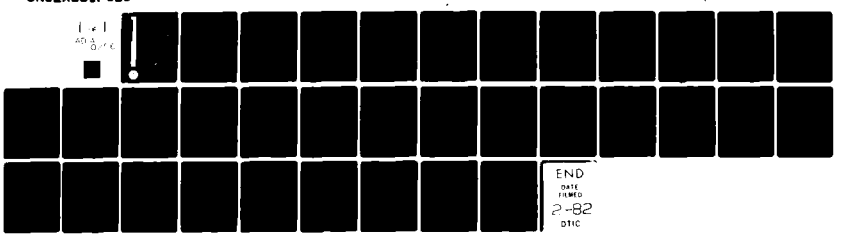
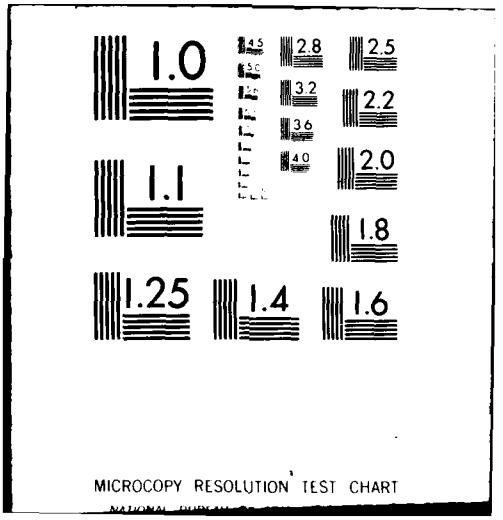
AD-A110 256 STATE UNIV OF NEW YORK AT BUFFALO AMHERST DEPT OF CO--ETC F/G 9/2
A MODULE TO ESTIMATE NUMERICAL VALUES OF HIDDEN VARIABLES FOR E--ETC(U)
NOV 81 N V FINDLER, J E BROWN, R LO, H Y YOU AFOSR-81-0220

UNCLASSIFIED AFOSR-TR-81-0873 NL



END
DATE FILMED
2-82
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
MICROCOPY RESOLUTION TEST CHART

AFOSR-TR- 81-0873

LEVEL II

(3)

A MODULE TO ESTIMATE NUMERICAL VALUES OF HIDDEN VARIABLES FOR EXPERT SYSTEMS*

BY

NICHOLAS V. FINDLER, JOHN E. BROWN,
RON LO, AND HAN YONG YOU

NOVEMBER 1981

DEPARTMENTAL TECHNICAL REPORT NUMBER: # 190

GROUP FOR COMPUTER STUDIES OF STRATEGIES
TECHNICAL REPORT NUMBER: # 4

*THE WORK DESCRIBED HAS BEEN SUPPORTED BY THE AIR FORCE OFFICE OF SCIENTIFIC RESEARCH GRANT [redacted] AFOSR-81-0220

STATE UNIVERSITY OF NEW YORK AT BUFFALO

Department of Computer Science

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER **AFOSR-TR- 81-0873**	2. GOVT ACCESSION NO. A110256 / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A MODULE TO ESTIMATE NUMERICAL VALUES OF HIDDEN VARIABLES FOR EXPERT SYSTEMS	5. TYPE OF REPORT & PERIOD COVERED INTERIM, 1 JUL 80 - 30 JUN 81
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Nicholas V. Findler, John E. Brown, Ron Lo, and Han Yong You	8. CONTRACT OR GRANT NUMBER(s) AFOSR-81-0220
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Computer Science State University of New York at Buffalo 4226 Ridge Lea Road, Amherst NY 14226	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F; 2304/A2
11. CONTROLLING OFFICE NAME AND ADDRESS Directorate of Mathematical & Information Sciences Air Force Office of Scientific Research Bolling AFB DC 20332	12. REPORT DATE NOV 81 / 13. NUMBER OF PAGES 28
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) UNCLASSIFIED / 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Decision-support systems, estimation techniques, inductive inference-making, learning programs, production systems, man-machine systems, distributed databases.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
In the area of strategic decision-making, the objective often is to achieve one's own goals and to prevent the achievement of the adversaries' goal. To do so, the decision-maker needs to know, as precisely as possible, the values of the relevant variables at various times. Some of these variables, the open variables, are readily measurable at any time. Others, the hidden variables, can be measured only at certain times, either intermittently or periodically. (CONTINUED)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 422759

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ITEM #20, CONTINUED:

The authors have implemented a module that can act as a decision-support tool for a variety of expert systems in need of estimates of hidden variables values at any desired time. The estimation is based on generalized production rules expressing stochastic, causal relations between open and hidden variables. The quality of the estimates improves through a multi-level learning process as both the number and the quality of the rules increase. The modularity of these causal relations make incremental expansion and conflict resolution natural and easy. Restricting the set and the domain of pattern formation rules to a reasonable size makes the system effective and efficient. Finally, the system can be easily employed for distributed database applications.

A MODULE TO ESTIMATE NUMERICAL VALUES OF HIDDEN VARIABLES FOR EXPERT SYSTEMS

Nicholas V. Findler, John E. Brown,

Ron Lo, and Han Yong You

Group for Computer Studies of Strategies

Department of Computer Science

State University of New York at Buffalo

ABSTRACT

In the area of strategic decision-making, the objective often is to achieve one's own goals and to prevent the achievement of the adversaries' goal. To do so, the decision-maker needs to know, as precisely as possible, the values of the relevant variables at various times. Some of these variables, the open variables, are readily measurable at any time. Others, the hidden variables, can be measured only at certain times, either intermittently or periodically.

We have implemented a module that can act as a decision-support tool for a variety of expert systems in need of estimates of hidden variable values at any desired time. The estimation is based on generalized production

rules expressing stochastic, causal relations between open and hidden variables. The quality of the estimates improves through a multi-level learning process as both the number and the quality of the rules increase. The modularity of these causal relations make incremental expansion and conflict resolution natural and easy. Restricting the set and the domain of pattern formation rules to a reasonable size makes the system effective and efficient. Finally, the system can be easily employed for distributed database applications.

KEYWORDS: decision-support systems, estimation techniques, inductive inference-making, learning programs, production systems, man-machine systems, distributed databases.

