2

NASA TECHNICAL MEMORANDUM

NASA TM X- 62,240

NASA TM X-62,240

EARTH RESOURCES GROUND DATA HANDLING SYSTEMS FOR THE 1980'S

Edgar M. Van Vleck, Kenneth F. Sinclair, Samuel W. Pitts, and Robert E. Slye

Ames Research Center
Moffett Field, Calif. 94035

152

TABLE OF CONTENTS

LIST OF TABLES

SUMMARY AND CONCLUSIONS

The purpose of this study has been to understand the system requirements of an operational data handling system for earth resources in the decade of the 1980's, focusing on the problems that will be encountered in meeting the stringent agricultural user requirements of that time frame. Such an understanding of requirements will be essential not only in designing the ground system that will ultimately handle the data but in design studies of the earth resources platform, the sensors, and the data relay satellites that may be needed.

The starting point was to determine the rate at which data would be received in the facility from such an operational system. After the anticipated data rate was analyzed, it was possible to study implications for the alternative systems that would process it; for example, to determine how preprocessing might be done in a system which handles a volume so much larger than present systems and produces an end product so much more sophisticated. It was also considered important to analyze in some detail the various algorithms that must be implemented on computers for automatically classifying the data; in particular, it was necessary to outline the present status of such algorithms and to describe both the progress that is being made in their development and the problems that stand in the way of further development.

In addition to the demands made on the software algorithms by the large projected data loads, it was necessary to analyze the demands made on the computers that implement the algorithms and on the computer memory needed to store the raw and processed data. The emphasis throughout was first on assessing the user requirements in order to project the data load, then on assessing the impact of the data load on the system requirements, and finally on analyzing the interaction between the user requirements and the system requirements.

It was found that although the present day user community is not well enough developed to provide firm user requirements, it is developing rapidly and a statement of the needs of this community sufficient to outline the major system requirements is possible. An analysis of these requirements indicates that a data rate on the order of 2×10^{11} bits per day will be generated, and a ground resolution on the order of ten to twenty meters will be required; this will handle some 75-90 percent of all requirements. Twelve spectral bands will be provided at the satellite, but these may be reduced to as few as three before the data is completely processed.

In studying a range of system alternatives, it was found that many of the apparently open alternatives in system design were foreclosed by certain user requirements and by limitations of available memory, computer, data transmission, and input/output technologies. An evolutionary system was therefore postulated which combines the various discrete alternatives in a way consistent with user requirements and technology capabilities.

Preprocessing and processing requirements were analyzed in sufficient detail to provide an indication of how these must be expanded beyond present capabilities. Preprocessing requirements will dictate that the existing systems must be scaled up significantly. Although the preprocessing loads foreseen are extremely heavy, it should be possible to handle them by the appropriate use of parallel digital operations and other techniques such as special digital logic. Techniques for automatic processing of the data, including classification and recognition, are now undergoing intensive research and will very likely be available when needed to handle the future processing load.

To ensure that these techniques may be used operationally, particular attention will have to be paid to the definition of the output required by the user; in addition, the selection of the proper computers for implementation of the algorithms will be critical. Particularly if the implementation takes place early in the decade, hybrid computers will

probably play a large role, while later, for flexibility, use will have to be made of high speed digital computers. Computer speed is increasing, while at the same time computers are becoming available that display lower and lower cost per calculation; both trends in the evolution of computing capability will be important if digital computers are ultimately to play a large role. In addition to the hybrid computer and the general purpose digital computer, special purpose digital logic will almost certainly be used both in early and later versions of the system. This is because the speed of such computing elements will permit them to solve such problems as likelihood ratio algorithms as much as 1,000 times faster than, for example, the IBM 360/75. Of the general purpose digital computers, parallel processors were shown to be superior to other computers for processing picture elements.

The various sections of the study combined to make it possible to present the following tentative conclusions:

1. It appears to be possible to outline in broad terms the requirements of a future data handling facility capable of handling 2×10^{11} bits per day. The study does not permit making detailed statements about the characteristics of such a system, but sufficiently precise statements can be made to identify major system elements that need further work.

