1977

Review of Ecosystem Progress Report No. 1

W. Vaughn Frick

REVIEW OF ECOSYM PROGRESS REPORT NO. 1

W. Vaughn Frick
Ann Arbor, Michigan
January 1977

INTRODUCTION

ECOSYM is a classification and information system for managers of wildland ecosystems. It is being developed by Utah State University with the support of the SEAM project of the U. S. Forest Service. This paper is a review of the first progress report by the ECOSYM project. It will consider the basic design concepts with particular emphasis on the design of the computer based delivery system.

A great deal of high quality work has been completed on ECOSYM. Following the time honored tradition, laudatory praise for a job well done will be held to a minimum. The best work will be covered lightly, if at all. Instead, this review will deal with those areas where the reviewer feels clarifications are needed, further work is required, or important considerations may have been overlooked or underestimated.

SYSTEM DYNAMICS

As any ecologist is well aware, for any system to endure it must be able to adapt to a changing environment. An information system is no exception. In some areas of the report, in particular during discussion of the classification system component set, this need has been noted by the authors. The entire system must be able to react to external stimuli. This capability to adapt must be designed into the system from its inception. This adaptability greatly increases the power of the system and is necessary for a variety of reasons.

No matter how complete the information need survey, ECOSYM cannot predict all future information needs. As our society changes, Congress and the Executive branch will place new responsibilities on the Forest Service creating new information needs that are unforeseeable at this time. The capability to incorporate these information needs must be designed into ECOSYM now or it will become obsolete when new needs are not met.

Specifically, ECOSYM must be able to dynamically add new information, elements of (X), and new data, elements of (E). A major advantage of this capability is that it makes ECOSYM in a sense self-building. ECOSYM's delivery system does not need to be aware of all potential elements of (X) or (E).[1] As new information and data needs are identified, or as previously unavailable data is brought on line, the delivery system's knowledge of the overall framework can be modified to incorporate it. This adaptability has other ramifications as well.

In the author's discussion of cost-effectiveness, they list three criteria an information system must meet to be cost-effective:

---

[1]This concept is covered in greater detail in the discussion of the delivery system below.

1) A given item of information is provided at the lowest cost per unit,

2) The maximum aggregate value of information is provided for a given system budget,

3) The value or benefits received are at least worth the cost of providing the information.

All of these criteria must be met before the system can be considered cost-effective, yet the second and third frequently can't be quantified. Validation of the first criterion requires knowledge of all methods of generating a given item of information. As research work continues, new methods will be developed. Determination of the cost-effectiveness of ECOSYM then becomes an ongoing empirical process.

In addition, it should be noted that the value of X_{ml} (X_{ml} = F(Eijk) will frequently be determined through statistical estimation techniques. All good planning systems used by wildland managers will require the gathering of control information. In other words, given that the information system has stated X will result if action A is taken at place P, the planning system should require the measurement of X to determine if it did indeed result from A at P. This control information can be used to increase the size of the sample through which X is estimated. In this way, the control information can be used to increase the accuracy of ECOSYM's estimate of X.

Information developed through research and control systems must be fed back into ECOSYM. Specific procedures must be developed to gather and utilize this information. Systematic use of this information will continually increase both the cost-effectiveness and accuracy of ECOSYM, but this input to the system must be organized.

There are major advantages in a design which allows the basic framework of ECOSYM to grow and adapt to changing information needs. This growth must be stable and controlled, however. As noted by the authors, large amounts of information have been collected by various groups concerned with land management. They have used various data collection rules and measurement standards resulting in duplication of effort and ineffective sharing of information. By allowing ECOSYM to modify its own framework, the possibility of this type of error is introduced within the system. To prevent this error from occurring, a procedure must be developed to centrally review proposed changes and additions to the data catalogued by ECOSYM. Ideally, this review process would be nationwide in scope and would be concerned with the general use of the data to be added or the information to be developed.

THE DELIVERY SYSTEM

User Level

The system should be designed with the local land manager as the primary client. He will be the most frequent user and will place the greatest demand on the system for precision. As the authors note, systems designed for the higher level user are seldom effective for the lower level user. It should also be noted however that systems designed for the lower level user are seldom inexpensive enough for the higher level user.