1. INTRODUCTION

Most expert systems are essentially aids to human decision-making concerning a sequence of actions to be taken. We have been particularly interested in strategic decision-making where the consequences of each action affect the final outcome of some confrontation among adversaries throughout the whole history of the confrontation. The actions aim at optimizing an implicitly or explicitly given objective function in a certain environment. The simultaneous achievement of one's own goals and the frustration of the adversaries' goals is a special, but frequent and important, case in strategic confrontations. (Even more special is the case in which Nature is the (non-competitive) opponent and the goal is to identify, say, the source of some malfunctioning or the location of a certain resource.)

Precise knowledge of the values of the relevant variables in the environment is necessary for the decision-making task. Some of these variables, called by us open variables (OV's), are observable and measurable at any time whereas the values of the hidden variables (HV's) can be identified only at certain times, either intermittently or periodically. Some examples of these two types of variables are as follows:

(i) Atmospheric conditions: Open variables are measured continuously at the Earth's surface while high-altitude variables, such as stratospheric wind velocity and air temperature, can be observed only when, for example, balloon-born instruments are sent up.

(ii) Material testing: Inexpensive and non-destructive testing of products can be performed at any desired time but costly or destructive tests are carried out sparingly.

(iii) Oil and mineral exploration: Evaluation of satellite photographs, seismic experiments, geological surface studies may be done with arbitrary frequency but deep-drill work is necessarily restricted.

(iv) Earthquake prediction: In addition to seismographical data, a number of OV types have been used and suggested as precursors of earthquakes. The HV's would be the location of the epicenter and the intensity of the earthquake.

(v) Specialist combat training: The length and intensity of training, the frequency of exercises are controlled, open variables while, for example, performance under actual battle conditions can be evaluated only at irregular times.

The estimation of the HV values is a problem in pattern-directed inductive inference-making. To provide such estimates, we have designed and implemented a system based on the assumption that certain OV's and HV's are stochastically and causally related, and both OV's and HV's

can be causes or effects. The system analyzes the (evaluated) behavior in the vicinity of the time or space at which the estimate is sought.

The basic idea is as follows. The results of measurements of OV's, and occasionally of HV's, are collected and input. A program constructs a unique mathematical description of the behavior of each OV by identifying the patterns prevailing over the domain of the independent variable. The latter is a time- or space-like variable assuming continuous or discrete values or it can be an event-counter. The mathematical description consists of an ordered set of parametrized basic patterns, called by us morphs, that fit the datapoints optimally. Optimality refers to the requirement that a minimum number of morphs are computed for a prespecified level of statistical significance (there is a tolerated level of "unexplained" variance around the morphs). As described in detail later, a morph can be a trend, a step function or a sudden change of only momentary effect. The predictor part of the stochastic causal relation (the "condition" part of the usual production rule) consists of:

(i) The values of one or several parameters of a morph describing the behavior of an OV;

(ii) The difference in space or time between the beginning of the morph and the point at which the value of an HV is known (or sought).

The predicted part of our generalized production rule (traditionally the "action") refers to a certain HV whose value is given or is to be estimated.

Each rule has an associated credibility level, a measure of its quality. The rules are ordered according to decreasing credibility levels for any given OV-HV pair. This arrangement resolves the conflict among several possible estimates as shown in Section 5 dealing with estimation.

We shall next discuss some theoretical issues related to our method, followed by a description of how the knowledge base is established, refined and utilized. The Appendix illustrates all this in more detail.

2. THE THEORETICAL BASIS

As outlined before, a sort of generalized production rule (GPR) expresses a stochastic, causal relationship between the (mathematical properties of the behavior of the) OV and the (value of a) HV. The form of these rules is [2,3,4]

$$W_r \;/\; M_{ijk} \;/\; T_{jm} \dashrightarrow V_m(H_n) \;:\; Q_r \qquad (1)$$

Here, W_r is the number of rules that have been pooled to form the r-th rule at hand; M_{ijk} is the i-th combination of the parameters of the j-th morph describing the behavior of the k-th OV. Some further explanation is necessary before

giving the meaning of the other symbols.

Our morph-fitting program (MFP) fits a minimum number of basic patterns, morphs, to a sequence of datapoints while maintaining a prespecified level of statistical significance, that is keeping the amount of "unexplained" variance of the datapoints around the morphs below a certain value [1]. A morph is one of the following three basic patterns (see Figure 1):

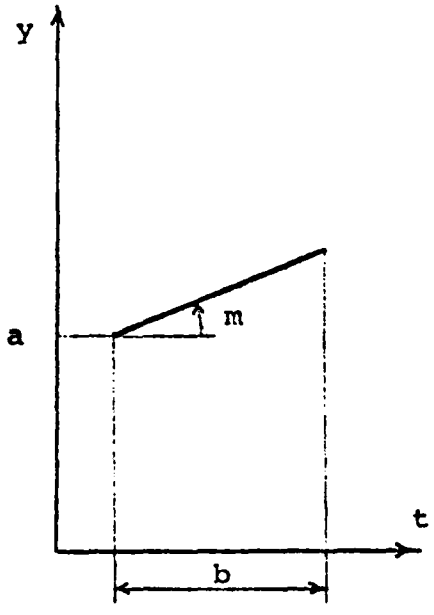
FIGURE 1 ABOUT HERE

.a trend is a monotonic change, a straight line, with three parameters: length, slope and base (starting) value;

.a step function connects the end point of a trend with the starting point of another if there is a discontinuity between two adjacent trends, and has two parameters: base value and change;

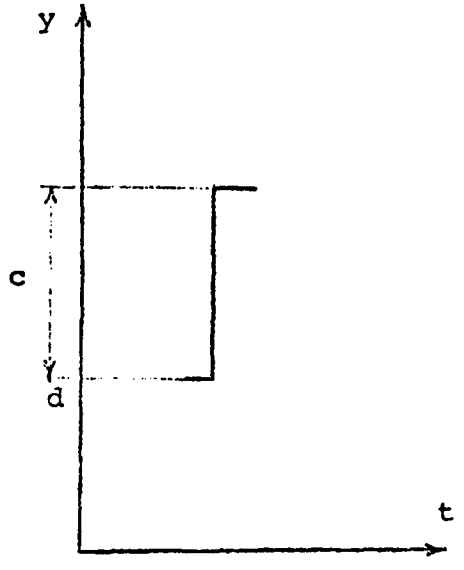
.a sudden change is a momentary jump superimposed onto a trend, with two paramters: base value and peak.

In addition to these three morph types, our MFP identifies a fourth basic pattern. It is called a delay function, essentially a period over which the datapoints are too "scattered" to be described mathematically. Its only parameter is its length. Since the delay function represents a lack of information about the OV behavior, it is not used in the formulation of the GPR's.



