Ecological Indicators: Software Development

Sergei Rodionov
Joint Institute for the Study of the Atmosphere and Ocean, University of Washington, Seattle, Washington

1. Upgrade to the Sequential Regime Shift Detection Method

The sequential regime shift detection method described in *Rodionov* (2004) was based on the assumption that observations in the series are independent of each other. Many ecological indicators, however, exhibit serial correlation (also referred to as red noise). Due to the presence of red noise, these time series are characterized by long intervals when the observations remain above or below the overall mean value. These intervals can be easily misinterpreted as genuine regimes with different statistics, as illustrated in Fig. 1.

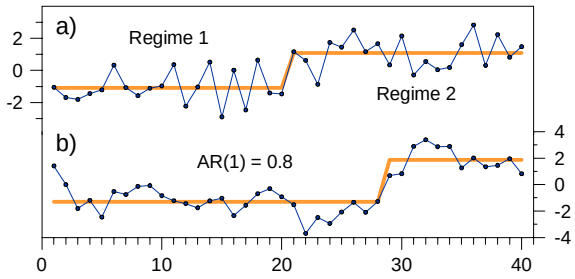

Fig. 1. Realizations of a) white noise process with a shift in the mean at t =21 from -1 to 1, and b) red noise process with AR1 = 0.8. The shift at t = 29 in the latter case would be statistically significant at the $3 \cdot 10^{-9}$ level, if the data points were independent.

There are two approaches to deal with the serial correlation. The first approach is to reduce the degrees of freedom in calculation the significance level of the shifts proportionally to the serial correlation. The second approach is to use a prewhitening procedure, which consists of removing red noise from the time series prior to applying a regime shift detection method. Both approaches require an estimation of lag-1 autoregressive coefficient (AR1). The problem is that when regime shifts are present, using the entire time series often leads to overestimation of AR1. A possible solution to this problem is to break the time series into subsamples, so that the majority of them do not contain change points, and then use the median value of all AR1 estimates.

It is well-known, however, that the conventional estimators, such as the ordinary least squares (OLS) or maximum likelihood techniques, yield biased estimates for AR1, particularly for small samples. Rodionov (2006) discusses two procedures of bias correction of the OLS estimator for short time series. The first procedure is called MPK after Marriott, Pope and Kendall, who proposed a formula for the expected value of the OLS estimator of AR1. The second procedure, called IP4 (Inverse Proportionality with 4 corrections), is based on the

assumption that the first approximation of the bias is approximately inversely proportional to the subsample size and is always negative. Both procedures are included in the new version of the sequential regime shift detection method (Fig. 2). The software can be downloaded from http://www.beringclimate.noaa.gov/regimes.

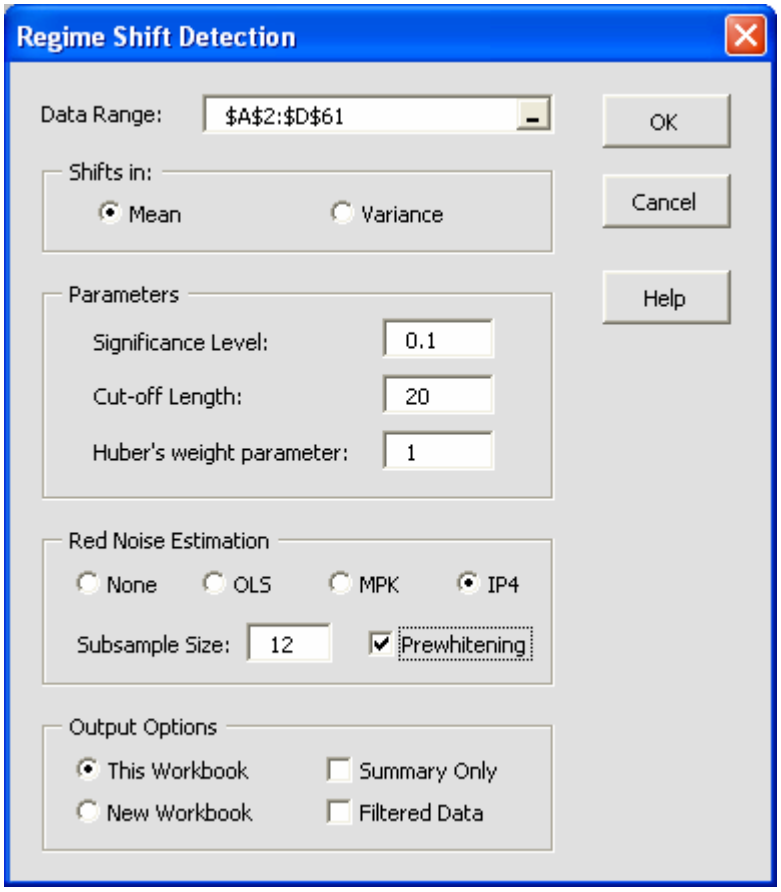

Fig. 2. Entry form of the regime shift detection method.

Extensive Monte Carlo experiments have demonstrated that the MPK and IP4 bias correction techniques produce similar AR1 estimates for subsample sizes greater than 10. For smaller subsample sizes, however, IP4 substantially outperforms MPK in terms of both the magnitude of the bias and variability of the estimates.