The higher level user normally will not need the accuracy required by the lower level user. There will be frequent applications where he needs only "ballpark" estimates without statistically precise answers. The system should be able to give him relatively accurate assessments of a large area without aggregating detailed low level information across the entire area. If the manager later decides a particular estimate is important, he could then go through the more expensive detailed aggregation for a statistically precise answer.

One method of reducing the cost of aggregation in such situations would be to limit the X_{m1} vector to one element, the mean value. This would reduce the amount of information aggregated by a factor of six. Another method would be maximal use of information already aggregated by lower level users.

Data Specification

In the section on computer software, the authors discuss the difficulties of software development with changing specifications for:

1) the kinds of data that will be collected,

2) rules for selecting relevant data from the totality of the data for a given query,

3) guidelines for displaying the relevant data on computer outputs,

4) rules for updating the data base.

The authors state: "If, however, the designer does not have sufficient detail and/or the specifications are changing over time, then MIS design exists under conditions of uncertainty and impossibility." The authors appear to have overestimated the difficulty of the problem. Consider the specifications point by point.

In the first point, the phrase "kinds of data" is nebulous. Since computer stored data is unitless, it makes no difference whether the data refers to vegetation or soil type. Both would be stored in the same manner. Let us assume, then, that the authors refer to problems such as the presence or absence of a particular Eij in the stored data or the particular data structure used (e.g., grid or polygon). Information such as this and more can be stored in an online data description for each data base and used by the delivery system as needed at execution time. In particular this allows the delivery system to operate in multiple data bases of different format and content. This is a proven technique that has been in use for more than a decade. This technique does not answer such questions as what types of data structures the delivery system should be able to manipulate (a consideration discussed below), but if proper programming techniques are used (e.g., modular program construction) the ability to manipulate new data structures can be added on an as-needed basis.

Regarding the second point, for any given query the user is requesting a specific X_{m1}. That X_{m1} is a function of <u>prespecified</u> elements of (E). For each X_m the specification of the elements can be determined from a stored description of that item of information.

Regarding the third point, the delivery system should allow the user to display any piece of information or data through any output mechanism in the system. In other words, the user determines what data is relevant. The user should also be given the ability to specify output formats through the use of general purpose report generators and mapping systems. Again, if modular programming techniques are used new output mechanisms (e.g., interactive CRT graphics) can be added as needed.

Regarding the fourth point, there are only two considerations relevant to data base updates: who can change the data and what possible values the data can have. Security can be handled by a simple password mechanism. Due to the hierarchical nature of the classification system, a simple range check can be used to determine data validity. The minimum and maximum values of k for a specific Eij can be stored as part of the data description of that Eij.

The design of the delivery system is not trivial, but it is far from impossible. It is well within the scope of current computer software technology. It can be conducted in parallel with the development of the classification and information framework, and (as discussed below) there are many reasons why it should be designed in parallel.

<u>Data to Information</u>

Given that the specific "k" values are known for the independent variables of $X_{m1} = F_E(E_{ijk})$, a method is needed to apply the function F_E to obtain the value of X_{m1} requested. The authors suggest using the values of k as indices into externally stored tables of X values. There are certain disadvantages to this method.

First it should be noted that as the number of independent variables increases arithmetically, the size of the needed table increases geometrically. The file storage needed for the table of values of a complex dependent variable of a complex dependent variable could be quite large.

The values in the tables will have been generated through some predictive estimation equation (probably developed from a regression model). If a method of encoding and using this equation were developed, the external storage for the tables would not be needed. This method could also be faster.

As a general rule of thumb, an Input/Output operation takes approximately one thousand times as long as an arithmetic operation. A computer can do a lot of calculating in the time it takes to read a record from the table file. Which method would be faster depends on the number of arithmetic operations necessary to perform the calculation, the number of times the calculation needs to be performed, and the number of usable pieces of information gained per I/O operation.

The delivery system could be designed to accommodate both methods and/or a combination of the two. Whether or not this is necessary could be determined by testing. (This would obviously be preceded by better specification of the functional forms and estimation of the frequency of information use.)