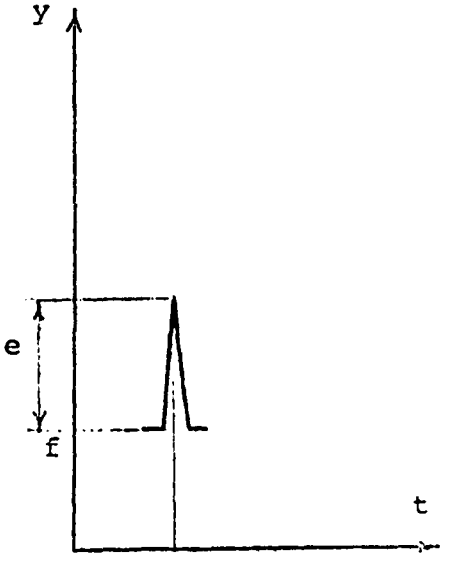
a: base value,
b: length,
m: slope;

trend



d: base value,
c: change;

step function



f: base value,
e: peak.

sudden change

FIGURE 1

A combination of morph parameters in (1) means that on the "left-hand side" of the rule, there can be one or two or, in the case of a trend, three parametric values of the morph.

T_{jm} is the difference in time (time lag) or in space (distance) between the start of the j-th morph (in case of a trend) or its occurrence (in case of a sudden change or step function), and the point of time or space at which the n-th hidden variable, H_n, assumes its m-th value, V_m. This difference may be positive--when the OV is the cause and thus precedes the HV, the effect--or negative in the opposite case. We shall use the common term 'lag' for T_{jm} whether it refers to a timelag or distance.

Q_r is the credibility level of the r-th rule. Its value is between 0 and 1, and depends on two factors:

.how well the morph in question fits the datapoints over its domain, and

.how many and how similar the rules were that have been pooled to form the rule at hand.

These issues will be discussed in a quantitative manner in the next two sections, dealing with establishing and merging rules.

Finally, it should be noted that initially, when the individual rules are first set up, a particular morph parameter combination may be assumed to be causally related to several values of several HV's and, vice versa, a

particular HV value may tentatively be associated with several morphs of several OV's.

3. ESTABLISHING THE KNOWLEDGE BASE

The following scheme of definitions should be helpful in the explanation to come:

knowledge base ::= <ordered set of pooled rules>;

pooled rule ::= <number of rules pooled>, <weighted average of rule parameters of similar rules>, <modified credibility level>;

similar rules ::= <rules with rule parameters in predefined proximity>;

rule parameters ::= <morph parameters>, <lag>, <value and type of HV>;

morph parameters ::= <type of OV>, <trend parameters | step function parameters | sudden change parameters>;

trend parameters ::= <{base value, slope, length}>;

step function parameters ::= <{base value, change}>;

sudden change parameters ::= <{base value, peak}>;

basic pattern ::= <morph | delay function | ...>.

Here, {...} means a non-empty subset of the set consisting of the elements in the braces.

As time proceeds, sets of datapoints, each specifying the value of a particular OV and the corresponding independent, time-like or space-like value, are entered in the system. At any desired time, the user may invoke the MFP, which converts the above "raw" data into basic pattern

descriptions. Similarly, values of HV's become available at times and are put on separate files. The user can then direct the system to set up all applicable rules, tentatively causal relations. To reduce the probability of a combinatorially explosive situation, associating every morph parameter combination of every OV with every value of every HV, he can specify

.which OV's and HV's are likely to be causally related;

.which is the cause and which is the effect in a given OV-HV combination (the sign of the lag);

.the minimum and maximum meaningful values of the lag between a given OV and HV (limits of relevance and physical possibility).

The system will then establish the initial knowledge base consisting of all possible rules satisfying the above restrictions. A rule is a list with the following sequence of elements on it:

(i) The number of rules pooled to form the rule at hand (initially one).

(ii) The type of the OV.

(iii) One of the possible combinations of the morph parameters. (Note there are seven rules set up with reference to a trend of three parameters, and three rules set up with reference to both a step function and a sudden change because they have two parameters.)

(iv) The number of datapoints within the domain of this morph.

(v) The values of the numerator and denominator of the F-ratio measuring the goodness of fit by this morph.

(vi) The lag between the start (or occurrence) of the morph and the point at which the value of the HV was measured. (Two rules are set up, differing only in the sign of the lag, unless the user tells the system whether the OV is the cause or the effect.)

(vii) The type of the HV.

(viii) The value of the HV.

(ix) The credibility level of the rule.

The last item needs explanation. When a new rule is established, it may be a sample of an indefinitely large population of rules whose rule parameters are normally distributed around the value of the causal relation. Or else, it could represent OV and HV events that happen to occur simultaneously. One would, therefore, tend to set its initial credibility level at 0.5 . However, the initial credibility level should also reflect how well the morph in question fits the OV datapoints. A higher level of scattering is equivalent to greater uncertainty about the values of the morph parameters. The goodness of fit, as computed by the MFP, is expressed by the F-statistics [5,6].

In our case,

$$F_{1,N-2} = \frac{s^2_{yreg-\bar{y}}}{s^2_{yreg-y}}(N-2) \qquad (2)$$

Apart from the factor (N-2), it is the ratio of the variance of the regression line (trend) points around the mean to the variance of the datapoints around the regression line. The number of degrees of freedom is always one for the numerator and (N-2) for the demoninator, where N is the number of datapoints.

A large F value indicates that the variation of the data over the independent variable is well-explained by the fitted trend. A small value of F yields uncertainty about the quality of the fit. In accordance with the other related measures discussed in the next session, we define a probabilistic measure of the credibility level of a newly established rule

$$Q_0 = 0.5 \cdot C_1 \qquad (3)$$

where

$$C_1 = \int_0^F G(F')\,dF' \qquad (4)$$

Here G(F) has the usual definition of the probability distribution of $F_{1,\nu}$ (for F>0),

$$G(F) = \Gamma\left(\frac{1+\nu}{2}\right) / \left[\Gamma\left(\frac{1}{2}\right) * \Gamma\left(\frac{\nu}{2}\right)\right] * (\nu.F)^{-\frac{1}{2}} . \left(1+\frac{F}{\nu}\right)^{-\frac{1}{2}(1+\nu)} \qquad (5)$$

and 0 for $F \leq 0$.