2. A system of the capability described above appears to be feasible, to the degree of certainty permitted by the broad nature of the study, but the feasibility is marginal and depends on many detailed tradeoffs that have yet to be made.

3. Pattern recognition algorithms that are capable of classifying the data are now available, but need much work before they will be useful in an operational system. Increases in speed and in accuracy will be required, particularly in speed; choice of computer for implementation will be critical, since some of the best algorithms require a digital computer and computer speeds may not be adequate for such algorithms.

4. Hybrid (digitally supervised analog) computers will probably have to be used for classification in any early systems. It will be highly desirable to implement digital computers, as their speed improves, for recognition algorithms to take advantage of their greater flexibility.

5. Much research is required in the area of unsupervised algorithms to provide rapid classification of data without extensive quantities of ground truth required by supervised algorithms. It would be highly desirable to be able to use unsupervised algorithms to provide initial classification into natural clusters before applying ground truth comparisons. However, unsupervised clustering algorithms generally are too slow in comparison with such supervised methods as likelihood ratio. Faster clustering techniques are required.

6. Because present methods of classification require extensive quantities of ground truth for accurate classification, there is a requirement for additional research in the area of signature extension techniques. This research should be combined with the research on unsupervised algorithms, since it is possible that the biggest breakthrough in this area may come with unsupervised algorithms.

7. There should be research into the combination of sequential classification techniques with ordinary recognition techniques and the automation of such combined schemes. The potentiality for increased accuracy of classification appears to be substantial, but the increased memory requirement must also be considered.

8. Very careful attention should be given to the various tradeoffs in memory allocation; large amounts of raw data may be stored, with computing on demand, or computing of all data may be done with storage of only processed data, or any of several intermediate schemes. It appears that in the 1980 period, large memories will be easier to obtain than additional computing capacity, so it would appear that the schemes employing large memory combined with computing-on-demand are more likely to be used in an operational system.

9. The user requirement for data in terms of quantity, timeliness, freshness, and completeness must be more carefully assessed. Various

types of system users must be distinguished and the relative difficulty of satisfying each type of user must be studied. Some users will wish to process data on demand from a wide selection of stored raw data; other users will be satisfied by a regularly computed and disseminated product much like the one the weather bureau now operates; for some other users it may be necessary to process everything received in their area. User requirements will be in existence for other than merely recognition data and these other user requirements for ancillary data handling programs will have to be studied in more detail, particularly in terms of their effect on system sizing.

10. An analysis of schemes and strategies for the operation of the large-scale memory systems will be essential to the design of the future system. User requirements for data storage will need to be more carefully assessed. It will need to be determined how long data should be stored for data at each degree of accessibility, how often memories at each level of accessibility are to be purged, etc.

11. In spite of all the remaining uncertainties in the memory requirements, there will very likely be a requirement for 10^{13} bit and larger accessible and erasable memories. If non-erasable memories of correspondingly larger size are available, they may be acceptable. The state-of-the-art in large-scale optical memories is already equal to this, although it may be necessary to increase the read-in rate by a factor of two.

12. Although bandwidth requirements will be severe, and there may be other reasons to do onboard processing, the data acquisition rate is so great that very little onboard processing will probably be done. The area of onboard processing should be studied more intensively, however, as a separate task since considerable savings may be made if onboard compression of the data turns out to be feasible.

13. There will probably be a need for a synchronous relay satellite, possibly using laser or millimeter transmission. This is due to the high bandwidth requirement of the system. The feasibility of such a relay satellite will need to be studied as a separate task, although it is

apparently feasible to modulate a laser with the one gigabit of data per second required by the system. This high transmission rate would occur during only a relatively few seconds of each pass. The probable requirement for a relay satellite follows from the fact that without the relay onboard storage on the order of 10^{11} bits may be required, which would be beyond the state-of-the-art for onboard memories.

