printed by NIOZ *texel*

A METHOD FOR VALIDATION

by

David Stroo

This report is the result of a 6 weeks study by David Stroo, student at the faculty of Industrial Engineering and Management Science at the Eindhoven University of Technology, in commission of Biological Research Ems-Dollard Estuary (BOEDE) (Netherlands Institute for Sea Research) under guidance of drs. P. Ruardij.

CONTENTS

1. Summary

Aim of this investigation was to give a contribution to the validation of the BOEDE simulation model. The BOEDE model is an ecosystem simulation model, simulating the carbonflux through the Ems-Dollard estuary system. There seems to be a variety of statistical tests to validate a simulation model but most of them have a restriction: autocorrelation. Most statistical tests assume the absence of autocorrelation. Still, autocorrelation is often present in sample data.

In this report a description is given of a method to calculate a figure for the "measure of validity" (usefulness) of a predictor compared to a set of observed values. There certainly is agreement between this figure and the (subjective) judgement of the BOEDE researchers. Though this figure gives an indication of the usefulness, the judgement of people who are directly involved with the actual process is always better.

The method consists of three parts:

1) A figure for the deviation between the actual and simulated data due to a random component.
2) A figure for the correlation.
3) A figure for the relative deviation.

The final figure is the average of those three parts.

Acknowledgements

I would like to thank all the people who helped me to make this a reality on very short notice, especially Piet Ruardij!

2. Aim of this investigation

My task was to give a contribution to the validation of the BOEDE simulation model. A division of the model into two groups, a valid and an invalid group, was certainly not desirable. Too much information would then be lost. A figure for the "measure of validity" was preferable. Such a figure could then be used as a substitution of the validation pictures (actual and predicted values in one picture) and to see the overall effect of changes in the model. So my aim was to develop objective criteria for validating this ecosystem model.

Validity however cannot be measured objectively. It is subjective both to the aim of the model and to the aim of the individual user. So, in a way, the validity figure must be determined by the user and his aim.

This investigation consisted of the following phases.

- Investigation of the literature about validation of simulation and forecasting models. Of course also non-biological literature. Especially economic literature proved to be usefull.
- Selection of a method or methods from the literature.
- Adaption of the method(s) to this specific model. (With the help of the makers/ users).
- Implementation.

3. Introduction to Biological Research Ems-Dollard Estuary

An estuary is an inlet of the sea reaching into a river-valley as far as the upper limit of the tidal rise. Estuaries and their environments have always been very important for human settlement. But with the enlargement of settlements and industrialization estuaries became more loaded with organic and inorganic waste. The BOEDE project was initiated in response to awareness of increasing pollution in the Ems- Dollard estuary in the Netherlands.

The aims of the project were threefold: (Biologisch onderzoek Eems-Dollard estuarium, 1983)

1) To trace and describe the factors that determine the structure and function of the Ems-Dollard estuarine ecosystem. To do this, the present ecosystem had to be quantitatively and qualitatively described and the processes in the ecosystem had to be studied.
2) To study the changes that have appeared in the ecosystem as a result of human influence, and to develop methods of predicting future changes, preferable using a mathematical model.
3) To use the results obtained from 1 and 2 to create a management strategy for the Ems-Dollard estuary and, if possible, for similar areas.

At the time of this investigation the BOEDE mathematical model was near to completion. The most important phenomena in the Ems-Dollard ecosystem can be described by the model. To make that model it was necessary to know the different organisms, their predator-prey or other relations and how they are influenced by the physical or chemical environment. All this resulted in a big and complex model and made a computer indispensable. So the model was programmed in the NIOZ computer system (Norsk Data Nord-100).

With certain inputs (light etc.) a prediction can now be made of the total organic carbon, phytoplankton, pelagic bacteria etc.. Refer to appendix 2 for some examples of the BOEDE variables. The predictions can be plotted with their concentration at the y-axis and the time at the x-axis. Field observations can be plotted into the same picture if wanted. This to compare field observations and predictions which can lead to acceptance or rejection of certain theories and/or assumptions in the model. Rejection can lead to a new or better investigation. Obviously this "validation" is very important.

4. The literature, a view of the possibilities and comments

4.1 Introduction to validation

Validation is the process of building an acceptable level of confidence that an inference about a simulated process is a correct or valid inference for the actual process. Validation clearly applies to a far more general environment than simulation; validation is a problem associated with all modeling. There is no such thing as "the test". The experimenter selects a set of tests from the many possible. (Horn, 1971).

Ideally a comparison test should handle nonstationarity, compensate for noisy data, simultaneously evaluate a number of output measures and work for small samples. Does such a test exist? The answer is yes if one is willing to define test very broadly. The test is simple. Find people who are directly involved with the actual process. Ask them to compare actual with simulation output. This test is sometimes attributed to Turing (1950).

But this Turing test is not what we are looking for. Our aim was to develope objective criteria!

Mankin makes a difference between a valid model and a useful model. A valid model has no behaviour which does not correspond to system behaviour; a useful model predicts some system behaviour correctly. Since no model is perfect, available ecosystem models may be described as invalid but useful models. (Mankin et al., 1975).

Validation is not something to be attempted after the simulation model has already been developed and only if there is time and money still remaining. Instead model development and validation should be done hand in hand throughout the course of simulation study. This recommendation is often not followed. (Law, 1982).

The problem of validation has at least two different points of view.

a) Criteria that consequence from the aim of the simulation model. These validity criteria are described in section 4.2.
b) Criteria that consequence from the theory of scientific inquiry. This because the development of a model and model testing is inseperable of one's theory of scientific

inquiry. These theories of scientific inquiry are described in section 4.3.

4.2 Types of validity criteria

A limited exploration of the types of validity criteria (Hermann, 1967).

1) Internal validity.
   All the exogenous inputs are held constant across all runs. The unexplained variance between these intended replications would provide a measure of reliability or what is called "internal validity". If the observed results (output) of an operating model can be attributed to extraneous factors rather than the specified relationships in the simulation, then its internal validity is low. If the inputs of the BOEDE model are being held constant it results in perfect replications. Consequently the internal validity of the model is high.
2) Face validity.
   Face validity is a surface or initial impression of a simulation's realism. Probably no approach to model validity is reported more frequently than the subjective estimates of experimenters or observers as to the correspondence between the model's operation and their perception of the phenomena which the simulation represent. But the face validity in its usual form suffers from the lack of explicit validity criteria. Face validity has been used quite often in the development of the BOEDE model.
3) Variable-parameter validity.
   This is the comparison of the simulation's variables and parameters with their assumed counterparts in the observable universe. Sensitivity testing is a feature of variable parameter validity. In repeated runs of a simulation the setting of a parameter or the range of values assigned a variable are systematically changed to determine what difference, if any, the alteration has on the operation of the model. The variable parameter approach has the advantage of isolating individual components of the simulation. It is thus possible to determine what particular features may be reducing the representativeness of the operating model. But the sensitivity procedure is quite laborious and for a complex model it can be almost endless. The BOEDE researchers haven't used

the sensitivity testing very thoroughly yet, if at all.

4) Event validity.
This approach employs "natural" events as criteria against which to compare outcomes occuring in the simulation. To the extent that events can be equated with the consequences or end products of a simulation, they provide the material of immediate relevance to the investigator with interest in prediction. Because (biological) events are the result of interaction among numerous elements in the mode, event validation may be quite useful for checking the total simulation, that is, the composite set of interrelationships. By the same reasoning, however, event validation may be less useful for discovering the exact parts of an operating model responsible for incongruities between the simulation events and those in the reality to which it is compared. Since biological models predict "natural events" the event validity is often used.

5) Hypothesis validity.
In this approach, hypothesized relationships become the validity criteria. If X is observed to bear a given relationship to Y in the observable universe, then X' should bear a corresponding relationship to Y' in a valid operating model. Hypothesis validity differs from parameter-variable validity or event validity in that the criteria are not individual entities, but connections between two or more units. This method is used in comparing and judging the relations of the predicted variables in the BOEDE model.

Despite the criteria mentioned above there are still no explicit criteria set to make a selection between a useful an a useless model. As Shannon (1975) writes: "Despite extensive literature dealing with validation procedures, the problem of validating simulation models remains as difficult and elusive as ever. But it is an issue we cannot avoid or push lightly aside".

Churchman (1968) maintains that one's theories of model building and model testing are inseperable aspects of one's theory of scientific inquiry. If this is true, then it would perhaps be useful to look briefly at some different theories of scientific inquiry.

4.3 Theories of scientific inquiry

1) Subjective vs. objective methods.
   There is an apparent conflict when we are designing and validating simulation models between the need to be objective and the need to make constructive and intelligent use of our subjective beliefs (insights, intuitions, impressions etc.).
2) Rationalist vs. empiricist.
   Throughout the history of science, rationalists and empiricists have battled over the correct method to conduct scientific inquiry. Both agree that science begins with the observation of selected parts of nature but there the agreement ends. Should the scientist, after observation, postulate the way elements of the system interact and then discover whether the facts fit the hypothesis (rationalist), or should he only include those interactions which can and have been tested empirically (empiricist).
3) Absolute pragmatists.
   The absolute pragmatist, in the pure form, says basically, "I am building the model for a specific purpose or use. If it fulfills that purpose, then it is a valid model".

Very seldom does one come across a pure rationalist, empiricist, or absolute pragmatist. Most experimenters find themselves in the position of being willing to use and be concerned with all three points of view.