The gamma-function is defined as

$$\Gamma(z) = \int_0^\infty e^{-t} \cdot t^{z-1} \cdot dt, \qquad Re(z)>0 \qquad (6)$$

The area under the tail of the curve G(F),

$$\int_F^\infty G(F').dF' \qquad (7)$$

gives the probability that the morph fitted to the datapoints is in reality unrelated to them. For example, if a trend fits the datapoints well, F will be large and we reject the null hypothesis of no relation between the trend and the datapoints. This is why the complement of the integral (6) appears in the formula for C_1, (4), which is a contributing factor to the credibility level of a rule just established.

4. OPTIMIZING THE KNOWLEDGE BASE

As time proceeds, the system receives further sequences of values of several OV's and occasional values of several HV's. The rules that represent 'real' causal relations will recur but the rule parameters will vary, as discussed above, due to statistical fluctuations (measurement errors, variations in the environmental conditions, etc.). Similar rules should then be combined and the credibility level of the "pooled" rule should be raised because an underlying causal association between the OV and HV in question has been corroborated by the new evidence. The rule parameters

of the pooled rule will be the weighted averages of the corresponding input values. Thus, the criteria for combining a new rule with one already in the knowledge base are:

.the rules considered relate the same type of OV and HV;

.the same type of morph describes the OV behavior;

.the same combinations of morph parameters appear;

.the corresponding rule parameters in the rules considered are within the allowed (user-specified) range from each other.

When these conditions are satisfied, a smaller number of rules, but of higher credibility levels, will constitute the knowledge base. Both the number of rules combined to form the current one, W_r, and the credibility level, Q_r in (1), will be updated. Another probabilistic measure will play a role in the new Q_r value. It measures the similarity between the rules pooled. We have said before that the rule parameters are assumed to be normally distributed around the respective population mean values. In view of the small sample sizes, we shall normalize the parametric values being compared by using Student's t-statistics [5,6]. Thus, the j-th rule parameter, P_j, is converted into a t-value

$$t_j = (P_{j,new} - \bar{P}_j)/s_j \qquad (8)$$

where $P_{j,new}$ is the parameter of the rule being considered for pooling; $\bar{P}_j$ is the average value of the same parameter over the W rules already pooled and the one, therefore, that

appears with the rule in the knowledge base,

$$\overline{P}_j = \frac{1}{W}\sum_{k=1}^{W} P_{j,k} \tag{9}$$

$$s_j^2 = \frac{1}{W-1}\sum_{k=1}^{W}(P_{j,k} - \overline{P}_j)^2 \tag{10}$$

The $\underline{t}$-values of all the rule parameters (morph parameters, lag and HV value) are then averaged to obtain $\overline{\underline{t}}$. The overall measure of rule similarity is found from

$$C_2 = 2*\int_{\overline{t}}^{\infty} G(t').dt' \tag{11}$$

where the factor 2 is due to the symmetry of the concept 'deviation from a value'; G(t) is the Student's $\underline{t}$-distribution [6],

$$G(t) = \Gamma(\frac{\nu+1}{2})/[(\pi\nu)^{\frac{1}{2}}.\Gamma(\frac{\nu}{2})]*(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}}, \quad -\infty<t<\infty \tag{12}$$

with being the number of degrees of freedom--in our case equal to W-1.

One has to update also the measure of goodness of fit, $\underline{C_1}$. We define the $\underline{F}$-value for the new, pooled rule as the ratio of the sum of the numerators to the sum of the denominators in the $\underline{F}$-ratios of the separate rules. Using obvious notation,

$$F_{new} = \frac{\text{numerator1+numerator2}}{\text{denominator1+denominator2}} \tag{13}$$

is substituted into (4).

Finally, the updating of the credibility level has to be discussed. The "percent measure" (in regard to the maximum allowed difference between rule parameter values), on which the decision as to pool or not to pool was based, may have been a poor choice by the user. It would then produce pooled rules that have low C_2 values. On the basis of experimentation, we have set a threshold value of 0.05 for the product of the new C_1 and C_2 values. Below 0.05, the scatter around the morphs and the lack of similarity between the pooled rules would probably result in poor estimates of hidden variable values. Therefore, the credibility level is to be decreased. When $C_1 * C_2 \geq 0.05$, it will be increased, according to the heuristic formulae

$$Q_{new} = \begin{cases} Q_{old} + \dfrac{20/19(C_1 . C_2 - 0.05)}{W+1}(1-Q_{old}) & \text{if } C_1 . C_2 \geq 0.05 \\ Q_{old} - \dfrac{20(0.05 - C_1 . C_2)}{W+1} Q_{old} & \text{if } C_1 . C_2 < 0.05 \end{cases} \qquad (14)$$

The factors 20/19 and 20 insure that the increments satisfy the restriction that the credibility level must lie between 0 and 1. The formula (13) also expresses the fact that the more rules have been pooled so far (the value of W), the less effect a new rule has.

5. ESTIMATING THE VALUE OF A HIDDEN VARIABLE

When an estimate of a HV value is desired at a certain value of the independent variable (time or space), the user has to provide a sequence of OV values in its vicinity. (Remember, for the lag to be meaningful for a given OV-HV type, it has to be less than a user-specified value.) These OV values are then submitted to MFP. Next, the system looks in the knowledge base for highest credibility level that

.connect the HV sought and the available OV's;

.refer to the same morph type;

.involve morph parameter and lag values that are "similar enough" to those in the query, that is within the user-specified range of pooling rules.

In fact, the user may request the N best estimates. Since the rules in the knowledge base are ordered according to decreasing values of credibility levels, it seems natural to return N values (or less if the knowledge base cannot provide enough), from the top rules satisfying the above criteria. However, the overall quality of the estimate, its confidence level, C_e, depends on two additional factors and, therefore, may well be in different order from the one noted above. Namely, also how well the new morph fits its datapoints, and how close its parameters are to those of the morph matched in the knowledge base, contribute to the confidence level of the estimate. These measures are again probabilistic and are formulated identically to the ones

used previously, C_1 in (4) and C_2 in (11), respectively, except that in the present case the closeness of HV is, of course, meaningless. Therefore, the *t*- value, as calculated in (8), for the rule parameter HV is identically zero. We also believe that out of the three factors contributing to the confidence level of the estimate, the most important should be the credibility level of the rule used, Q. This belief is expressed by the formula

$$C_e = Q.\sqrt{C_1 .C_2} \qquad (15)$$

One can now see that the quality measures of the estimates provided by several rules are not necessarily in the same order as the rules providing them. For example, let us compare two adjacent rules in the knowledge base--the one higher up has a higher credibility level--and let both be able to yield an estimate for the same type of HV and at (about) the same lag value. It is quite possible that the parameters of the morph in the query are closer to those of the second rule than to those of the first one. Furthermore, the morph in the query, matching the morph of the second rule, may fit its datapoints better than the corresponding morph, matching the morph of the first rule, fits its datapoints. This is how the confidence level of the estimate provided by the second rule can be higher than that provided by the first rule. Therefore, the top N matching rules may not yield the N best estimates. However,

we know that always $C_e \leq Q$, because $C1$, $C2 \leq 1$. The system is to abandon the search for more estimates when

(i) the number of rules matched is at least N , and

(ii) the credibility level of the rule considered is less than the lowest confidence level of the estimates obtained so far, or

(iii) there are no more rules to consider.