14. Because of the possibility of cloud cover during the transmission, there would need to be several widely scattered ground receiving stations. Because the bandwidth is so great, it would probably be impractical to relay the data from these to the data center via terrestrial lines, and therefore the relay satellite should be considered for this relay function as well as its primary relay function. The data would be transmitted to a ground station that is not obscured by clouds and recorded there. As soon as both the ground station and the data center station are free of clouds, the data would be transferred from the remote station to the data center. This dual function would more effectively utilize the data relay satellite, which otherwise would be used only for a few minutes each day. It is highly likely that a relay satellite of this capacity would find cooperative users who could utilize the capacity when not in use for the data handling system. This prospect deserves separate attention.

INTRODUCTION

Planned and projected experimental earth resources satellites will acquire data at a rate far higher than that of any other peaceful space program to date. The first operational systems (ref. 1) are expected to generate even larger amounts of data. It seems clear that if this vast amount of data is to be made useful, the key to any successful earth resources monitoring system, our methods of handling and disseminating data must be vastly improved. This is especially true in the context of an operational system where, by definition, the data collected are destined for use in predefined applications in which timeliness and completeness are paramount. The purpose of this study then is to define

in preliminary fashion the scope of the operational earth resources satellite data handling problem and to consider the various alternatives for making this data available to the ultimate users in a timely and satisfactory manner.

To limit the problem, but hopefully not the utility of the results, a single discipline, agriculture, has been chosen for analysis. The requirements for agricultural data appear to be unique. First of all, relatively high spatial resolution is required. Second, extensive coverage is needed, and third, the timeliness requirements are demanding. These factors, coupled with the fact that multispectral data appears to be required for the kinds of analysis important to agriculture, combine to make this discipline the dominant data producer for an operational earth resources system. For the time frame under consideration here (1980 to 1990), operations have been limited to the United States and its territories. This is not to suggest that some coverage in countries other than the United States will not be available and the acquired data provided to foreign governments, but rather that the data handling complex will be primarily sized to accommodate users in this country and any large scale foreign use would probably involve development of separate data facilities in each of the user nations.

In addition to the constraints outlined above, several other limitations have been imposed on the study. Onboard data processing has been considered only briefly. Optical processing has been omitted, either onboard or in the ground system; rapid development in this field makes it likely that some or all of the earth resources data processing jobs will be possible using optical techniques in the latter part of the 80's, but there are not enough studies available at present that meaningfully analyze such optical processing methods. The experimental and evolutionary nature of the earth resources program at this time must also be stressed. There are large uncertainties in a complex, developing field of this type, which involve a many-faceted user community that does not yet clearly understand the potential or limitations of synoptic

space-acquired data. The critical nature of the data handling problem makes it essential to begin consideration of the systems and technology that may contribute to a solution.

Over the next five years, our understanding of space-collected earth resources data should improve dramatically: Ways people handle and think about remotely-collected data will change as new tools are developed and new insights are provided by this continuing stream of imagery from space. Thus, any estimate of the data yield of a future operational earth resources satellite has much uncertainty; but changes greater than a factor of 2-10 from the nominal values postulated appear to be extremely unlikely.

This paper consists of the following sections: data anticipated from an operational satellite system; data system concepts potentially capable of handling this data flow; preprocessing requirements for data that will be either manually interpreted or machine interpreted; methods of machine processing at various levels of sophistication; current and projected state-of-the-art for digital and analog computers and input/output devices; and analysis of the conceptual systems using the results of the subsequent sections of the report.

The overall system that is considered in this report is diagrammed in figure 1. The system consists of an earth resources satellite relaying its data to a command and control center via a synchronous relay satellite through a mission control center to a central data processing facility. At the same time, data is relayed from an aircraft and from ground truth sites to the central data facility via courier, radio, and terrestrial lines. These elements are described in more detail in the section on system alternatives but are here for convenient reference in reading the study.