To illustrate the effect of prewhitening on regime shift detection, the method was applied to annual series of the Pacific Decadal Oscillation (PDO) index, 1900-2005. Figure 3 illustrates changes in AR1 estimates depending on the bias correction technique and subsample size. The MPK and IP4 estimates are practically the same for subsample size $m > 11$. The estimates remain relatively stable at about 0.45, as m increases to 27. For greater m, AR1 estimates jump to a higher level of about 0.60. This behavior of AR1 is typical for the time series that represent a

mixture of red noise with shifts in the mean. It shows that a characteristic time scale of the PDO regimes is about 25-30 years.

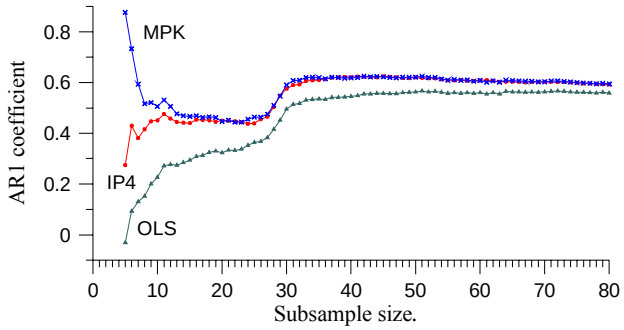

Fig. 3. OLS estimates of the annual PDO index with no bias correction and using the MPK and IP4 techniques.

After prewhitening, statistically significant (at $p < 0.01$) regime shifts in the PDO are still detected in 1948 and 1976, although their magnitudes are smaller than those in the observed time series (Fig. 4). The red noise component (Fig. 4c), which accounts for about 25% of the total variance in PDO, enhances the shifts. The overall conclusion is that the PDO appears to be more than just a manifestation of red noise, as was suggested in some recent publications (Rudnick and Davis, 2003; Hsieh et al., 2005).

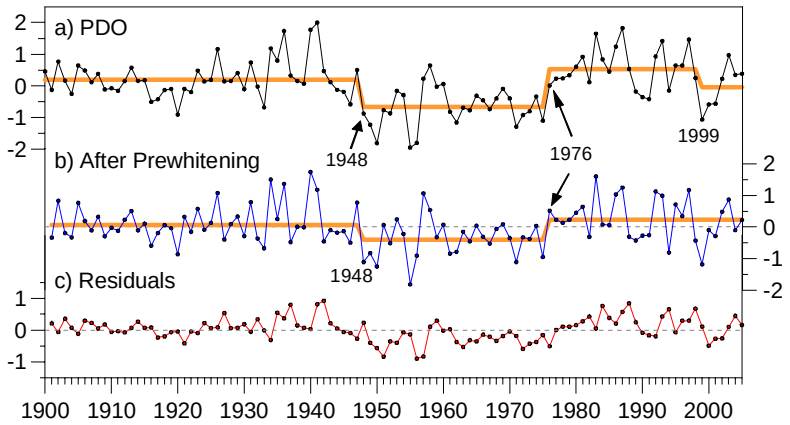

Fig. 4. a) Annual PDO index, 1900-2005, with a stepwise trend, b) the same time series after prewhitening, and c) difference between the time series in a and b.

2. Vitus: Knowledge Management System for the Bering Sea

An increasingly large number of ecological indicators call for methods to deal with the information overload. One large group of methods tries to resolve this problem by reducing the dimensionality of the system. This group includes principal component analysis, singular value decomposition, multidimensional scaling and other methods. These methods proved to be useful in analysis of large sets of indicators (e.g., Hare and Mantua, 2000), although there is often a

problem in interpreting the results. Another important drawback of those methods is that they do not preserve information about the relationships between the indicators.

An alternative approach to the information overload is to use a tool that can help manage information in such a way that only the information relevant to the problem or question at hand is provided to the user at any given point of the analysis. With this in mind, a prototype of a knowledge management system for the Bering Sea ("Vitus") has been developed. The system itself is far from its completion, its data and knowledge bases are not filled, but about 80% of its functionality is in place. It is written in VB.NET with the use of several off-the-shelf Microsoft products: Word, Excel, Access, and Visio.

The major components of Vitus are: Data Explorer, Rule Explorer, Inference Engine, Graphical Interface, Search and Reporting Facilities. In many respects, Vitus is similar to an expert or decision support system, but unlike those commercial expert systems that I am familiar with, both the knowledge presentation and inference process are more transparent to the user and designed to be used in environmental research.

The Data Explorer (Fig. 5) organizes information about indicators based on geographical hierarchy. The user can easily create his/her own geographical domain with the necessary level of details. The data for each variable is kept in a separate Excel file and the descriptive information in a Word file. The user can see a list of rules, for which a selected variable participates in the IF or THEN clauses (Fig. 6). With a click of the mouse, the variable can be inserted into the project, which is visualized as an influence diagram (Fig. 7).