Finally, the estimates are returned by the system only if the average credibility level, weighted by the rule weights W, of all the rules used for estimation is at least 0.75 . This requirement will balance highly credible rules of small weight (formed by pooling few rules) and less credible rules of large weight (rules "blurred" due to many scattered contributions to it).

6. THE PROGRAM

The system is implemented in ALISP, except for the morph-fitting program, MFP, which is written in FORTRAN IV. An extensive use of overlays has forestalled garbage collection problems. The system is highly interactive, which has helped avoid the potentially disastrous effects of combinatorial explosion. As discussed before, the user's judgement is asked for in regard to

.which OV's and HV's may be causally related,

.which of the above is the cause and which is the effect,

.what the maximum and minimum values of the lag between them are,

.what the limits of similarity are for rules to be pooled.

In addition, the system allows the user either to provide values for OV and HV datapoint coordinates or else to specify the names of the datafiles in which these have been stored before. If the morph-fitting program has not been invoked yet, the user can freely add data to datafiles at anytime. The user has several important options. The first two are relevant to distributed processing and intelligence. Let us consider a star-like computer network configuration. In its center, there is a usually larger machine able to communicate with a number of usually smaller, satellite computers. Each of these smaller computers receives OV and HV data from its vicinity, and processes them to form a regional knowledge base. The user has the option of merging either source files, containing morph descriptions and HV datapoints, or regional knowledge bases. The conditions for merging source files are:

(i) the morphs describing the same OV in the two files are assumed by the user to be potentially causally related to the same HV's;

(ii) the limits governing the pooling of rules (maximum and minimum lag values, degree of similarity between rule parameters) are the same in the two files.

Furthermore, it is assumed that the user-specified statistical measures for fitting morphs (such as the minimum number of datapoints able to start a trend, minimum significance allowed for adding/dropping points at the two ends of an iteratively established trend, etc.; see [1] for details) were the same when the source files were generated from the "raw" datafiles. Similarly, the user has the option of merging knowledge bases governed by the same rule-pooling laws. (Our system checks automatically for this and the above conditions.)

We emphasize that merging source files and knowledge bases are user options. Even if the conditions noted before are satisfied, the user may decide not to proceed with the merging because the source files/knowledge bases were generated by two sites separated by a long enough distance or time period to render the results incompatible.

As another option, the user can display any set of rules in an English-like transcription. He can also eliminate from the knowledge base any rule he considers wrong. He can get the User's Manual on the screen and is guided continually in responding to system questions. The user options are explained briefly in the Appendix.

7. SUMMARY

We have implemented a noise-tolerant, pattern-directed inference system capable of making inductive generalizations concerning rules of hidden variables--variables that can be measured only intermittently or periodically. The implementation is a strongly-coupled interactive system that makes full use of the human's knowledge about physical relations and limitations in order to avoid combinatorially explosive situations. Some of the user options, such as merging source files and knowledge bases under certain conditions, would make our system useful also for an environment of distributed processing and intelligence.

A wide variety of expert systems that need numerical estimates of variables which cannot be measured at arbitrary times, could make use of our system as a computational module.

8. ACKNOWLEDGEMENTS

The project has been supported by AFOSR Grant 81-0220. The interface between FORTRAN IV and ALISP was written by Don McKay. Ernesto Morgado has programmed the morph-fitting package. George Sicherman has been a constant discussion partner and critic. Michael Belofsky did the word-processing for the manuscript.

9. REFERENCES

[1] Findler, N.V. and E. Morgado: Morph-fitting--An

effective technique of approximation (To appear in IEEE Trans. on Pattern Analysis and Machine Intelligence, Special Issue: Computer intelligence--three decades, March 1982)

[2] Findler, N.V.: A multi-level learning technique using production systems (To appear in Cybernetics and Systems, March 1982).

[3] Findler, N.V.: Pattern recognition and generalized production systems in strategy developement (Proc. of Fifth Internat. Conf. on Pattern Recognition, Vol. I, pp. 140-145, 1980).

[4] Findler, N.V.: An expert subsystem based on generalized production rules (Submitted for publication).

[5] Freund, J.E. and R.E. Walpole: Mathematical Statistics (Prentice-Hall: Englewood Cliffs, N.J., 1980).

[6] Wonnacott, T.N.: Introductory Statistics (Wiley: New York, 1977).

APPENDIX

The following briefly explains the user's possible actions. The interaction is highly systems-aided. The system asks the user to confirm each response to avoid mistakes.

To enter the Generalized Production Rules (GPR) system, the user types "-ON,GPRS". This command begins a procedure which puts the user into the LISP system and also loads the top-level functions of the GPR system. After that, the user simply types "(BEGIN)" to initiate a run. The system will respond by asking the user to specify which of several possible actions he wishes to take. The options available to him are:

(1) INTRODUCTION--A short description of the objectives and the user environment of the GPR appears on the screen.

(2) USER'S MANUAL-- A detailed set of instructions on how to input datapoints for OV's and HV's, how to respond to system request, how to initiate the MFP, and how to obtain estimates of HV values appears on the screen.

(3) LOADING DATA FOR OV'S--The system guides the user step-by-step. He can invoke MFP at any desired time. He would also specify the limits of meaningful lag values.

(4) LOADING DATA FOR HV'S-- Here again the user is guided step-by-step. As with (3), data can be loaded either into a new datafile or into one used before. The user has to state which OV's and HV's may be causally related. Rules are automatically established, pooled when possible, and displayed. The knowledge base named by the user is updated. The user can eliminate a rule after it is displayed, if he so desires.

(5) PREDICTION--The user is asked to specify OV data potentially related to the HV whose estimate(s) are sought at a given value of the independent variable, and the number of estimates wanted. The system returns as many estimates as possible up to the desired number, and the rules used for estimation.