Naylor and Finger (1967) have designed a method which contains the validity criteria as well as the theories of scienific inquiry. It's called the multistage verification and is described in section 4.4.

4.4 Multistage verification

Stage 1

This stage is to seek face validity of the internal structure of the model based upon a priori knowledge, past reseach and existing theory. In general, most complex simulation models consist in modeling a large number of "simple" processes. When they are combined the large number of possible interactions makes understanding of the behavior of the total system impossible. Thus, the first stage of validation entails looking at each of the

"simple" processes modeled to ensure that the building blocks, so to speak, are the best possible. Any hypothesis that can be rejected upon the basis of a priori knowledge or past research should be so rejected until additional research or experience modifies our belief.

Stage 2

The second stage is also concerned with the validation of the internal structure of the model, and consists in empirically testing, wherever possible, the hypothesis used. In this stage, we attempt to verify as many as possible of the assumptions that survived stage one, by subjecting them to vigorous emperical testing. The theory of statistics, as it pertains to estimation and hypothesis testing, provides the framework for this stage of validation based upon the empiricists viewpoint.

Stage 3

The third stage attempts vigorously to verify the model's ability to predict the behavior of the real world system. Here we are faced with convincing the user that our model does what we claim it will do, that it is useful. In general, it entails comparing the input-output transformations generated by the model with those generated by the real world system. (Event validity).

In the case of the BOEDE simulation model, stage 1 and 2 highly concern biological theories and theses. As I am not a biologist, I will leave those stages to the BOEDE researchers. Stage 3, comparing the input-output transformations, is more general. Thus, my research will be directed towards stage 3. I know that all three stages occurred in an iterative manner throughout the model development process. But stage 3 only in a "Turing"-like test (Thuring, 1950). (Using the impressions of the researchers). Stage 3 has been described in section 4.5.

4.5 Comparing the input-output transformations (Stage 3)

Suppose that Ai (actual) are observations from a real-world system and that Pi (predicted) are output data from a corresponding simulation model. We would like to compare the two data sets, in some sense, to determine whether the model is an accurate representation of the real- world system. The

first approach which comes to mind is to use one of the classical statistical tests (t, Mann-Whitney, two-sample chi-square, two-sample Kolmogorov-Smirnov etc.) to determine whether the two data sets can be safely regarded as being the same. However, the output processes of almost all real-world systems and simulations are nonstationary and autocorrelated and thus none of these tests is directly applicable (Law & Kelton, 1982). When autocorrelation is present in sample data, the use of classical statistical estimating techniques (which assume the absence of autocorrelation) will lead to underestimates of sampling variances (which are unduly large) and inefficient predictions. (Naylor & Finger, 1967) The BOEDE model is also highly autocorrelated.

There is however one statistical test which might be suitable. It is postulated by prof. dr. R. Doornbos (Bosch et al., 1982), Technological University Eindhoven. I will call this test 1.

Test 1) Suppose a certain predicted and actual data set is available (Pi resp. Ai). Suppose the number of Ai is n. Divide the n Ai's into k groups and assume that the Ai's in the group are destributed normally.
For instance:

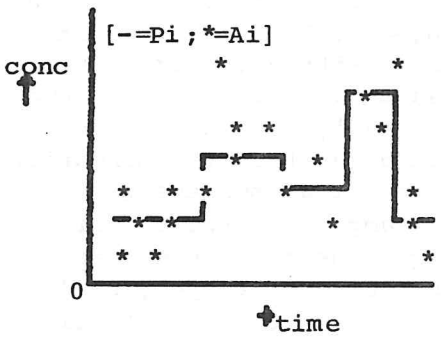


Fig. 1.

Here the Ai's can be divided into 5 groups. (Conc means concentration).

Then you can test $\frac{S_1^2 - S_2^2 * (n-k)}{S_2^2 * (k-1)}$ having a $F_{(n-k)}^{(k-1)}$ distribution.

With $S_1^2 = \Sigma(Ai-Pi)^2$ and $S_2^2 = \Sigma(Ai-\bar{A})^2$ (summation per group) and

$\bar{A} = \frac{\Sigma(Ai)}{n}$.

The assumption of the normal distribution is not even the weakest point of the F-test. The division into k groups is a very subjective one! Still this method can be applied if wanted.

Naylor & Finger (1967) described three other tests in "Verification of computer simulation models".

Test 2) Regression analysis.

It is the possibility of regressing actual series on the generated series and testing whether the resulting regression equations have intercepts which are not significantly different from zero and slopes which are not significantly different from unity. Obviously the regression of the actual series on the perfect predictor will give intercept 0 and slope 1.

Test 3) Spectral analysis.

Data generated by computer simulation experiments are usually highly autocorrelated. Spectral analysis considers data arranged according to historical time. It is essentially the quantification and evaluation of autocorrelated data at which spectral analysis is aimed after the data have been transformed into the frequency domain. Spectral analysis provides a means of objectively comparing time series generated by computer model with observed data. More information can be found in Fishman & Kiviat (1967). But as van Horn (1971) stated, spectral analysis faces several problems. First it requires a large number of observations. The cost of data collection on an actual process may preclude obtaining a sufficient sample for the use of spectral techniques. Another requirement is even more restrictive. The procedures described above apply to "covariance stationary" processes. Also a very high level of mathematical sophistication is required in order to apply it and this method can be expensive to apply in terms of computer time or storage. (Law & Kelton, 1982).

Test 4) Theil's inequality coefficient (Theil, 1961).

A technique developed by Theil has been used by a number of economist to validate simulations with econometric models. Theil's inequality coefficient "U" provides an index which measures the degree to which a simulation model provides retrospective predictions of observed historical data. U varies between 0 and 1. There is no obvious reason why this method cannot be used to validate biological models.

Theil's coefficient (test 4) and the regression analysis (test 2) (which is also published by Theil) seem to be the most promising tests.

5. Theil methods and additions

5.1 Theil's methods

Theil (1961) suggests that it is not unreasonable to measure the seriousness of a given forecast error by its square. Consider then $U1^2=1/n*(\Sigma(Pi-Ai)^2)$ which is the mean square prediction error (MSE) for the set of all n observations. Theil defines the Ai's and Pi's as actual and predicted changes because those are often used in an economic model. But as the BOEDE model calculates values we prefer to use values. And I see no reason why this method cannot be used with the actual and predicted values.

Let us define a perfect predictor as a predictor which can predict the actual values exactly. An optimum predictor is the best predictor for a given actual data set. It is obvious that MSE=0 for the perfect predictor. But an optimum predictor will not necessesarily induce the mean square prediction error to be zero. Almost always some kind of "noise" will be involved (for instance random observation noise).

For instance:

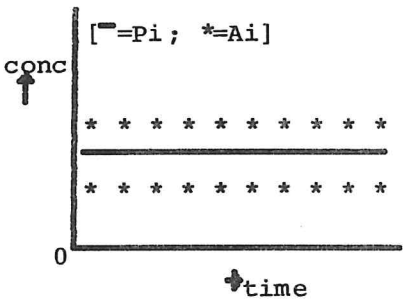


Fig. 2.

(Obviously this picture is not representative for a BOEDE simulation.)
This Pi is the optimum predictor for Ai without the mean square prediction error to be zero. But the MSE is minimal!

Theil stated the following. Suppose that the forecaster will frequently distinquish between a systematic an a non-

systematic part of the realized changes (values) and that his prediction efforts will then be concentrated on the systematic part. Suppose further that the nonsystematic part can be regarded as a random variable with zero expectation. This means that each realized change (value) Ai consists of a systematic part which coincides with the corresponding predicted change (value) Pi and a nonsystematic part which is zero on the average and completely unrelated to Pi. In other words, if we could compute a least-squares regression of the observed changes (values) Ai on the predicted changes (values) Pi the result would - ideally - be of the form Ai=Pi + residual; that is the regression coefficient of Pi would be one and the constant term zero.

Consider the following decomposition (also see appendix 1):

$$\frac{1}{n}\Sigma(Pi-Ai)^2=(\bar{P}-\bar{A})^2 + (Sp-R*Sa)^2 + (1-R^2)Sa^2.$$

with $\bar{P}=\frac{1}{n}\Sigma Pi$; $\bar{A}=\frac{1}{n}\Sigma Ai$; $Sp^2=\frac{1}{n}\Sigma(Pi-\bar{P})^2$; $Sa^2=\frac{1}{n}\Sigma(Ai-\bar{A})^2$;

$$R=\frac{1}{n}*\frac{\Sigma(Pi-\bar{P})(Ai-\bar{A})}{Sp*Sa}$$

Divide the three terms of the decomposition by MSE and we have three "inequality" proportions.

$$\text{So: } \frac{(\bar{P}-\bar{A})^2}{MSE} + \frac{(Sp-R*Sa)^2}{MSE} + \frac{(1-R^2)Sa^2}{MSE} = 1$$

The first term describes the mean and is called the mean component (MC). The second is called the slope component (SC) and the third is called the residual component (RC).