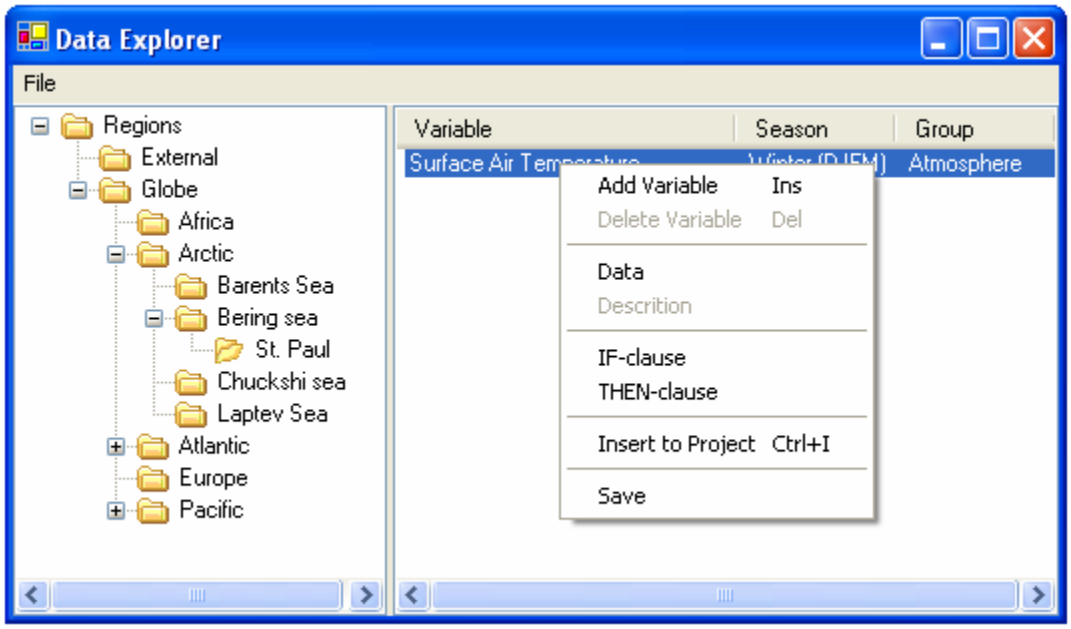

Fig. 5. Data Explorer interface.

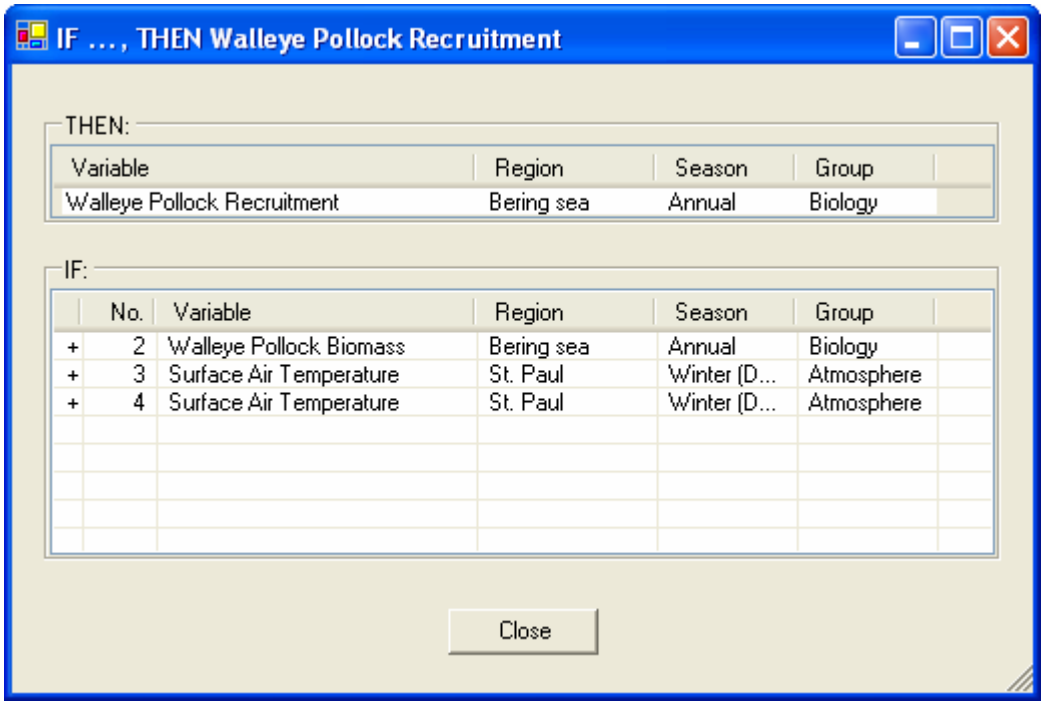

Fig. 6. A list of rules that describe factors affecting walleye pollock recruitment.