(6) DISPLAYING RULES IN THE KNOWLEDGE BASE(S)--The system asks the user which open and hidden variables are connected and what type of morph appears in the rules he wants to have displayed. He can continue to the next rule to be displayed (if any), eliminate the rule being displayed, or stop the display action.

(7) MERGING TWO KNOWLEDGE BASES--If the conditions listed in Section 6 are satisfied, the system follows the user's instruction about merging knowledge bases. The combined knowledge base can replace either of the initial ones, another existing one, or form a newly created one.

(8) MERGING TWO SOURCE FILES--If the conditions listed in Section 6 are satisfied, the system follows the user's instruction about merging source files, i.e. files containing morph descriptions of open variable data and hidden variable datapoints.

(9) ENQUIRING ABOUT THE CONTENTS OF SOURCE FILE(S)--At user's request, the system displays the names of open variables, the hidden variables to which they may be causally related, the limits of the meaningful lag values,

and the minimum degree of similarity for pooling rules.

(10) END-OF-SESSION--Exit.

Figures 2 to 4 show the results of fitting morphs to the datapoints of the open variable OV1, and a few values of the hidden variable HV1. (Note that each figure represents information on a separate source file. The knowledge bases generated from these are then merged into a single one and used for estimation.) Figure 5 contains the morphs that fit OV1 datapoints in the vicinity of the time point 113.11 at which an estimate of the HV1 value is sought. Finally, the estimation is performed to yield the value 2.2503 with a confidence level of 0.7565 .

FIGURES 2 TO 5 ABOUT HERE

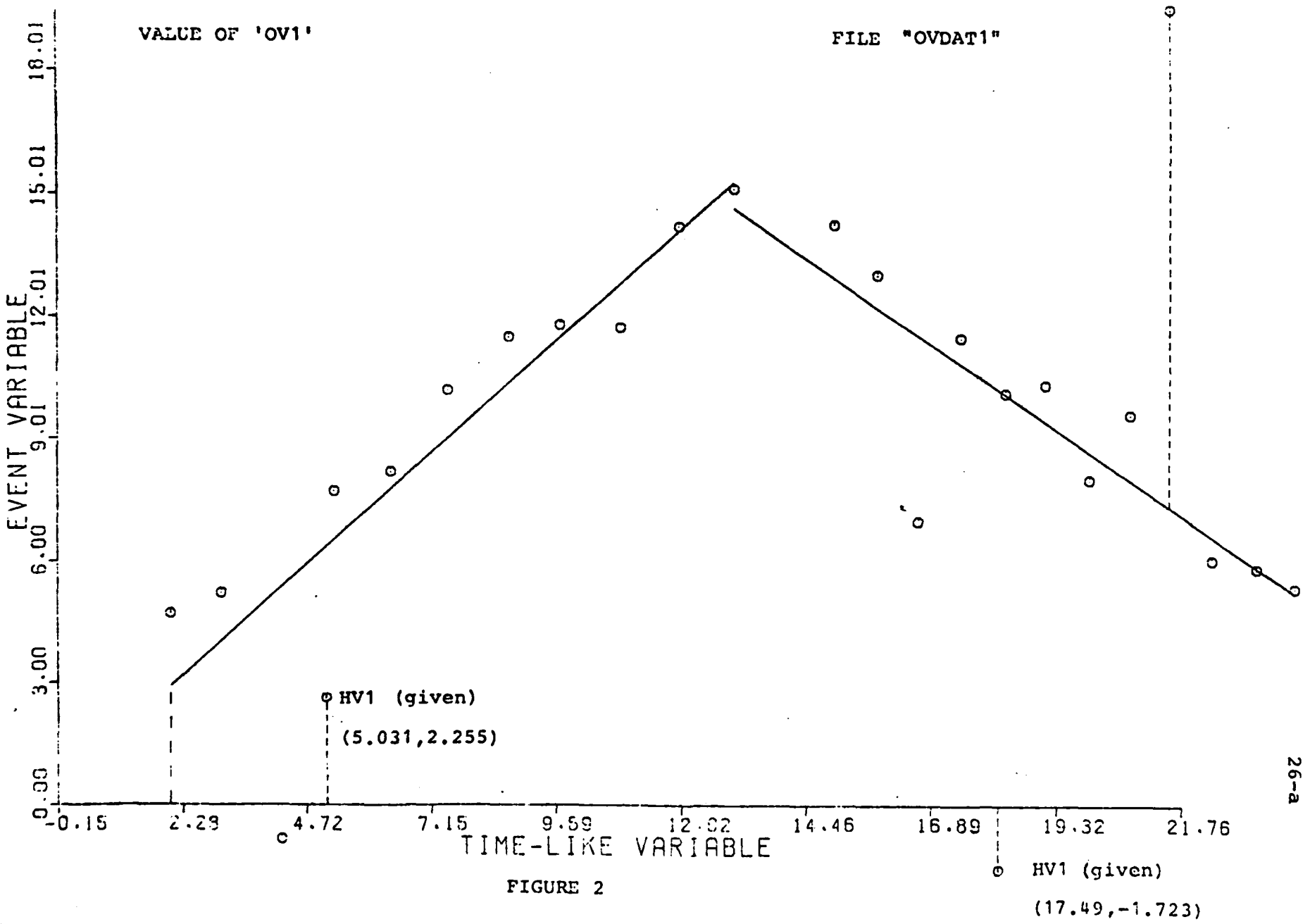

VALUE OF 'OV1'
FILE "OVDAT1"
EVENT VARIABLE
0.00
3.00
6.00
9.01
12.01
15.01
18.01
-0.15
2.29
4.72
7.15
9.59
12.02
14.46
16.89
19.32
21.76
TIME-LIKE VARIABLE
HV1 (given)
(5.031,2.255)
HV1 (given)
(17.49,-1.723)