Definitions:

$$MC=\frac{(\bar{P}-\bar{A})^2}{MSE} \qquad SC=\frac{(Sp-R*Sa)^2}{MSE} \qquad RC=\frac{(1-R^2)Sa^2}{MSE}$$

Let us consider this decomposition in relation to the regression Ai=Pi + residual (the optimum predictor). Since the residuals have zero mean, the mean of the A's and P's must be equal, so that the first term (MC) vanishes. Furthermore, the regression coefficient takes the form:

$$\frac{\Sigma(Pi-\bar{P})(Ai-\bar{A})}{\Sigma(Pi-\bar{P})^2} = \frac{R*Sa}{Sp}$$ (Also refer to Bosch et al., 1982)

If this coefficient is indeed 1, the second term (SC) vanishes. So, when perfect predictions (MSE=0) cannot be obtained, the desirable distribution of MSE over the three sources is MC=0, SC=0 and RC=1 indicating that errors are not systematic.

When researchers seek the optimum predictor if noise is involved then this method is the best. For instance:
MC=0, SC=0, RC=1 means: the optimum predictor has been found.
RC<1 means: there is still a systematic error left.
MC=0.8, SC=0.1, RC=0.1 means: a better predictor exists;
80% of the MSE is due to the mean component. (Prediction is too high/low).
10% of the MSE is due to the slope component. (Prediction has the wrong slope).
10% of the MSE is due to a random component.

It is clear that RC can be used as a figure for the prediction. Obviously it varies between 0 and 1 in wich 0 means bad and 1 means good.

There are however four problems. If those problems can be solved, if only partly, then the RC can be a basis for the judgement of the "usefulness" of the BOEDE simulations.

Those four problems are:

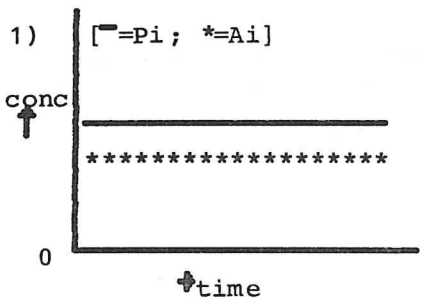


Fig. 3.

This will result in RC=0 because there is no MSE due to a random component. Here 100% of the MSE is due to the mean component. Biologists however may consider this predictor as rather good because the slope is correct. In other words, the

predictor gives the "right" direction and therefore deserves more than 0.

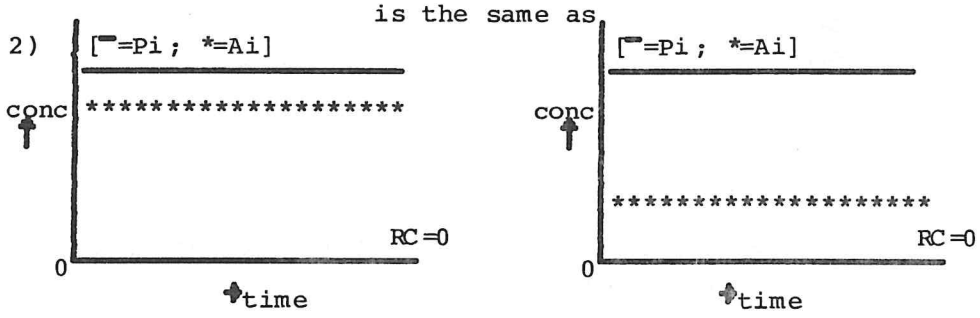


Fig. 4. Fig. 5.

This method does not take into account the (relative) amount of MSE.

3) Even when there is a little noise this method can be dangerous because a slight deviation might be judged too severely.

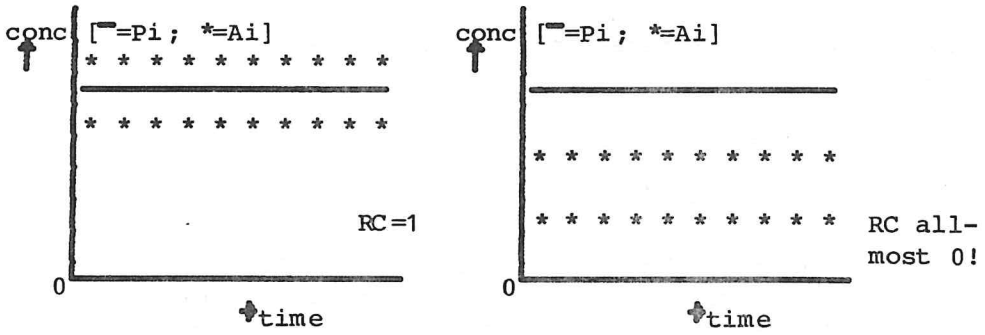


Fig. 6. Fig. 7.

4) RC is not defined for MSE=0.

5.2 Additions

The RC-component as a judgement of the usefulness of a model can be improved by two additions.

Addition 1 The correlation index (R).

R can be written as $R = \frac{Sap}{\sqrt{(Saa*Spp)}}$

with $Saa = \Sigma Ai^2 - ((\Sigma Ai)^2/n)$ and $Spp = \Sigma Pi^2 - ((\Sigma Pi)^2/n)$
and $Sap = \Sigma Ai*Pi - (\Sigma Ai \Sigma Pi)/n$.

R varies between -1 and 1. We want R to be a figure between 0 and 1 so we write $R^* = (R/2) + 0.5$. Now R^* varies between 0 and 1 in wich 0 means bad and 1 means good.

Little noise but a good direction will cause R^* to be high! So R will be high in Fig. 3, 4 and 5.

Addition 2 An index for the relative MSE.

In "Applied economic forecasting" Theil (1966) described:

$$U^2 = \frac{\Sigma(Pi-Ai)^2}{\Sigma(Ai)^2}$$

And in "Economic forecasting and policy" Theil (1961) described:

$$U^2 = \frac{\Sigma(Pi-Ai)^2}{\Sigma(Pi)^2 + \Sigma(Ai)^2}$$

Both figures describe a sort of relative MSE.
Both methods take the actual data set into account as "weight" to calculate the U^2. This is not advisable.
For instance.

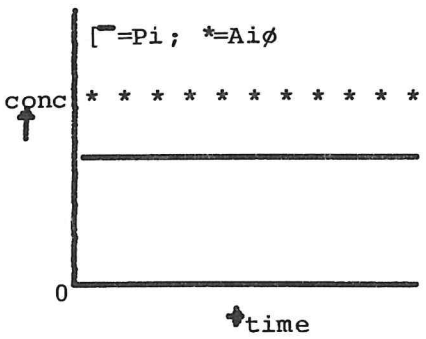


Fig. 8.

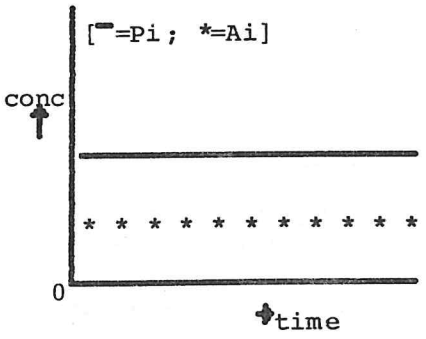


Fig. 9.

So in fig. 8 both calculations of U are being divided by a bigger number than in fig. 9. Since the deviation of Ai in relation to Pi is the same the outcome of U should be the same in both cases.

I would prefer

$$U^2 = \frac{\Sigma (Pi-Ai)^2}{\Sigma (Pi)^2}$$

because this method gives the same value of U for a deviation of Ai below and above Pi.

It is clear that U has no upper limit so define U* as

$$U^* = \frac{1}{1+\sqrt{\frac{\Sigma (Pi-Ai)^2}{\Sigma (Pi)^2}}}$$

U* now varies between 0 and 1 in wich 1 means good and 0 means bad.

From the previous text it seems that there are 3 components that are likely to contribute to the figure that is to be determined.

$$RC = \frac{(1-R^2)Sa^2}{MSE}$$

$$R^* = R/2 + 0.5$$

$$U^* = \frac{1}{1+\sqrt{\frac{\Sigma (Pi-Ai)^2}{\Sigma (Pi)^2}}}$$

With

$$R = \frac{\Sigma AiPi - (\Sigma Ai \Sigma Pi)/n}{(\Sigma Ai^2 - ((\Sigma Ai)^2/n))(\Sigma Pi^2 - ((\Sigma Pi)^2/n))}$$

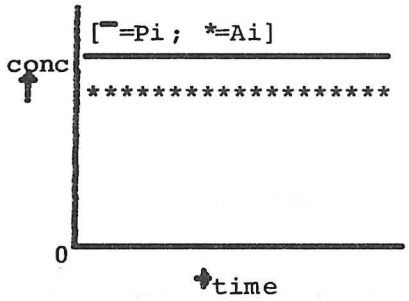
and $Sa^2 = 1/n \Sigma (Ai-\bar{A})^2$

6. Results

6.1 Characteristics

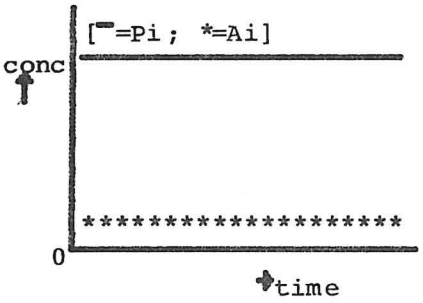
A combination of RC, R* and U*, as defined before can probably be used as a figure for the usefulness of the BOEDE model. Some characteristics can be checked out by the following pictures.

[+=high (good); -=low (bad)]



RC	R*	U*
-	+	+

Fig. 10.