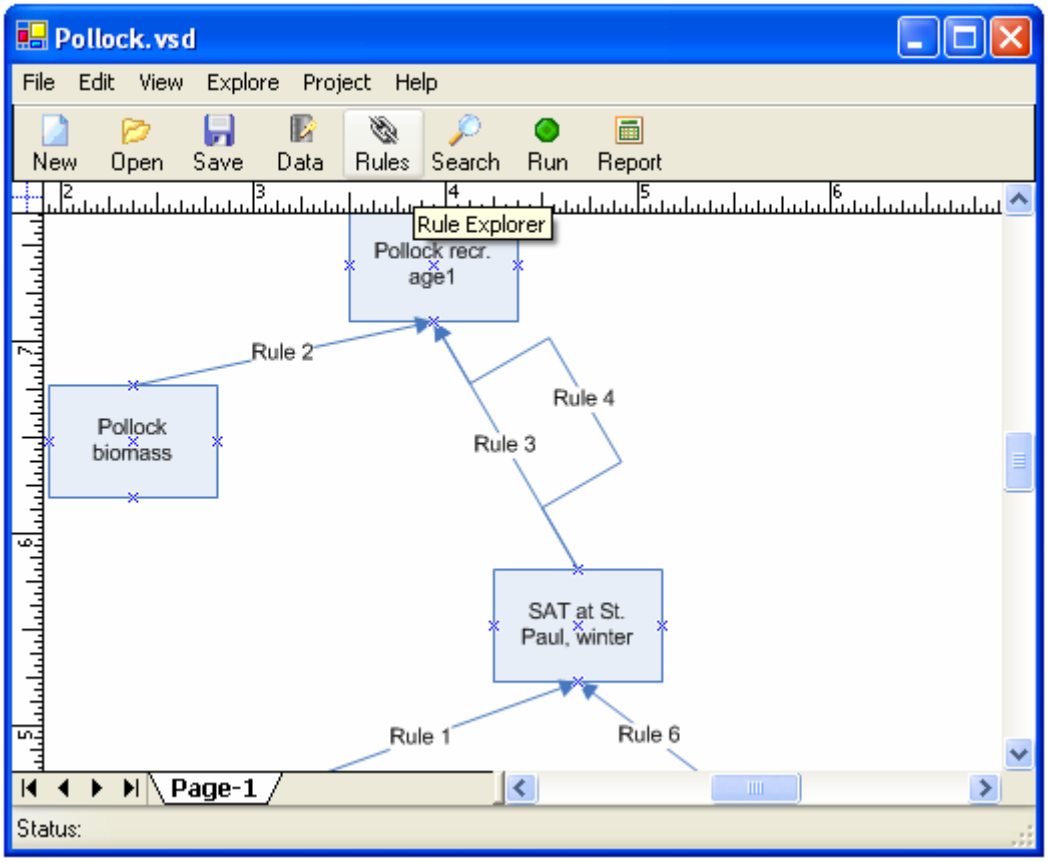

Fig. 7. Part of the influence diagram for walleye pollock recruitment.

The domain knowledge is presented in the form of IF-THEN rules and is controlled via the Rule Explorer (Fig. 8). The number of variables in the IF part of a rule is unlimited. For example, a rule may look like

```
IF ENSO event = warm,
AND Aleutian low circulation type = W1,
THEN SAT at St. Paul = above normal; CF = 10.
```

Here CF is the confidence factor for the rule (more about it is below). It is important to note that the data and code for each rule is placed in a separate Excel file. Therefore, although the IF-THEN form is default, the user can write his/her own code to express the relationship between the IF and THEN variables. For example, the user can program the Ricker stock-recruitment formula, or use linear regression instead of a simple IF-THEN relationship. Another advantage of this rule information storage is that the user can easily experiment with each rule separately and develop a better feeling of confidence in it.

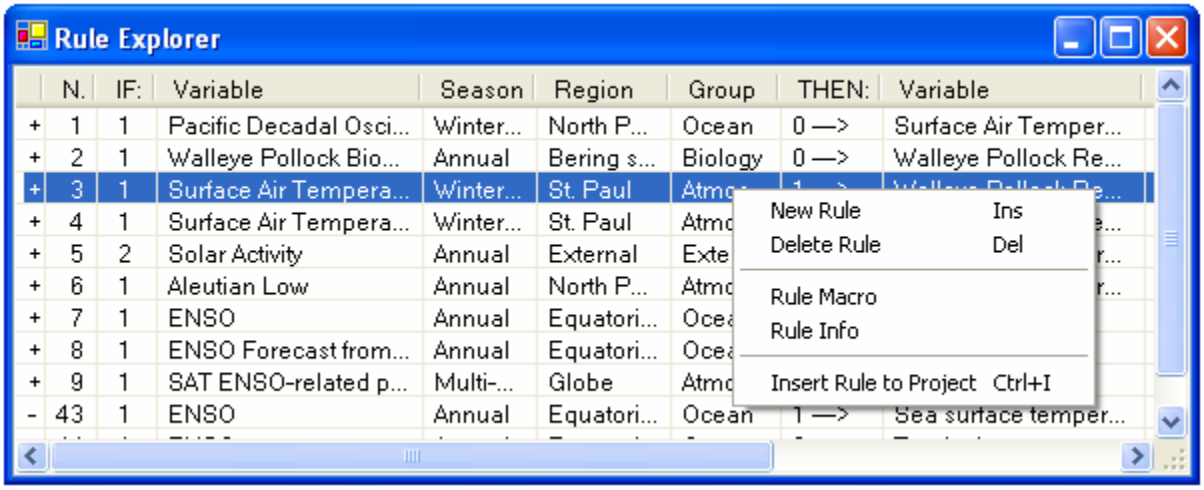

Fig. 8. The Rule Explorer.