FIGURE 2

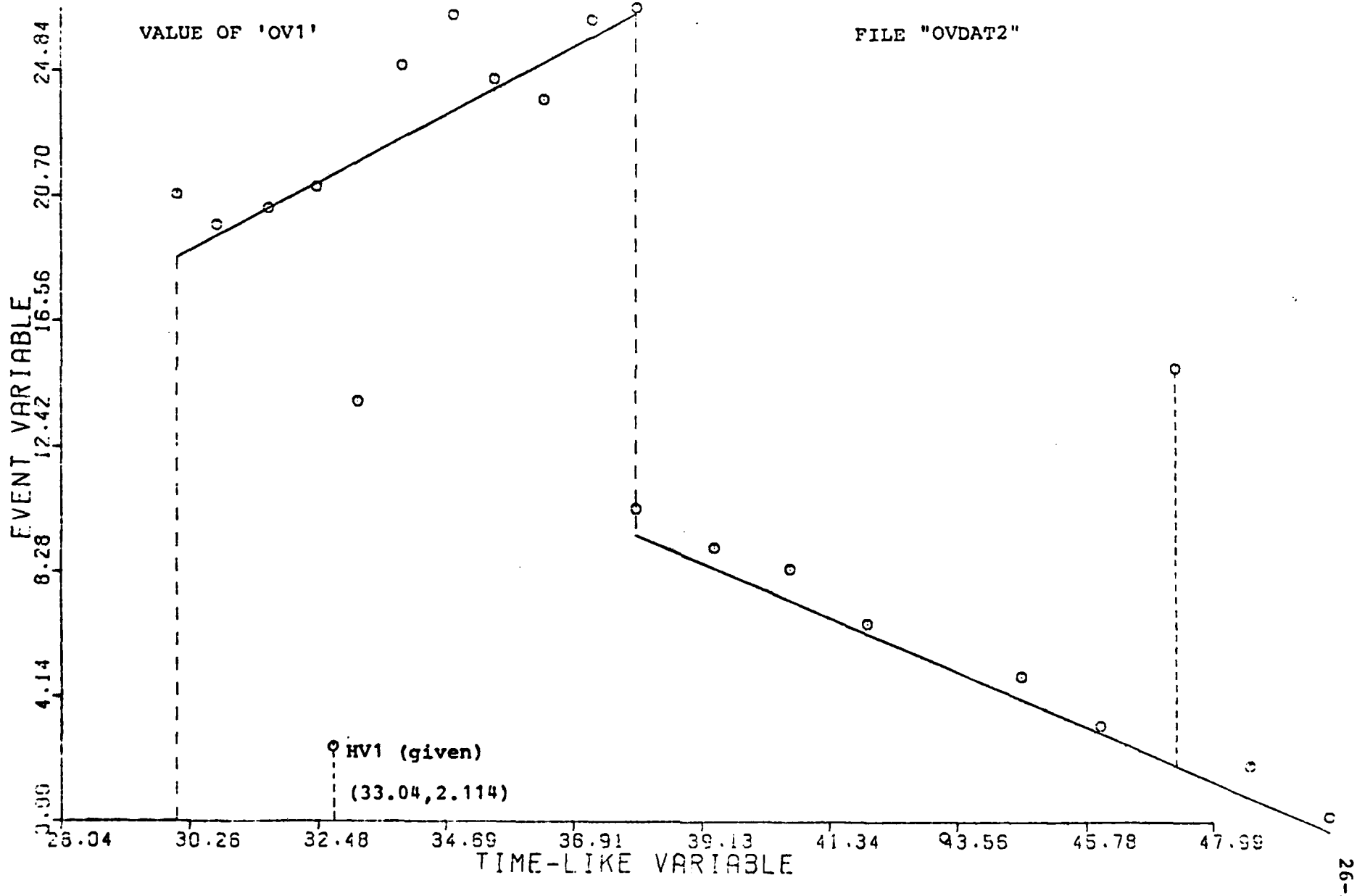

VALUE OF 'OV1'
FILE "OVDAT2"
EVENT VARIABLE
24.84
20.70
16.56
12.42
8.28
4.14
HV1 (given)
(33.04,2.114)
28.04
30.26
32.48
34.69
36.91
39.13
41.34
43.56
45.78
47.99
TIME-LIKE VARIABLE