RC	R*	U*
-	+	-

Fig. 11.

So in fig. 8 both calculations of U are being divided by a bigger number than in fig. 9. Since the deviation of Ai in relation to Pi is the same the outcome of U should be the same in both cases.

I would prefer

$$U^2 = \frac{\Sigma (Pi-Ai)^2}{\Sigma (Pi)^2}$$

because this method gives the same value of U for a deviation of Ai below and above Pi.

It is clear that U has no upper limit so define U* as

$$U^* = \frac{1}{1+\sqrt{\frac{\Sigma (Pi-Ai)^2}{\Sigma (Pi)^2}}}$$

U* now varies between 0 and 1 in wich 1 means good and 0 means bad.

From the previous text it seems that there are 3 components that are likely to contribute to the figure that is to be determined.

$$RC = \frac{(1-R^2)Sa^2}{MSE}$$

$$R^* = R/2 + 0.5$$

$$U^* = \frac{1}{1+\sqrt{\frac{\Sigma (Pi-Ai)^2}{\Sigma (Pi)^2}}}$$

With $$R = \frac{\Sigma AiPi - (\Sigma Ai \Sigma Pi)/n}{(\Sigma Ai^2 - ((\Sigma Ai)^2/n))(\Sigma Pi^2 - ((\Sigma Pi)^2/n))}$$

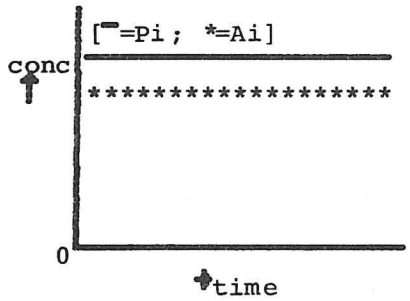
and $Sa^2 = 1/n\Sigma (Ai-\bar{A})^2$

6. Results

6.1 Characteristics

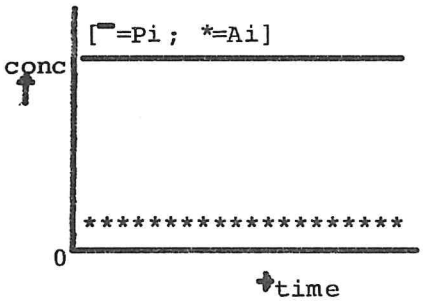
A combination of RC, R* and U*, as defined before can probably be used as a figure for the usefulness of the BOEDE model. Some characteristics can be checked out by the following pictures.

[+=high (good); -=low (bad)]



RC	R*	U*
-	+	+

Fig. 10.



RC	R*	U*
-	+	-

Fig. 11.

So in fig. 8 both calculations of U are being divided by a bigger number than in fig. 9. Since the deviation of Ai in relation to Pi is the same the outcome of U should be the same in both cases.

I would prefer

$$U^2 = \frac{\Sigma (Pi-Ai)^2}{\Sigma (Pi)^2}$$

because this method gives the same value of U for a deviation of Ai below and above Pi.

It is clear that U has no upper limit so define U* as

$$U^* = \frac{1}{1+\sqrt{\frac{\Sigma (Pi-Ai)^2}{\Sigma (Pi)^2}}}$$

U* now varies between 0 and 1 in wich 1 means good and 0 means bad.

From the previous text it seems that there are 3 components that are likely to contribute to the figure that is to be determined.

$$RC = \frac{(1-R^2)Sa^2}{MSE}$$

$$R^* = R/2 + 0.5$$

$$U^* = \frac{1}{1+\sqrt{\frac{\Sigma (Pi-Ai)^2}{\Sigma (Pi)^2}}}$$

With

$$R = \frac{\Sigma AiPi - (\Sigma Ai \Sigma Pi)/n}{(\Sigma Ai^2 - ((\Sigma Ai)^2/n))(\Sigma Pi^2 - ((\Sigma Pi)^2/n))}$$

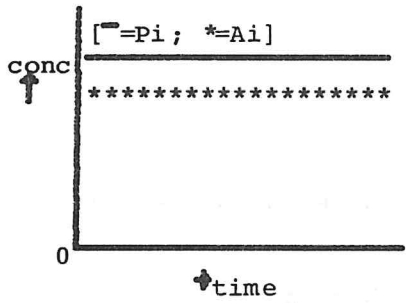
and $Sa^2 = 1/n \Sigma (Ai-\bar{A})^2$

6. Results

6.1 Characteristics

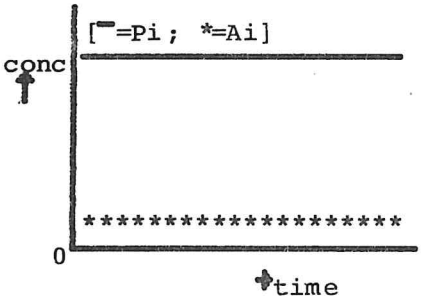
A combination of RC, R* and U*, as defined before can probably be used as a figure for the usefulness of the BOEDE model. Some characteristics can be checked out by the following pictures.

[+=high (good); -=low (bad)]



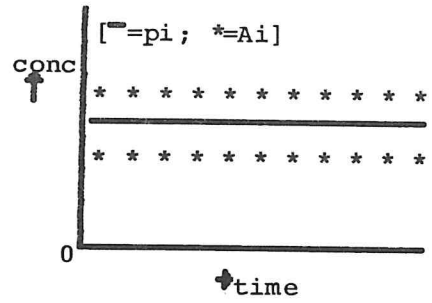
RC	R*	U*
-	+	+

Fig. 10.



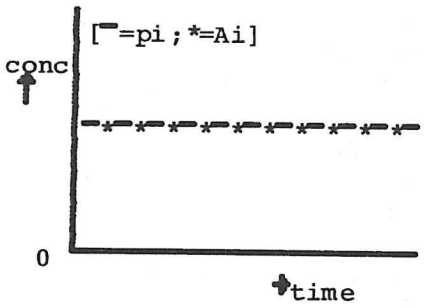
RC	R*	U*
-	+	-

Fig. 11.



RC	R*	U*
+	+/-	+/-

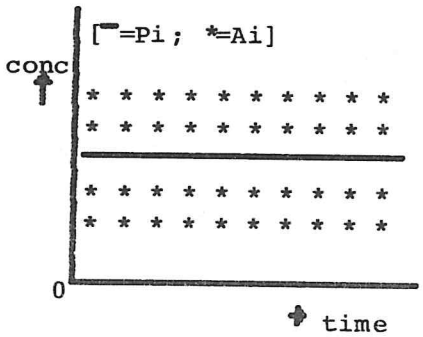
Fig. 12.



RC	R*	U*
not def!	+	+

Fig. 13.

This is an ideal situation. It is the best possible. A figure should be as high as possible. RC can not be calculated because MSE=0.



RC	R*	U*
+	-	+/-

Fig. 14.

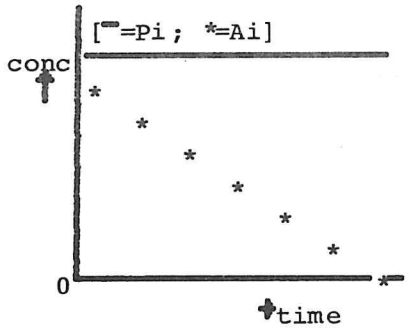


Fig. 15.

The average of RC, R* and U* proved to be a good estimator of the biologists opinion of the usefulness. But I would like to advise BOEDE to use the following definition: let F=10(RC+R*+U*)/3 be the figure for the usefulness. F now varies between 0 and 10 since many people often express results as a figure between 0 and 10. This figure defines the usefulness at an ordinal scale so it can be used to compare different predictors and submodels or the same submodel under different situations.

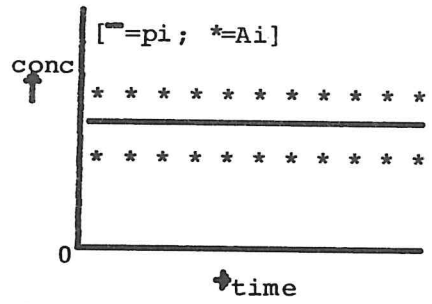
If wanted the BOEDE researchers can weigh those three components for instance by using F=10(a*RC+b*(R*)+c*(U*))/3 with a+b+c=1. But I have seen no reason to do so (so far).

Before using this method I would like to state a few warnings.

1) F can not be used when MSE=0
2) F can be used to develop a better (or worse) predictor for a certain set of observations but in that case it is simpler to minimize the MSE or maximize RC.
3) The calculated F can never be exactly 10 or 0. But F will be made a 10 when MSE=0.
4) F less than 10 does not necessarily mean that the optimum predictor has not been found! F will only be 10 when predictions and observations are exactly the same. Only when RC=0 het optimum predictor has been found.

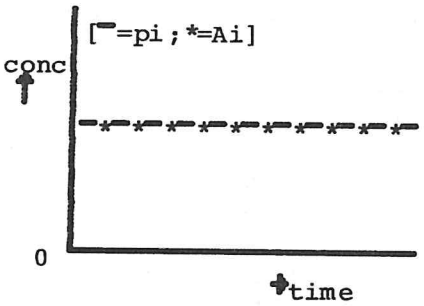
6.2 Application

A computer program has been made to calculate F for various variables of the BOEDE model. Refer to appendix 2 for some examples. The final program should contain:



RC	R*	U*
+	+/-	+/-

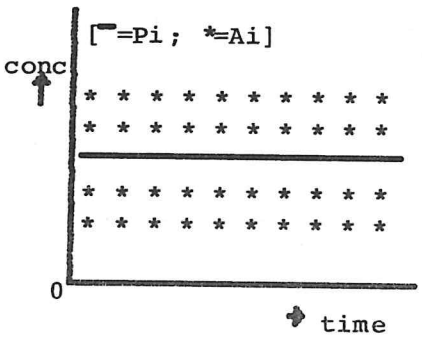
Fig. 12.