When the influence diagram is prepared, the user may run the project in the forecast or hindcast mode to infer the value of the target variable in a given year (Fig. 9). During this process, the system asks for the information about the variables in the terminal nodes of the diagram (Fig. 10). To facilitate the answer to those questions, the user is provided with the access to the data and descriptive information about the variable and related rule. The user can also search for any other pertinent information (Fig. 11).

Previous experience of working with climatic expert systems (Rodionov and Martin, 1996; 1999) showed that, in assigning confidence factors to the rules, it is important to maintain the relative importance of each rule in the system. In other words, it is not the numbers themselves, but the consistency in procedure of their assignment, should be of major concern to the user. Therefore, although the CF is equivalent to the subjective probability, whenever possible, it is recommended to estimate its value based on the formula:

$$CF = (P(C \mid e) - P(\neg C \mid e)) * 100\%,$$

which is the difference between the probability of category C of the variable given the evidence e and probability of any other category of the variable given the same evidence e, expressed in

percent. When CF = 0, it means that observing e will not change our prior confidence (if any) in C. The value of CF = 100 means that we can be 100% confident in C, given the evidence e. The confidence factors, calculated using the above formula, should be adjusted for the number of observations. The formula for adjustment (A) used here is as follows:

$$A = 100 - \log (N)/2 * 100,$$

where N is the sample size.

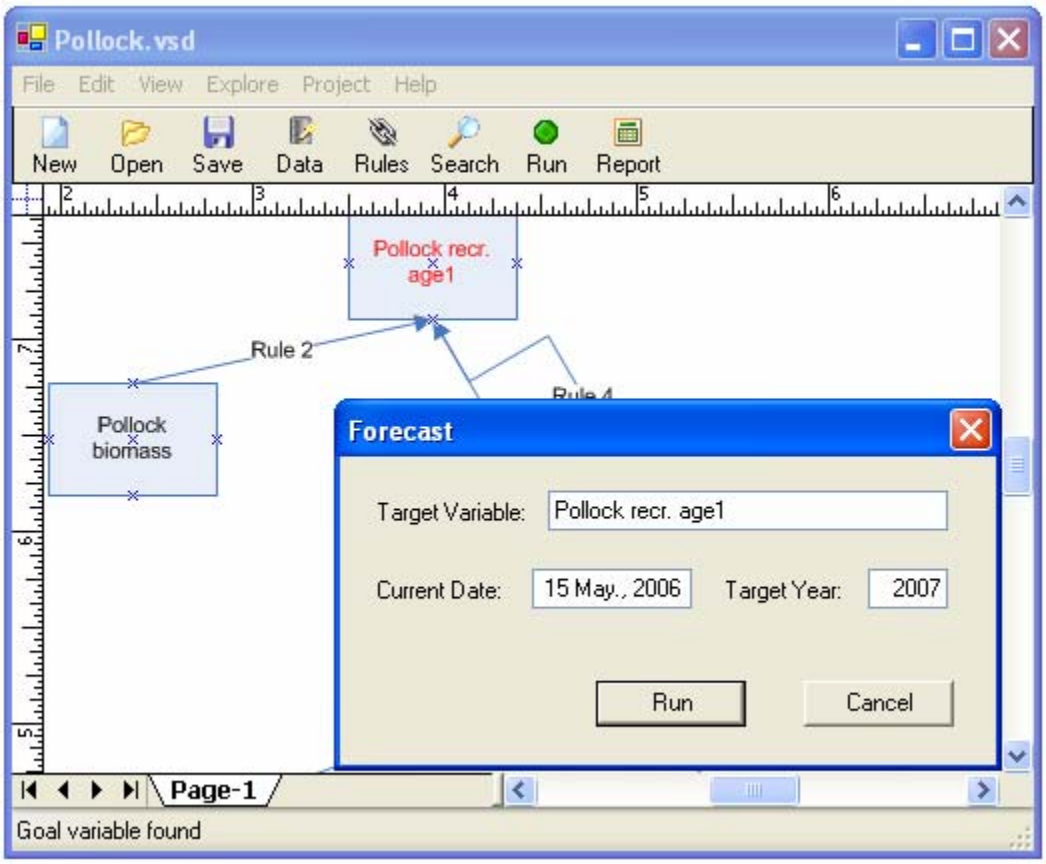

Fig. 9. Running the project in the forecast mode.

As an example, Table 1 shows the contingency table for the Pacific/North American (PNA) teleconnection index and North-South winds at St. Paul. The CF for anomalously strong northerly wind in the case of positive PNA will be

$$CF (Wind+ \mid PNA+) = (17/24 - 7/24) * 100\% = 42,$$

and after adjustment

$$CF_{adj} (Wind+ \mid PNA+) = 42 - 100 - \log (24)/2 * 100 = 42 - 31 = 11.$$

The value of CF_{adj} (Wind– | PNA–) is calculated similarly, so that the rule for these two variables will be as follows

```
IF PNA index = positive (negative),
THEN NS wind anomaly = positive (negative); CF = 11 (9).
```

Table 1. Contingency table for the Pacific/North American teleconnection index and North-South winds at St. Paul (Pribilof Islands). Both variables are broken into two categories of above and below normal values. Data: 1949-2005.

