the ARMAX(2,1,1,1) described above, with the influence of the external signals subtracted. The predictor is using 6 old process values and the prediction horizon is 2. This idea was further developed in Svensson & Holst, to cover ARMAX and SETARX processes when the external signals are known and the process is stationary.

5. Alarm system using predicted external signals

Since predictions of the process values are needed in the level crossing predictor, also predictions of the external signals are needed when the external signals are not known in advance. In case of known external signals, the effect can be included in the catastrophe level, giving a catastrophe level that changes through time, see Svensson & Holst. However it is not that simple in case of stochastic external signals. A few assumptions on the signals have to be added in order to get an explicit level crossing predictor.

If we assume that the external signals and the process are stationary Gaussian processes, the covariance of the process value predictor \(\text{Cov}(\hat{x}_{t-1}, \hat{x}_t)\) will include both the effects of the process noise and the external signals.

Suppose the process can be written,

\[
A(z^{-1})X_t = B(z^{-1})u_t + C(z^{-1})\epsilon_t \\
A_1(z^{-1})u_t = C_1(z^{-1})w_t.
\]

Due to linearity, \(X_t\) can be decomposed into one part, \(X_{u,t}\), describing the influence of the external signals and one part, \(X_{e,t}\), describing the influence of the system noise.

\[
X_t = X_{u,t} + X_{e,t} \\
A(z^{-1})X_{e,t} = C(z^{-1})\epsilon_t \\
A(z^{-1})A_1(z^{-1})X_{u,t} = B(z^{-1})C_1(z^{-1})w_t.
\]

The same deductions can be done for the predictions, leading to

\[
\hat{X}_t = \hat{X}_{u,t} + \hat{X}_{e,t}.
\]

If the noise processes \(\epsilon_t\) and \(w_t\) are assumed to be independent, the covariance of the predictions \(\hat{x}_{t-1}, \hat{x}_t\) is

\[
\text{Cov}(\hat{x}_{t-1}, \hat{x}_t) = \text{Cov}(\hat{x}_{u,t-1}, \hat{x}_{u,t}) + \text{Cov}(\hat{x}_{e,t-1}, \hat{x}_{e,t}).
\]

This means that if \(X_{u,t}\) can be optimally predicted, the technique presented in Svensson & Holst, can still be used and thus the resulting critical levels for the stochastic part of the process will be

\[
[L_{\text{cat}}(t-1|t-k), L_{\text{cat}}(t|t-k)] = [L(t-1), L(t)] - [\hat{x}_u(t-1|t-k), \hat{x}_u(t|t-k)].
\]
The catastrophe level $L(t)$ for the original process, is assumed to be deterministic and
known, and need not be predicted. The part of the process that is due to the external
signals, influences the mean value of the process and will thus enter as an addition to
the catastrophe level. Predictions for times $t - 1$ and $t$ are needed and the information
is available up until $t - k$.

One model, the ARMAX$(2,1,1,1)$-model with the external signals modelled as AR$(3)$
and AR$(1)$ as above, have been tested and the results are shown and compared to the
other alarm systems in Table 4, Table 5 and Table 6. The alarm system works well for
the data set that was used for estimating the model, but poorer for the other two data
sets. The reason for this could be that the fixed models for the process and the external
signals are not totally correct. This is similar to the alarm system where the external
signals are not predicted, which is expected since almost all the variability is due to
process noise.

In order to check how much the departures from normality and model type influence
the performance, a smaller simulation study based on the ARMAX$(2,1,1,1)$ model
above with the external signals simulated as AR$(3)$ and AR$(1)$, was also performed.
It shows that the process is very hard to predict, and will give a large amount of
false alarms if a high detection probability is desired. An alarm is denoted false if
it does not predict the catastrophe probability exactly in time. The influence of the inputs are
rather easy to predict when the prediction horizons are short, leading to almost the
same alarm system as for known inputs. The variability of the predictions of process
values is almost entirely due to the influence of the system noise, $e_t$. The results from
the simulation are shown in Table 7. When the wrong model is used the detection
probability can become a lot lower than calculated. This is obvious, especially for the
SETARX model. The performance would have been better if the models had been
estimated on the simulated data and not on the water data. Worth noting is that the
maximal detection probability for the naive-naive alarm system is 0.29, so it is not
comparable to the other alarm systems.