RC	R*	U*
not def!	+	+

Fig. 13.

This is an ideal situation. It is the best possible. A figure should be as high as possible. RC can not be calculated because MSE=0.



RC	R*	U*
+	-	+/-

Fig. 14.

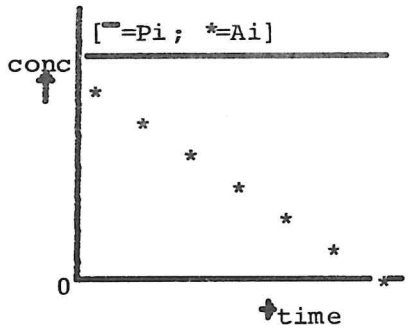


Fig. 15.

The average of RC, R* and U* proved to be a good estimator of the biologists opinion of the usefulness. But I would like to advise BOEDE to use the following definition: let F=10(RC+R*+U*)/3 be the figure for the usefulness. F now varies between 0 and 10 since many people often express results as a figure between 0 and 10. This figure defines the usefulness at an ordinal scale so it can be used to compare different predictors and submodels or the same submodel under different situations.

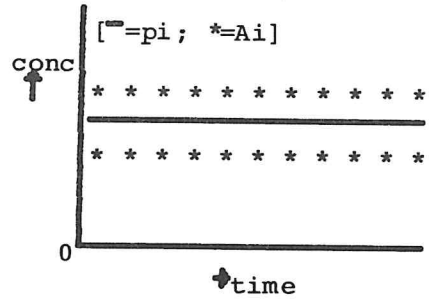
If wanted the BOEDE researchers can weigh those three components for instance by using F=10(a*RC+b*(R*)+c*(U*))/3 with a+b+c=1. But I have seen no reason to do so (so far).

Before using this method I would like to state a few warnings.

1) F can not be used when MSE=0
2) F can be used to develop a better (or worse) predictor for a certain set of observations but in that case it is simpler to minimize the MSE or maximize RC.
3) The calculated F can never be exactly 10 or 0. But F will be made a 10 when MSE=0.
4) F less than 10 does not necessarily mean that the optimum predictor has not been found! F will only be 10 when predictions and observations are exactly the same. Only when RC=0 het optimum predictor has been found.

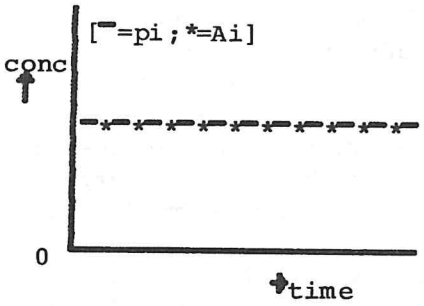
6.2 Application

A computer program has been made to calculate F for various variables of the BOEDE model. Refer to appendix 2 for some examples. The final program should contain:



RC	R*	U*
+	+/-	+/-

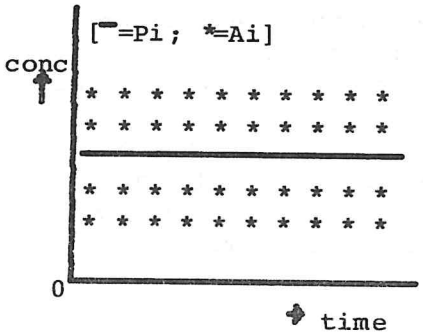
Fig. 12.



RC	R*	U*
not def!	+	+

Fig. 13.

This is an ideal situation. It is the best possible. A figure should be as high as possible. RC can not be calculated because MSE=0.



RC	R*	U*
+	-	+/-

Fig. 14.

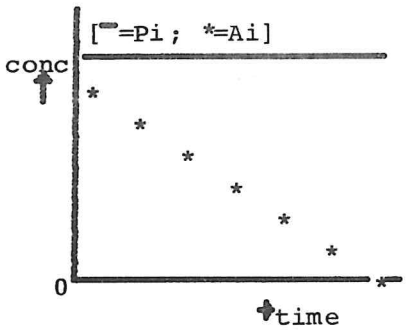


Fig. 15.

The average of RC, R* and U* proved to be a good estimator of the biologists opinion of the usefulness. But I would like to advise BOEDE to use the following definition: let F=10(RC+R*+U*)/3 be the figure for the usefulness. F now varies between 0 and 10 since many people often express results as a figure between 0 and 10. This figure defines the usefulness at an ordinal scale so it can be used to compare different predictors and submodels or the same submodel under different situations.

If wanted the BOEDE researchers can weigh those three components for instance by using F=10(a*RC+b*(R*)+c*(U*))/3 with a+b+c=1. But I have seen no reason to do so (so far).

Before using this method I would like to state a few warnings.

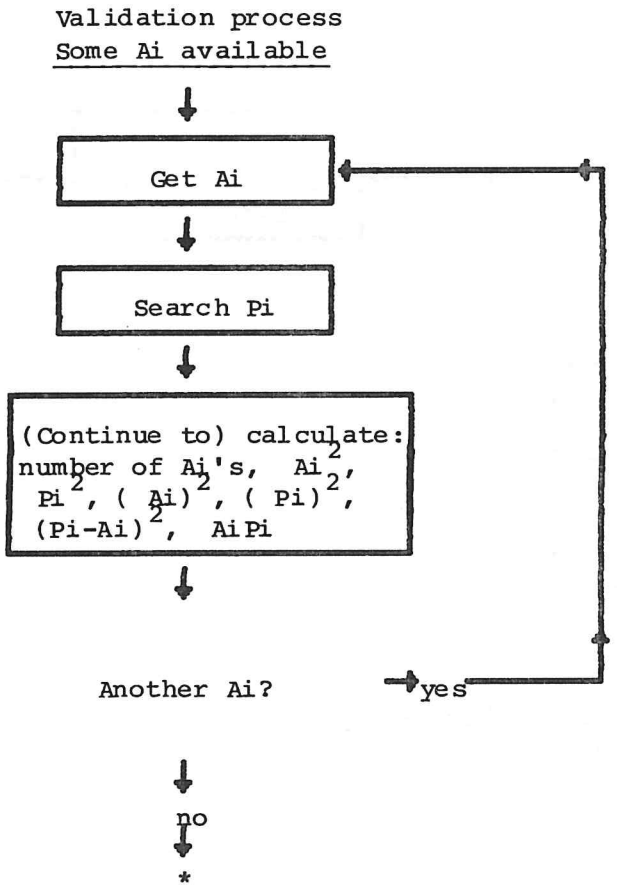
1) F can not be used when MSE=0
2) F can be used to develop a better (or worse) predictor for a certain set of observations but in that case it is simpler to minimize the MSE or maximize RC.
3) The calculated F can never be exactly 10 or 0. But F will be made a 10 when MSE=0.
4) F less than 10 does not necessarily mean that the optimum predictor has not been found! F will only be 10 when predictions and observations are exactly the same. Only when RC=0 het optimum predictor has been found.