NS wind anomaly	PNA +	PNA −	Total
Wind +	17	11	28
Wind −	7	22	29
Total	24	33	57

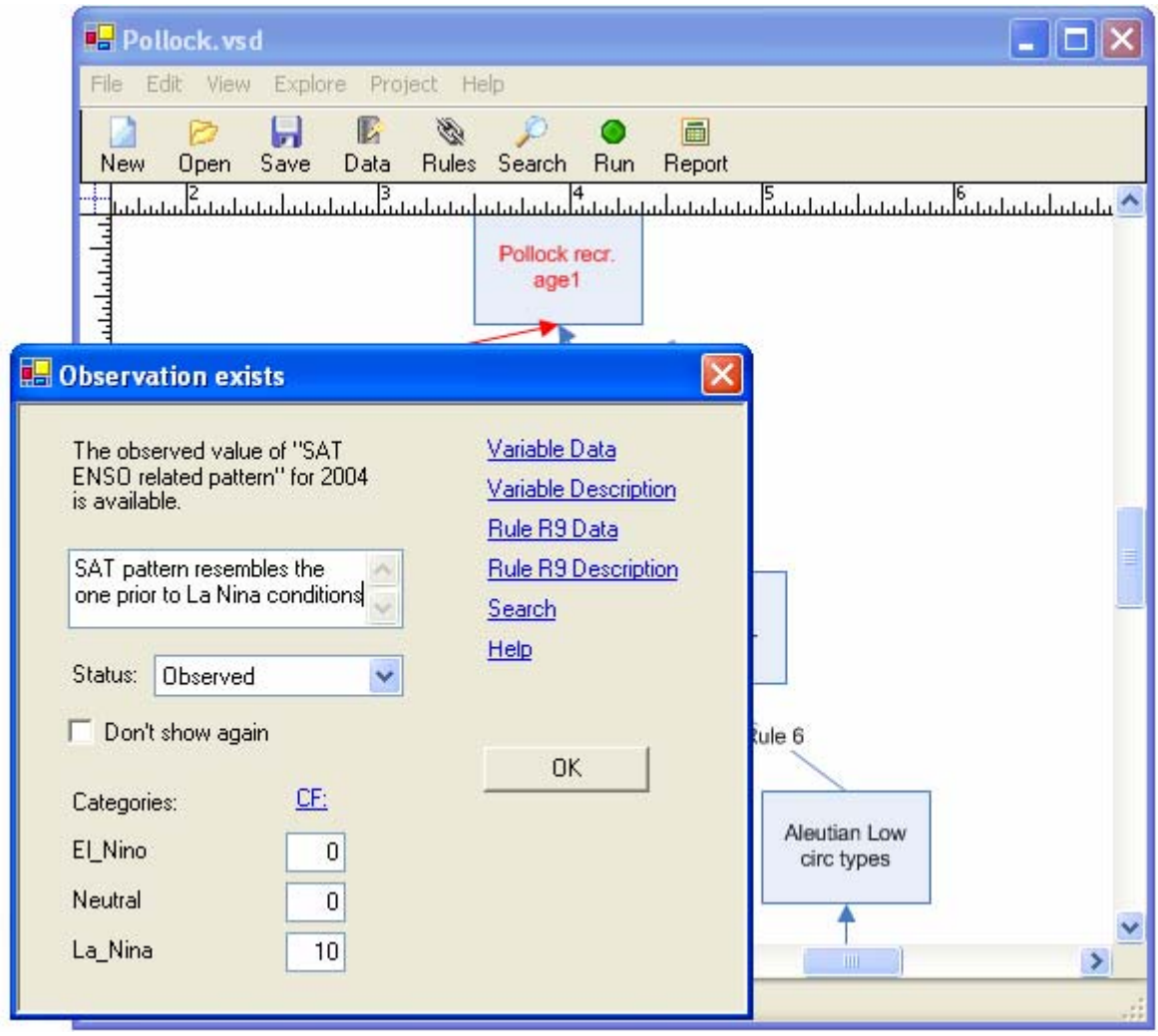

Fig. 10. The system asks questions about the variables in the terminal nodes.