FIGURE 3

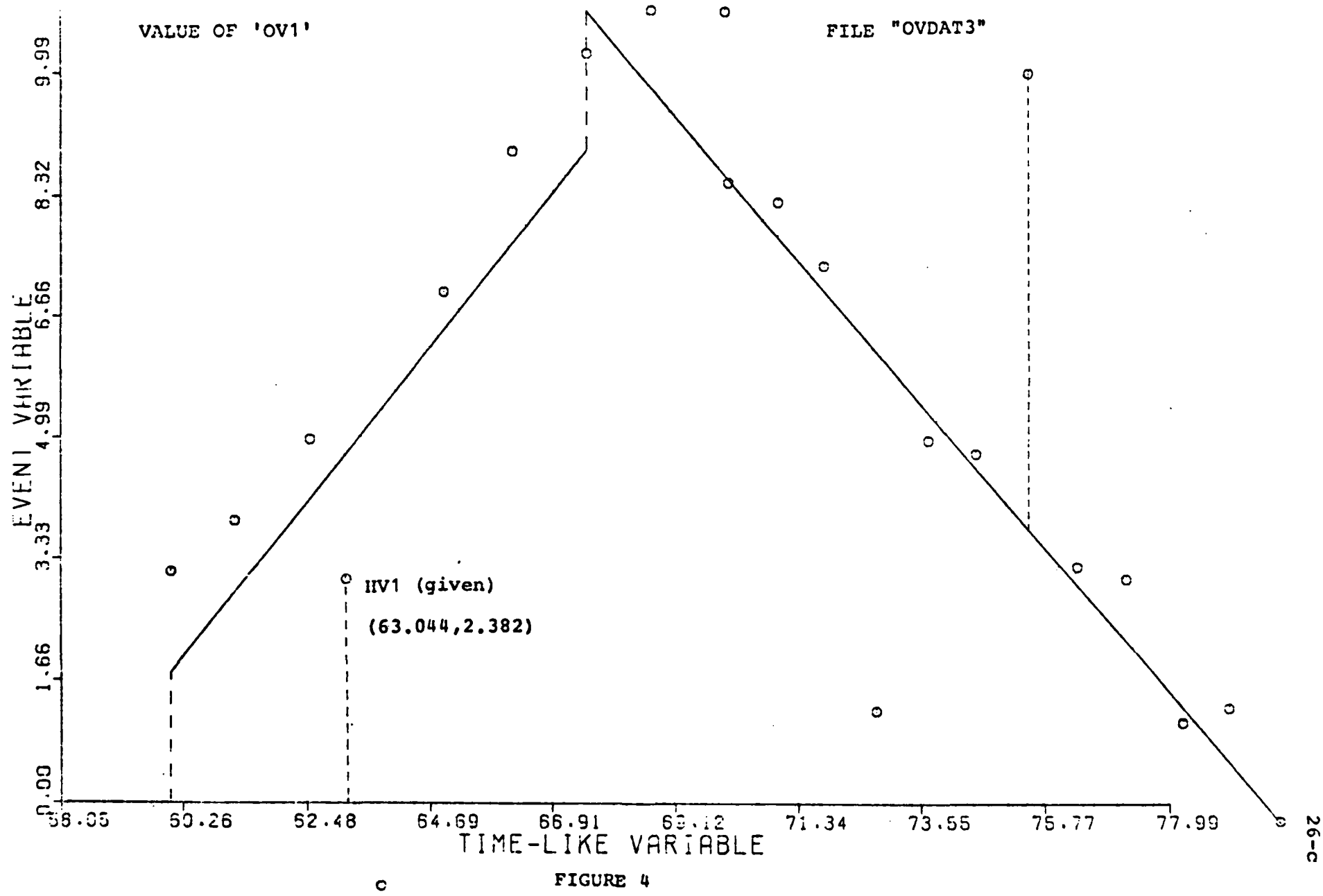

VALUE OF 'OV1'
FILE "OVDAT3"
EVENT VARIABLE
0.00
1.66
3.33
4.99
6.66
8.32
9.99
HV1 (given)
(63.044,2.382)
58.05
60.26
62.48
64.69
66.91
69.12
71.34
73.55
75.77
77.99
TIME-LIKE VARIABLE

FIGURE 4

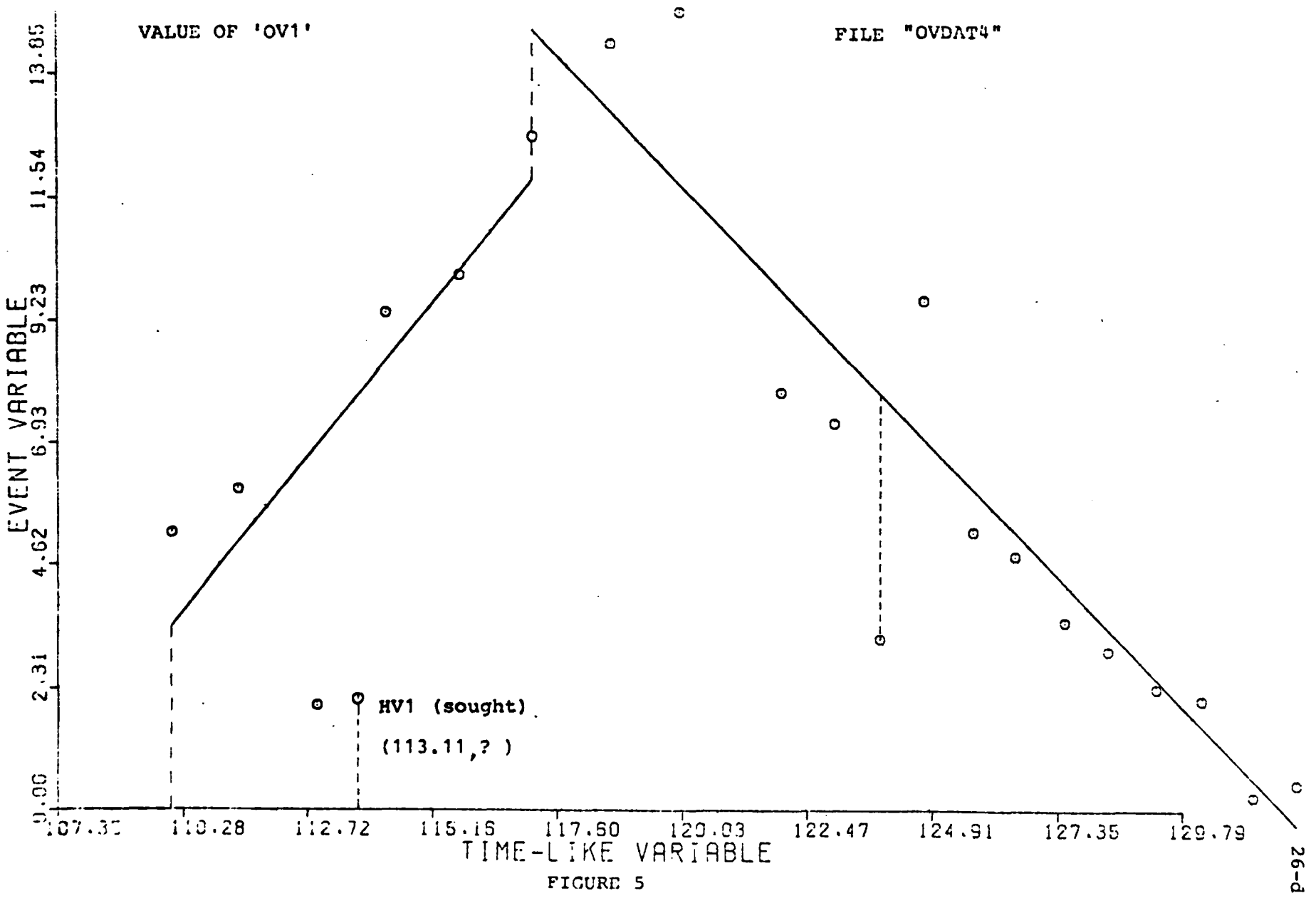

VALUE OF 'OV1'
FILE "OVDAT4"
EVENT VARIABLE
0.00
2.31
4.62
6.93
9.23
11.54
13.85
107.37
110.28
112.72
115.16
117.60
120.03
122.47
124.91
127.35
129.79
TIME-LIKE VARIABLE
HV1 (sought)
(113.11,?)
26-d

FIGURE 5

LEGEND FOR FIGURES

Figure 1 -- The three morph types and their parameters.

Figure 2 -- Plot of the contents of source file OVDAT1: a sequence of datapoints for the open variable OV1, the morphs fitting them, and two datapoints for the hidden variable HV1.

Figure 3 -- Plot of the contents of source file OVDAT2: a sequence of datapoints for the open variable OV1, the morphs fitting them, and one datapoint for the hidden variable HV1.

Figure 4 -- Plot of the contents of source file OVDAT3: a sequence of datapoints for the open variable OV1, the morphs fitting them, and one datapoint for the hidden variable HV1.

Figure 5 -- Plot of the contents of source file OVDAT4: a sequence of datapoints for the open variable OV1 and the morphs fitting them in the vicinity of the point at which an estimate of the value of the hidden variable HV1 is sought.


DAT
FILM