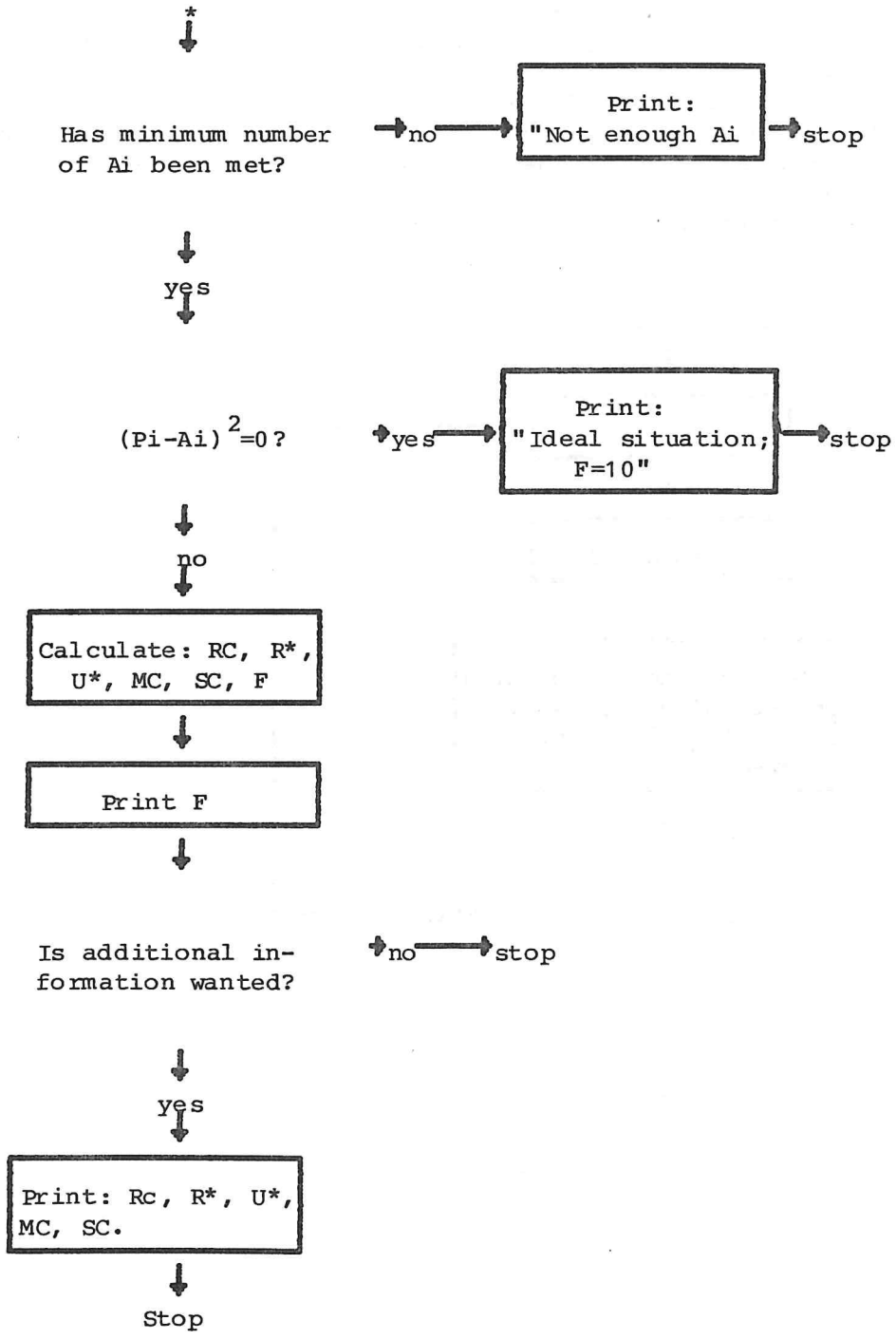
6.2 Application

A computer program has been made to calculate F for various variables of the BOEDE model. Refer to appendix 2 for some examples. The final program should contain:

1) A control for the minimum number of observations required. This minimum might be somewhere between 5 and 20, subject to the distribution of the field observations. If this minimum is not met a warning should be given.
2) A calculation of MSE.
   If MSE=0 the figure F can not be calculated though it should be 10!
3) The calculation of the figure.
4) The mean component (MC) and the slope component (SC) can be given as additional information about the systematic error.

The program flowchart of the final program could be:




*
Has minimum number of Ai been met?
no
Print: "Not enough Ai
stop
yes
(Pi-Ai)2=0?
yes
Print: "Ideal situation; F=10"
stop
no
Calculate: RC, R*, U*, MC, SC, F
Print F
Is additional information wanted?
no
stop
yes
Print: Rc, R*, U*, MC, SC.
Stop

7. Discussion

By calculating RC, U* and R* we can judge the measure of validity objectively. Though taking the average of these 3 components in order to calculate a figure is a very subjective decision. Thus in a way all 3 components are being weight equally.

Calculating RC can cause problems when alle Pi and Ai have about the same value. In the method used by Theil to calculate the RC it can be a problem to find the best fit of the regression line, as is shown in fig. 16.

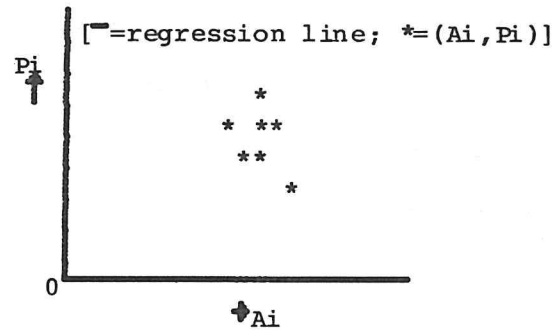


Fig. 16.

8. Appendix 1

Theil's formulas.

Prove that $1/n(\Sigma(Pi-Ai)^2)=(\bar{P}-\bar{A})^2 + (Sp-R*Sa)^2 + (1-R^2)Sa^2$

$1/n(\Sigma(Pi-Ai)^2=1/n(\Sigma((Pi-\bar{P})-(Ai-\bar{A}))^2) + 1/n(\Sigma(\bar{P}-\bar{A})^2$

$= (1)\ 1/n(\Sigma(\bar{P}-\bar{A})^2) +$

$(2)\ 1/n(\Sigma(Pi-\bar{P})^2)-2/n(\Sigma((Pi-\bar{P})(Ai-\bar{A}))+\dfrac{1/n(\Sigma((Pi-\bar{P})(Ai-\bar{A}))^2}{\Sigma(Pi-\bar{P})^2}$

$(3)\ 1/n(\Sigma(Ai-\bar{A})^2) - \dfrac{1/n(\Sigma((Pi-\bar{P})(Ai-\bar{A}))^2}{\Sigma(Pi-\bar{P})^2}$

$= (1)\ (\bar{P}-\bar{A})^2 +$

$(2)\ Sp^2 - 2R*Sa*Sp + R^2Sa^2$

$(3)\ Sa^2 - R^2Sa^2$

$= (\bar{P}-\bar{A})^2 + (Sp-R*Sa)^2 + (1-R^2)Sa^2$

With $R = \dfrac{1/n(\Sigma(Pi-\bar{P})(Ai-\bar{A})}{Sp*Sa}$

and $Sa^2 = 1/n(\Sigma(Ai-\bar{A})^2)$ and $Sp^2 = 1/n(\Sigma(Pi-\bar{P})^2)$

9. Appendix 2

Variables used (and predicted) in the BOEDE model are:

BAL = Thickness of the aerobic layer
BMEI = Meiobenthos
BPYR = Benthic pyrite (FeS)
BSUL = Benthic sulfide
CDIA = Concentration of the benthic diatoms in the most upper 1 cm of the tidal flat
DOC = Dissolved Organic Carbon
EMAC = Epibenthic macrobenthos
MBOX = Measured benthic oxygen production
MBPP = Measured benthic primairy production(=phytobenthos production)
MPPP = Measured pelagic primairy production(=phytoplankton production)
OX = Oxygen in the water
PBAC = Pelagic bacteria
PCOP = Pelagic copepods (zooplankton)
PHYT = Phytoplankton
PO4 = (Pelagic) phosphate
SILIC = (Pelagic) silicate
SLOCA = Oxygen demand of the anaerobic layer
SLOCB = Oxygen demand of the aerobic layer
SUSM = Suspended matter
TOC = Total Organic Carbon

The Ems-Dollard estuary has been divided into five compartments as shown in the figure (Fig. 17).

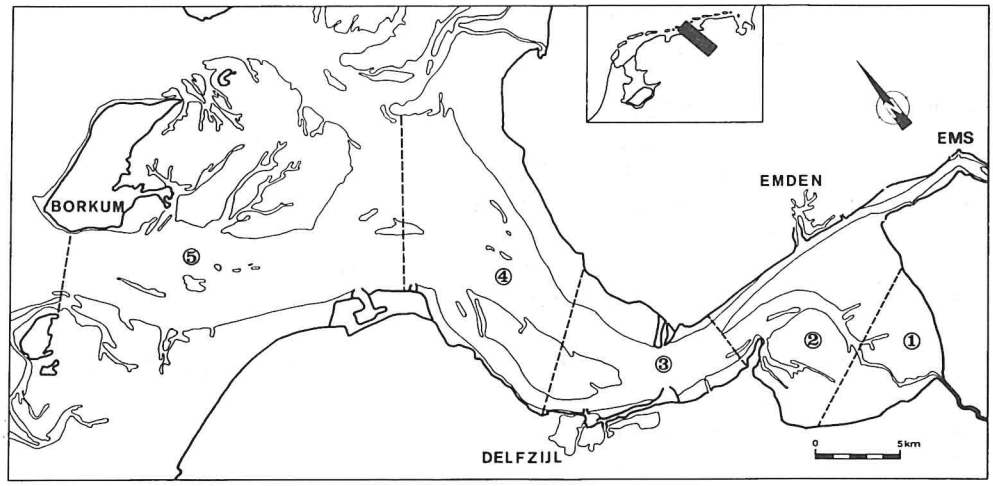


Fig. 17.

Now some examples will be given of predicted variables, field observations and the figure for the usefulness. The figures between brackets refer to the compartments in the estuary. Therefore result of a model simulation are used of a model version of May 1984. Results of these run may be differ of results which are obtained from later versions.