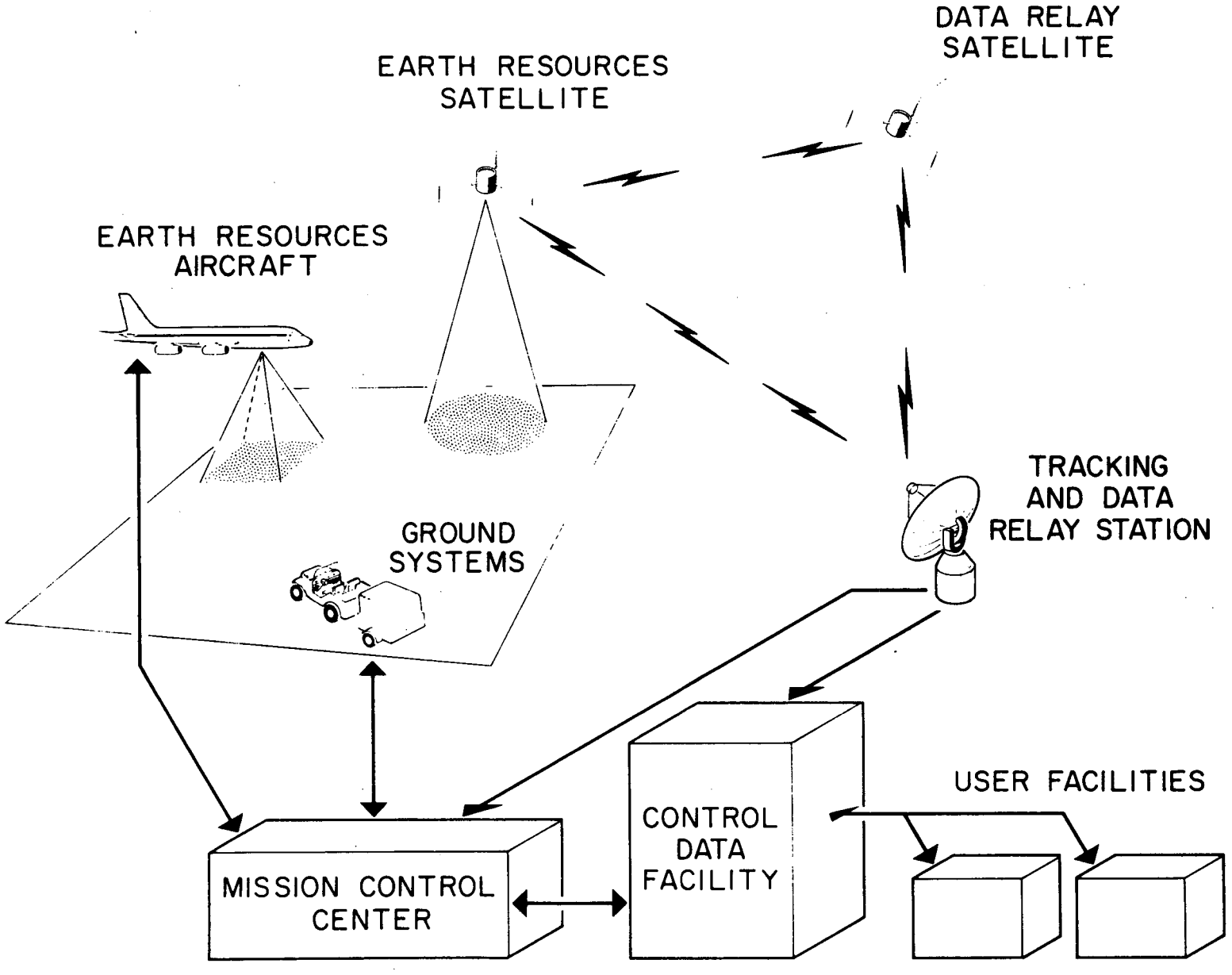

Figure 1. - Earth Resources Data Handling System--Overview.

DATA CHARACTERISTICS

The design of the data processing system depends on the types and volume of data anticipated from an operational satellite system. It is important to clearly define what is meant here by the term operational. First of all, the term is used to clearly distinguish this kind of system from the experimental program now underway with ERTS: Data is assumed to be collected only for predetermined and very specific purposes and it is assumed that the nature of the ultimate user product is well defined. Collected data not related to specific resource requirements will not be processed. The system has the inherent capability to provide the required data as needed.