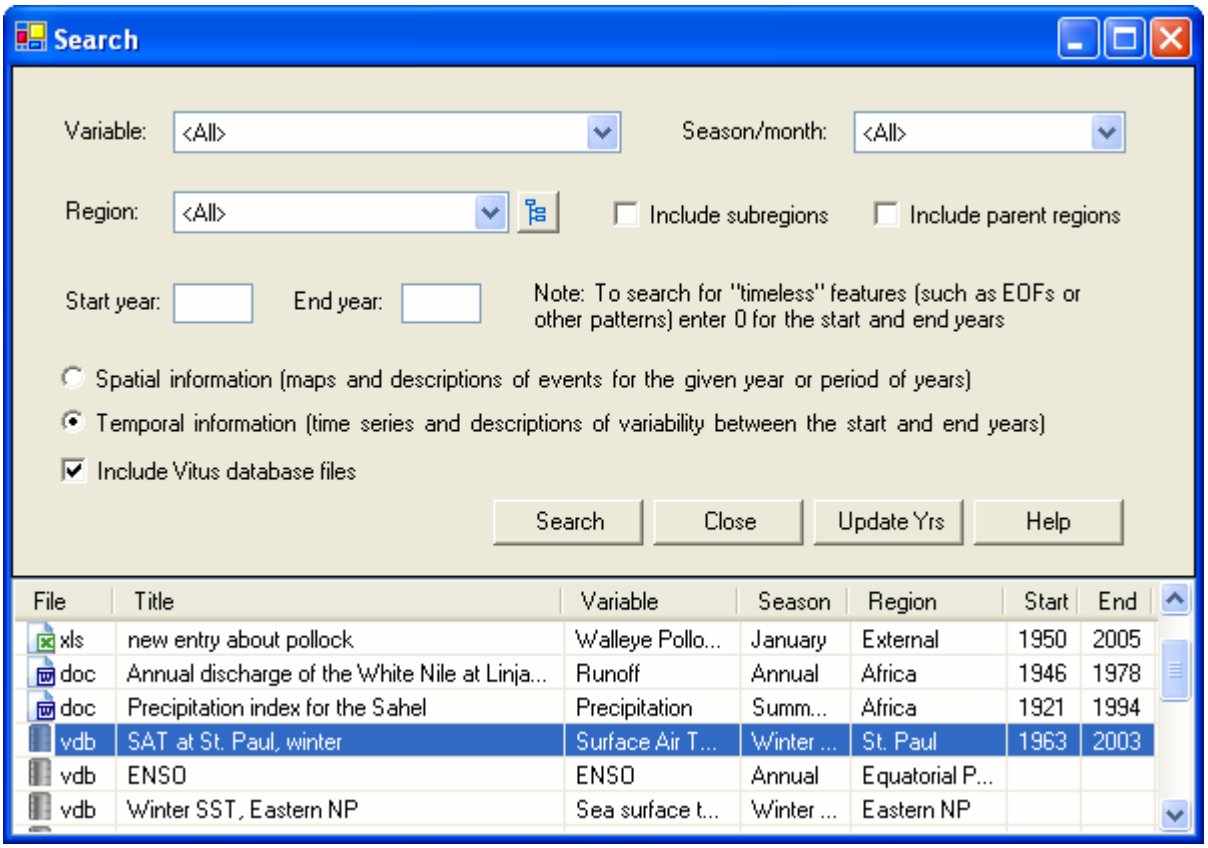

Fig. 11. Search for the relevant information.

The evidence from different sources is combined using the following formula:

$$CF_{comb} = CF_{old} + CF_{new} - (CF_{old} * CF_{new})/100.$$

In addition to the CF algebra, Bayesian inference technique may be added later. When all the evidence is collected, a forecast for the target variable is issued either in the form of odds (e.g., strong versus weak year class of walleye pollock) or probabilities. The user can also open the Custom Property window (Fig. 12) and check the information about individual variables and rules, or open the report that traces the logic behind the forecast (Fig. 13).