Appendix 2, example 1

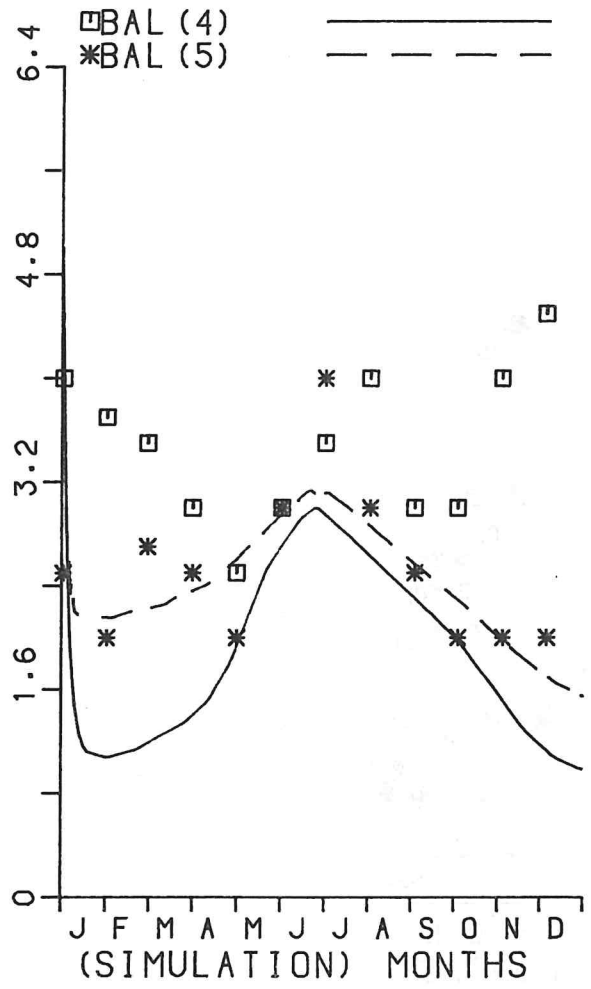


Fig. 18.

Var	F	RC	R*	U*	MC	SC
BAL(4)	3.9	0.10	0.51	0.56	0.63	0.27
BAL(5)	7.3	0.65	0.74	0.80	0.01	0.34

Appendix 2, example 2

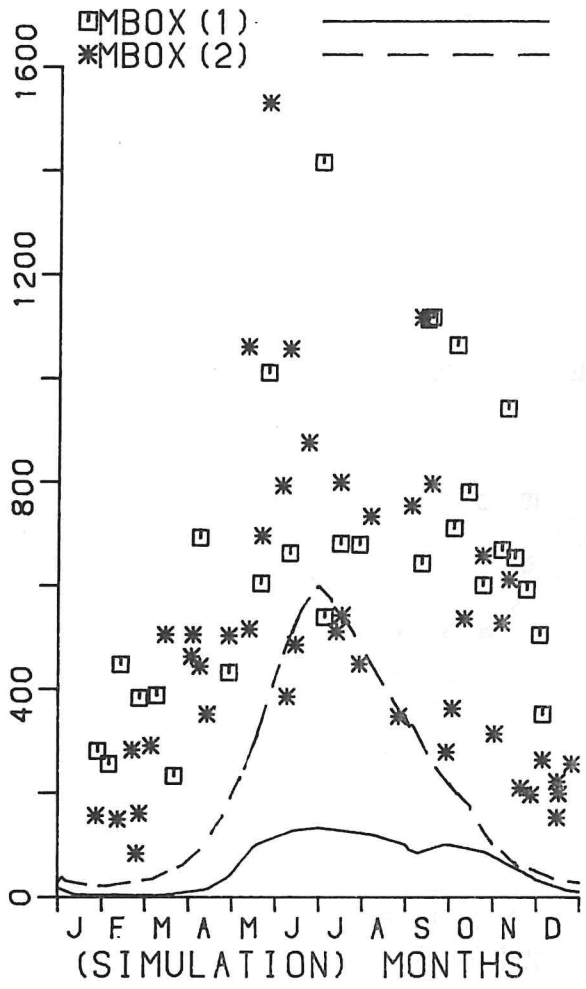


Fig. 19.

Var	F	RC	R*	U*	MC	SC
MBOX(1)	3.5	0.11	0.83	0.11	0.84	0.05
MBOX(2)	5.6	0.43	0.80	0.44	0.57	0.00

Appendix 2, example 3

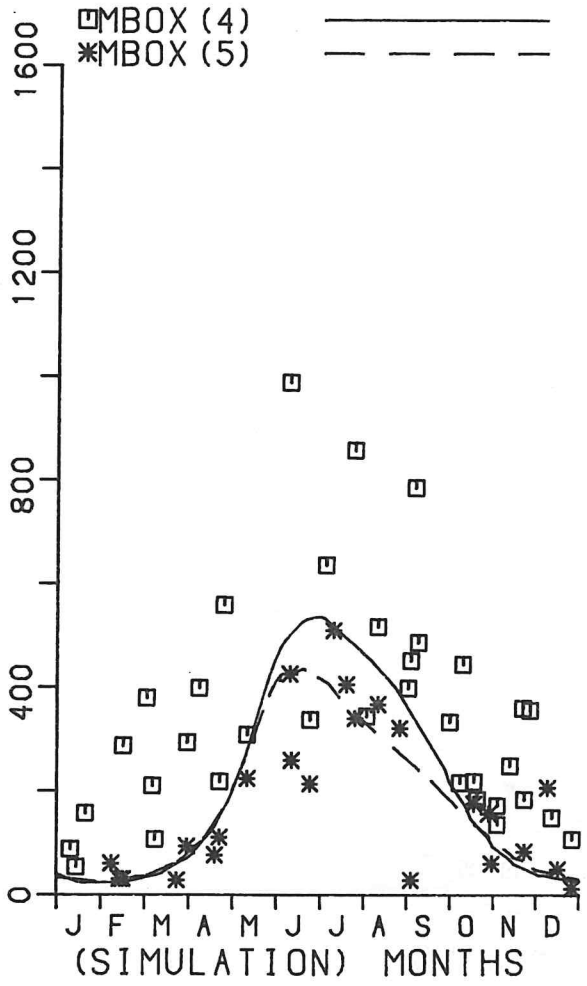


Fig. 20.

Var	F	RC	R*	U*	MC	SC
MBOX(4)	6.3	0.48	0.87	0.54	0.52	0.00
MBOX(5)	8.4	0.87	0.91	0.73	0.02	0.11

Appendix 2, example 4

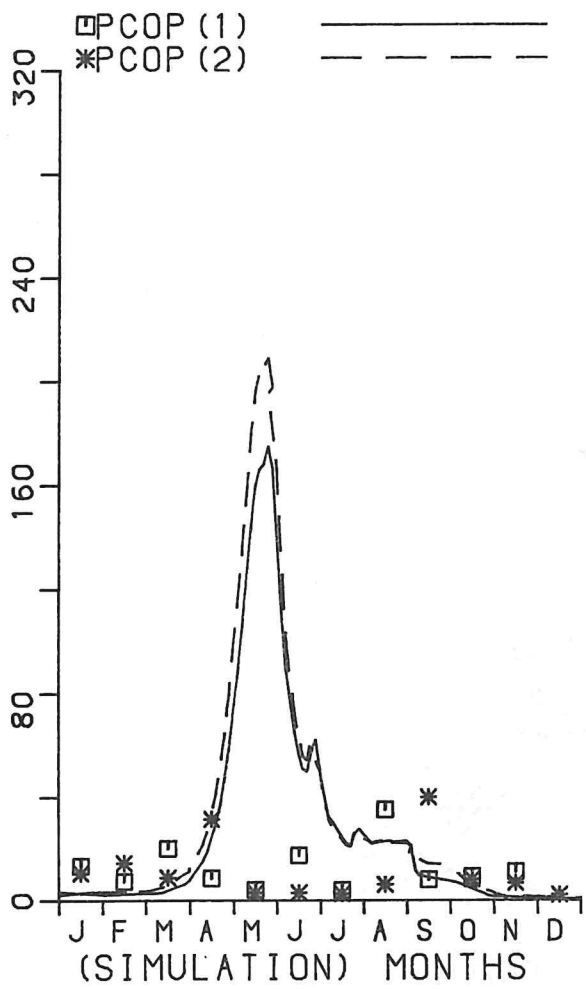


Fig. 21.

Var	F	RC	R*	U*	MC	SC
PCOP(1)	3.1	0.00	0.39	0.51	0.10	0.90
PCOP(2)	3.1	0.00	0.39	0.51	0.10	0.90

Appendix 2, example 5

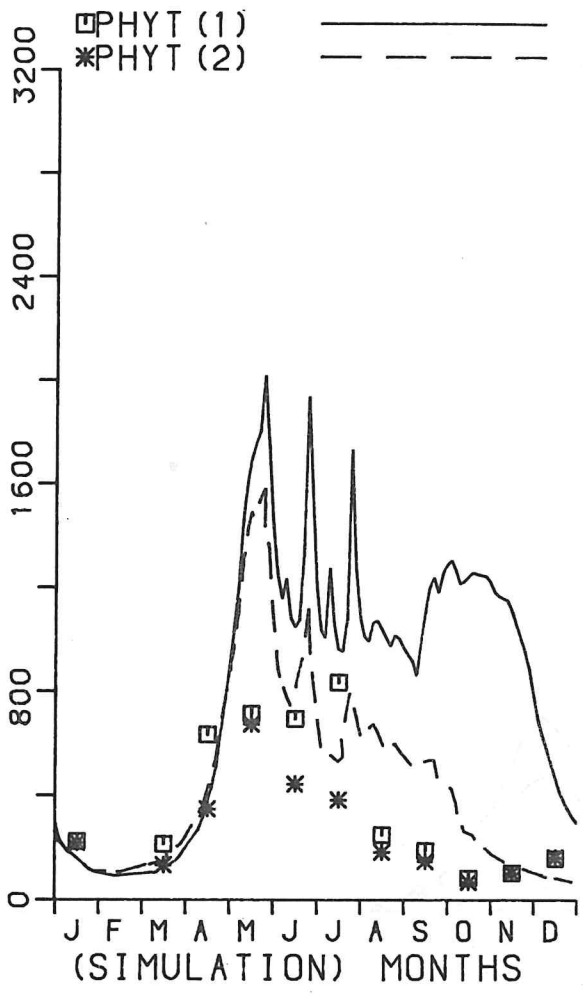


Fig. 22.