6. Results

The optimal alarm systems for the different models were compared to some simpler
alarm systems. The simplest alarm system, called the naive-naive alarm system, gives
an alarm when the process value $k$ steps before a possible catastrophe crosses a certain
level. This alarm system did work, but not as well as the optimal alarm systems.
The most important disadvantage is that the naive-naive alarm system will have a
maximum detection probability, that cannot be exceeded and is rather low.

Another simple alarm system models the process and gives alarm when the predicted
process values crosses a level, that was determined from the data sets. This alarm
system did not work, hence it has not been included in the tables below.

The optimal alarm system has a nonlinear alarm region, that changes depending on
the catastrophe level and the process values. This makes the optimal alarm system
rather complex. In many cases when the performance is important this is the alarm system that ought to be used. In other cases it might be good to compare a proposed simpler alarm system to the optimal in order to check how close to the optimal the simpler alarm system is.

The parameters describing the process have all been estimated on the data set from 1978, and then tested on all three data sets. The alarm level for the naive-naive alarm system has been optimized over the three data sets together. As can be seen in Table 1 the naive-naive alarm system has a rather low maximal detection probability, and thus is not possible to use if a high detection probability is required. The performance of the optimal alarm systems for these three data sets does not differ very much from each other and they have almost the same number of false alarms. The detection probabilities used are shown in parenthesis. They were in most cases set to 90%. ARMAX2111pred is the alarm system where also the external signals are predicted.

The optimal alarm systems with the highest detection probabilities have quite a few false alarms according to the strictest definition, where an alarm is considered false if it does not predict the catastrophe exactly right in time, but it could be questioned if all of these should be considered false. In Figure 1 it can be seen that a few of the so-called false alarms are early alarms, or alarms given when the levels are still critical. In case of early alarms, at least for one or two steps early which means 3-6 hours early, the additional cost should not be too large. Also, the confidence in the alarm system will not be damaged too much. In the case of alarms when still over the critical level, it means that it will take a little longer to get back to normal state from the emergency state, caused by the process being alarmed. The cost should be small compared to the cost of the catastrophe. If these ideas, i.e. one and two steps early alarms are counted as correct alarms and alarms given when in catastrophe state are not counted at all are taken into account, Table 1 will turn into Table 2.

The alarm level for the naive-naive alarm system is optimized over all three sets. It only reaches a total detection probability of approximately 40 %, which is far below the detection probabilities reached by the different optimal alarms. However, a higher detection probability will inevitably lead to more false alarms, and that is a trade-off that has to be made in each individual case.

In Figure 2 close-ups at some different times are shown to give an explanation for the rather high rate of false alarms. 95% one-dimensional confidence intervals based on the one and two step predictions are also shown. The process is rather hard to predict which leads to wide confidence intervals and a high number of false alarms if a high detection probability is wanted.

7. Conclusions

This paper has presented an optimal alarm for processes described by linear or piecewise linear processes applied to prediction of high water levels in the Baltic. The
optimal alarm technique gives as few false alarms as possible for a given probability of detecting the catastrophes.

Data are collected in the southern part of the Baltic and high water levels in Rødbyhavn in Denmark are to be predicted.

The models that are used to describe the water levels all contain external variables, with future values that are unknown at prediction time. This means that also these external signals have to be predicted, which influences alarm levels and probabilities for detection and for alarm. Three different models for the water level have been considered.

The optimal alarm systems presented in the paper work well, and have the ability to reach any specified detection probability. The more conventional alarm algorithm that the optimal alarm is compared to, i.e. the alarm is sounded when the process reaches a certain level, has a maximal detection probability which in these cases is rather low. This means that if a high detection probability is required, the optimal alarm system has to be used. A drawback with a high detection probability is that the number of false alarms also becomes rather large, even though the optimal alarm systems give a minimum of false alarms. In particular the SETARX model for the water level shows this balance, it has a fast response and detect almost all catastrophes on all datasets, but at the expense of giving a high amount of false alarms, in particular on a dataset (from 1980) to which the model was not adapted.

A possibility to lower the number of false alarms is to find a better model, e.g. by using more external information for the predictions or by taking the timevariations of the water level process into account. Furthermore, in the flooding data case the prediction errors are not exactly normally distributed, which introduces further approximations in the calculations.

ÖZET

Bu çalışmada Gaussian ARMAX süreçleri için yanlış alarmların sayılarının minimize edilmesi anlamında optimal kestiriciler incelenmiştir. Sonuçlar Baltık Denizi'ndeki su seviyelerinin kestirimleri için uygulanmış ve burada verilen optimal alarm sisteminin daha basit sistemlere göre daha iyi sonuç verdiği gözlemlenmiştir.