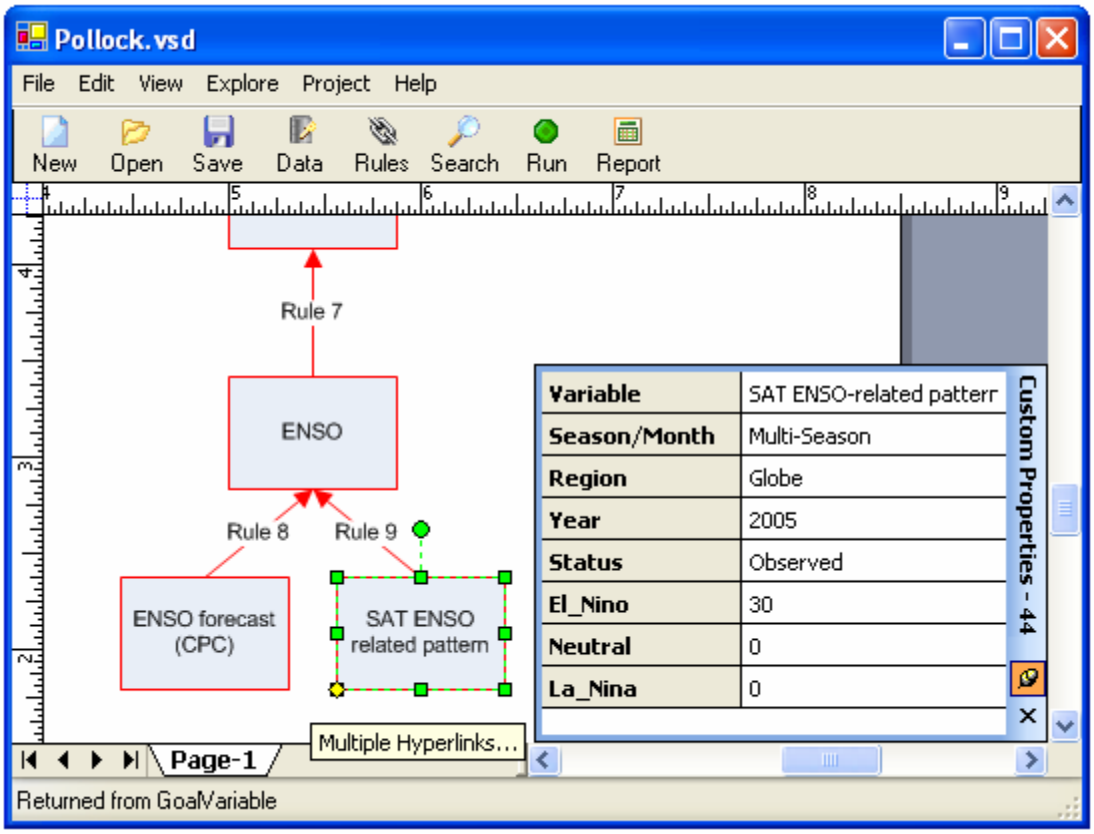

Fig. 12. Displaying information about the variables and rules in the property window.

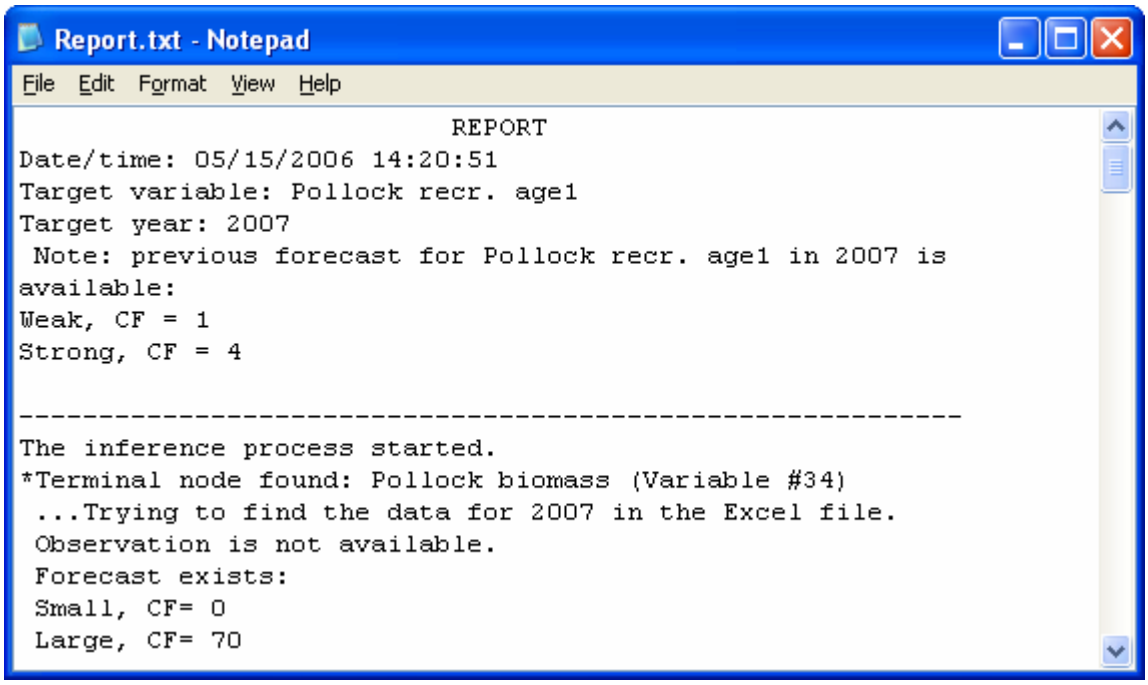

Fig. 13. An example of the report that traces the inference procedure.

References

Hare, S. R. and N. J. Mantua, 2000: Empirical evidence for North Pacific regime shifts in 1977 and 1989, *Progr. Oceanog.*, **47**, 103-146.

Hsieh, C. H., S. M. Glaser, A. J. Lucas, and G. Sugihara, 2005: Distinguishing random environmental fluctuations from ecological catastrophes for the North Pacific Ocean, *Nature*, **435**, 336-340.

Rodionov, S., 2004: A sequential algorithm for testing climate regime shifts, *Geophys. Res. Lett.*, **31**, doi:10.1029/2004GL019448.(L09204).

Rodionov, S., 2006: The use of prewhitening in climate regime shift detection, *Geophys. Res. Lett.* (in print).

Rodionov, S. N. and J. H. Martin, 1996: A knowledge based system for the diagnosis and prediction of short term climatic changes in the North Atlantic, *J. Climate*, **9**, 1816-1823.

Rodionov, S. N. and J. H. Martin, 1999: An expert system-based approach to prediction of year-to-year climatic variations in the North Atlantic region, *Int. J. Climatol.*, **19**, 951-974.

Rudnick, D. I. and R. E. Davis, 2003: Red noise and regime shifts, *Deep-Sea Research*, **50**, 691-699.