Var	F	RC	R*	U*	MC	SC
PHYT(1)	4.5	0.15	0.62	0.59	0.48	0.37
PHYT(2)	5.5	0.07	0.93	0.65	0.38	0.55

Appendix 2, example 6

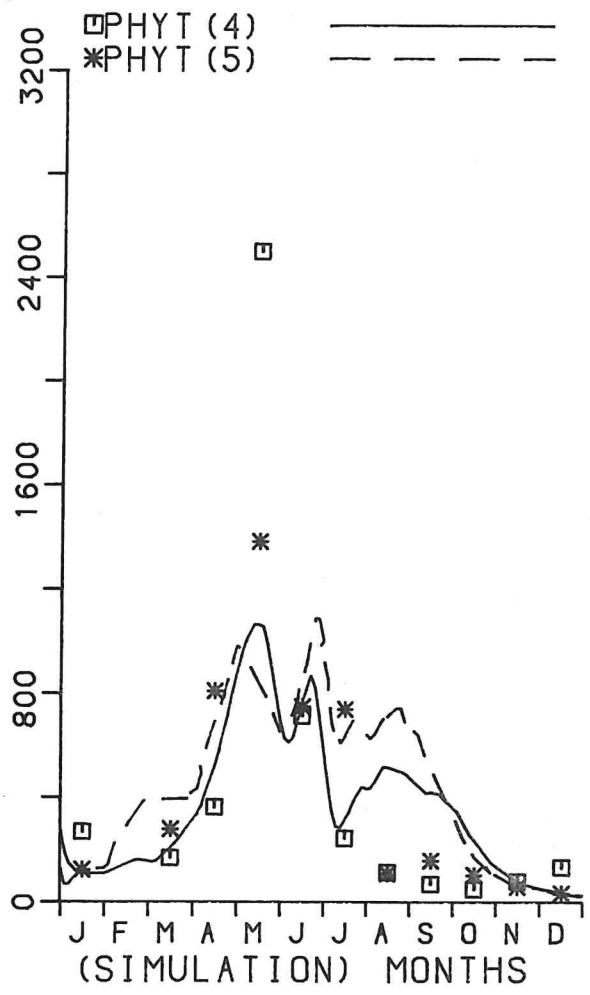


Fig. 23.

Var	F	RC	R*	U*	MC	SC
PHYT(4)	7.1	0.70	0.91	0.51	0.01	0.29
PHYT(5)	8.4	0.98	0.88	0.66	0.02	0.00

Appendix 2, example 7

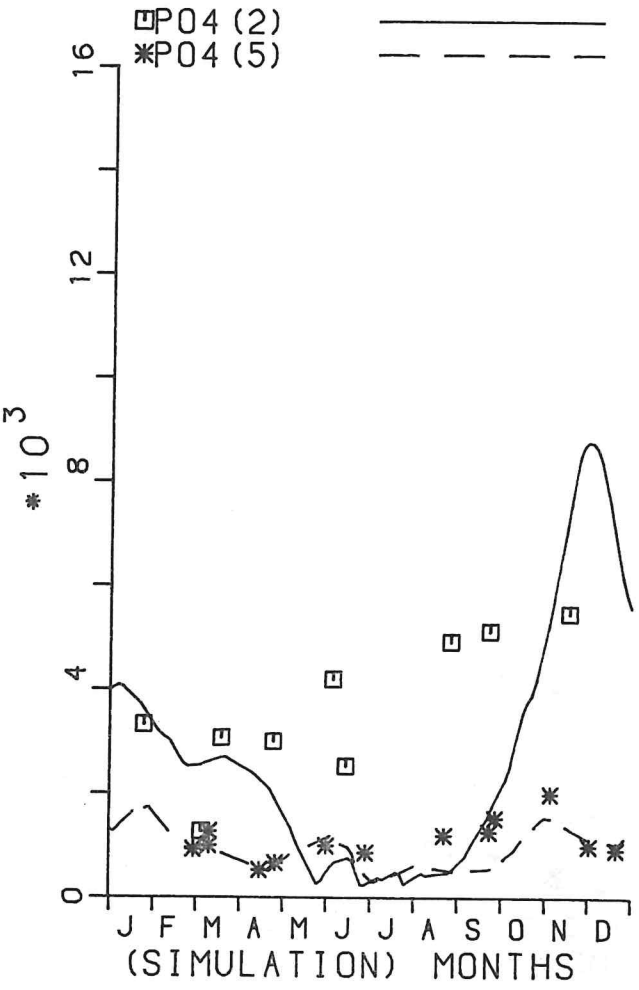


Fig. 24.

Var	F	RC	R*	U*	MC	SC
PO4(2)	4.8	0.27	0.62	0.56	0.25	0.48
PO4(5)	6.7	0.56	0.75	0.69	0.31	0.13

Appendix 2, example 8

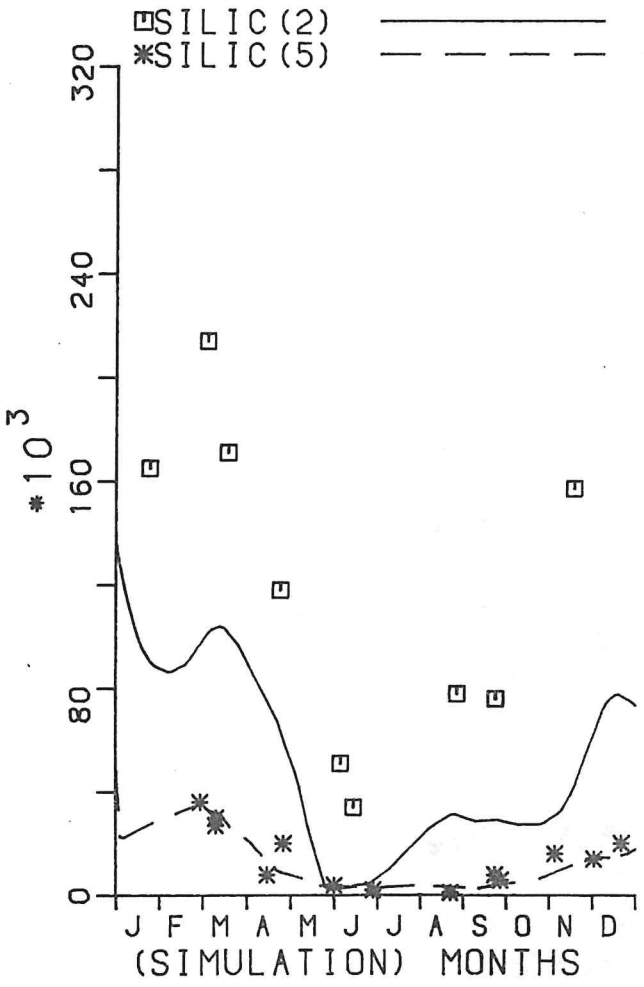


Fig. 25.

Var	F	RC	R*	U*	MC	SC
SILIC(2)	5.1	0.10	0.96	0.47	0.85	0.05
SILIC(5)	8.6	0.84	0.95	0.78	0.06	0.10

10. Literature

BOEDE-GROEP, 1983. Biologisch Onderzoek Eems-Dollard Estuarium. BOEDE publicaties en verslagen, 1983-1.

BOSCH, A.J., R. DOORNBOS, H.N. LINSSEN & J.Th.M. WIJNEN, 1982. Syllabus toegepaste statistiek. Dictaatnummer 2230, Technische Hogeschool Eindhoven.

CHURCHMAN, C.W., 1968. Challenge to reason. Mc Graw-Hill book company.

FISHMAN, G.S. & P.J. Kiviat, 1967. The analysis of simulation-generated time series. Management science, 13(7), 525-557.

HERMANN, C.F., 1967. Validation problems in games and simulations with special reference to models of international politics. Behavioral science, 12, 216-231.

HORN, R.L. van, 1971. Validation of simulation results. Management science, 17(5), 247-258.

LAW, A.M. & W.D. KELTON, 1982. Validation of simulation models. In: Simulation modeling and analysis. Mc Graw-Hill Book Company, 333-348.

MANKIN, J.B., R.V. O'NEILL, B.W. RUST & H.H. SHUGART, 1975. The importance of validation in ecosystem analysis. From: New directions in the analysis of ecological systems, part I, G.S. Innis, ed., Simulation councils proc. ser., 5(1), 63-71.

NAYLOR, T.H. & J.M. FINGER, 1967. Verification of computer simulation models. Management science, 14(2), 92-106

PLATT, T., K.H. MANN & R.E. ULANOWICZ, 1981. Ch. 1.4.3: Validation. In: Mathematical models in biological oceanography. The unesco press, 46-47.

SHANNON, R.E., 1975. Ch. 6: Validation and analysis. In: Systems simulation The art and science. Prentice-Hall, inc., 208-242.

THEIL, H., 1961. Economic forecasts and policy. North Holland publishing company.

THEIL, H., 1966. Applied economic forecasting. North Holland publishing company.

TURING, A.M., 1950. Computing machinery and intelligence. Mind 59, 433-460.