Delivery System Dynamics

One major advantage of using stored descriptions of the elements of (X) and (E) is that it allows only data that is being used to be kept on-line. Moreover, this technique builds in the capability to adapt to changes and additions to the overall ECOSYM framework. If the system is designed to allow the user to easily add more data and information items, it not only will respond to future physical-biotic information needs, but it could also incorporate more volatile-socio-economic data should it become available. There is no possibility of the delivery system being cost-effective unless it can manipulate data whose form and content is not yet envisioned.

Data Structures

As mentioned earlier, the delivery system can be designed to operate on multiple types of data structures. In the past the Forest Service has used many mapping systems such as MIADS, WRIS, COMLUP, GELO, CLOVER, and GRID. Not many have been data base independent in design and none have used both grid and polygon data structures. Viewed as a whole, these systems can be a source of information for the delivery system designer on the advantages and disadvantages of grid versus polygon storage techniques.

In general, the polygon systems need less file storage and are usually more geographically precise. On the other hand, grid systems can manipulate the data more efficiently, particularly as the complexity of the inquiries increases. More precise information on the relative tradeoffs can be gained from study of and experimentation with some of the mapping systems mentioned above. Ultimately, however, the choice of whether the delivery system should use grid structures, polygon structures or both will depend on such information as the frequency of data reference and the type of data manipulation required. The ECOSYM project is uniquely qualified to supply such information. This alone is a major reason why the design of the delivery system should be done in parallel with the design of ECOSYM. Many such questions will arise in the design of the delivery system which only those developing the information framework of ECOSYM will be able to answer.

Interactive Use

In the discussions of the use of the delivery system, attention centered on the generation and display of information, specifically elements of (X). Frequently the user will be able to determine that areas under study have a low potential for the development of a certain resource simply by mapping combinations of physical characteristics (e.g., locating certain soil types on steep slopes). How low the potential is may not concern him. If the delivery system can manipulate and display elements of (E) directly, the question can be answered without the expense of translating the data into specific information.

The user should also be able to minimize the area to which his question pertains. For example, he may want to perform the following sequence:

1) Calculate the growth potential on area P_1.
2) Find the areas within P_1 where the growth potential is greater than X_1 and the slope is less than E.
3) Name those areas P_2.
4) Summarize growth for P_2.
5) Summarize soil loss for P_2.
6) Code soil loss in P_2 into increments of X_2 and call the resulting variable ABC.
7) Map ABC in P_2.

By restricting the data set in Step 2, the user has gained two important advantages. First, he has reduced the amount of data the delivery system must process in subsequent operations. Second, he has increased the relevance and reduced the amount of information to cross his desk. The ability to use dynamically produced subsets of a parent data set is gained through stored (and for siblings, dynamically generated) data descriptions. Note that in Step 5 the user has asked for a previously unreferenced element of (X) for his subset area. This requires that the delivery system go back to the parent data base and extract those elements of (E) necessary to calculate soil loss. A detailed explanation of how this could be done is not within the scope of this review. Basically, it involves keeping track of parent-sibling relationships as a tree structure whose nodes are data base descriptions. Suffice it to say that data reduction techniques are available that are feasible, powerful and cost effective.

External Ties

Although it was not mentioned in the report, it is assumed that methods will be availablethat will easily allow a non-programming user to get data into the system. One area that is frequently overlooked, however, is the ability to get subsets of the data base back out again.

As mentioned previously, planning control information will be available. It will consist of measured values of elements of (X). If the measured values of X differ significantly from the predicted values supplied by ECOSYM, someone will want to get the relevant elements of (E) out of ECOSYM in a form suitable for the local statistical analysis routines. The delivery system should supply this capability

CONCLUSIONS

With the exception of the delivery system, the internal design of ECOSYM appears to be superlative. More attention needs to be given to ECOSYM's ties to its operating environment. A determination of not only what information is needed but the frequency and manner of use needs to be made. This does not mean simply how many questions a data item is relevant to, but how many times each question will be asked. This information would be invaluable to the delivery system designer andit could be a relatively inexpensive byproduct of the ECOSYM research.

It should also be recognized that while this information can greatly contribute to the cost-effectiveness of the initial design, information need will change. ECOSYM should be designed to adapt to these changes.