This precludes the experimental use of the system to the extent that special types of processing would not be available, but bulk data in either the raw or preprocessed (rectified, photometrically corrected, and related to the map base) form would be available to experimenters for use in their own processing facilities. Operational timeliness requirements generally do not apply to experimental work, for which the general purpose digital computers usually available to researchers at colleges and government laboratories should be suitable. The data characteristics for the various resources disciplines will next be discussed together with the data characteristics of an operational earth resources satellite.

Data Generation

The data generation characteristics for the various resources disciplines are determined by the resolution requirements, the area to be covered, the spectral bands required, and the repetition rate needed. To ascertain the data production characteristics of an operational earth resources satellite, these factors must be understood for the various disciplines. Unfortunately, conclusive data regarding these requirements are not yet available because of the evolutionary

nature of the program. However, the range of uncertainty is being reduced as users interact with the space and high altitude data now available.

The presently-estimated requirements for key earth resources disciplines are shown in table 1,* which also shows the relative data production rate of these disciplines and illustrates the dominance of the agricultural requirement.

Certain types of measurements have been excluded from table 1. These include virtually all requirements calling for an overflight of the same region at intervals of a few days or less. Both emergency situations and some routine measurement problems, e.g., some oceanographic and meteorology objectives fall into this category. Some or all of the excluded oceanographic and meteorology objectives could conceivably be met from a geosynchronous satellite system because of the typically lower spatial resolution requirements of these disciplines. Thus, much greater coverage could be obtained from a single frame of imagery. This class of measurement has not been considered here.

The requirements for earth resources are shown in figure 2 as a graph of fraction of resource requirements that are satisfied versus resolution. The percentage of requirements satisfied is from 60 percent at very good resolution to 95 percent with the coarsest resolution curve if the resolution is ten meters as assumed in this report.

---

*The data rates in table 1 are derived from the other columns of the table as follows:

$$\text{Data Rate } \left(\frac{\text{Bits}}{\text{Day}}\right) = \frac{\text{Area } (\text{km}^2) \times \text{Bands} \times 10^6 \times 8 \text{ Bits}}{\text{Interval (Days)} \times \text{Resolution}^2 \text{ (meter}^2)}$$

The minimum data rate column is derived by using the combination of factors that give the minimum data rate; correspondingly for the maximum data rate column. The data rates have been rounded to nearest integer.

TABLE 1 REQUIREMENTS FOR EARTH RESOURCES DISCIPLINES

Discipline	Resolution (Meters)		Coverage Interval (Days)	Area Covered* km^2	Bands/ Sensors	Data Rate (Bits/Day)	
	Detailed Survey	Recon. Survey				Min.	Max.
Agriculture	10-30	30-100	7-21	3×10^6	12	2×10^{10}	5×10^{11}
Cartography	3-20	20-200	1825	9×10^6	3	3×10^8	2×10^{10}
Forestry/ Range Land	10-50	50-200	7-30	3×10^6	8	3×10^9	3×10^{11}
Geography	6-30	6-100	365	9×10^6	3	1×10^9	3×10^{10}
Geology	6-100	30-200	365	2×10^6	4	2×10^8	6×10^{10}
Hydrology	3-100	50-250	10-20	1×10^6	4	2×10^8	4×10^{11}
Meteorology	1000-2000	1000-4000	.25-1.0	30×10^6	2	1.0×10^8	2×10^9
Oceanography	20-300	200-1000	14-30	15×10^6	4	1×10^8	1×10^{11}

* Reference 1, Data for 48 states.

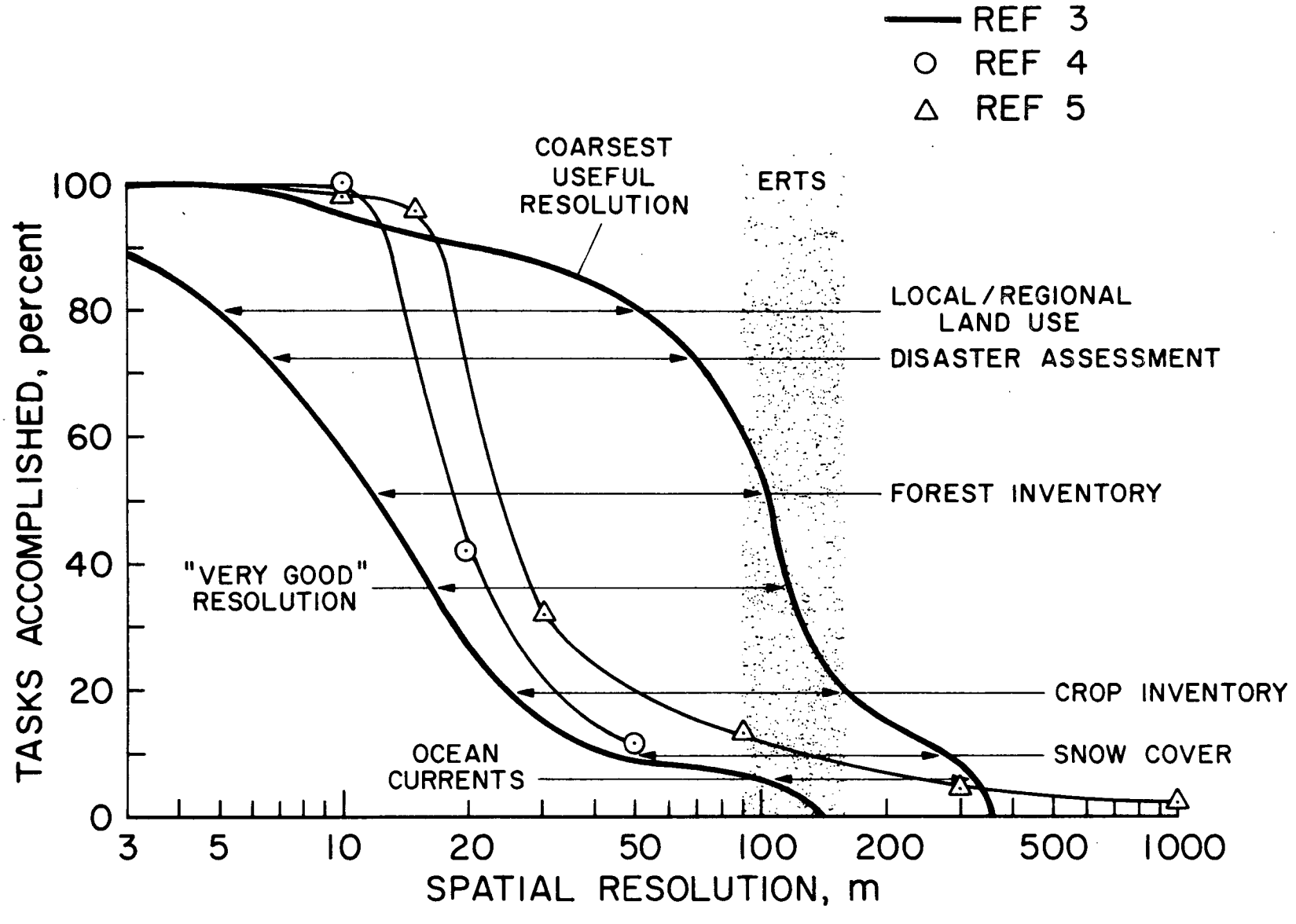

Figure 2. - Percent of Tasks Accomplished Versus Resolution.

It is commonly thought that the data rate of a system is inversely proportional to the square of the resolution; one of the remarkable facts that can be derived by also including such a curve as the present one in the consideration of data rate is that the data rate actually increases faster than the inverse square power of the resolution, since improved resolution actually will increase the demand for the data and will thus act to increase the output requirements of the processing system.

Monitoring of emergency situations, i.e., floods, earthquakes, forest fires, etc., does not appear to be a suitable application area for earth resources satellites. These discrete events generally require essentially immediate coverage with high coverage repetition and high spatial resolution. It would be merely fortuitous if a satellite, even in a multi-satellite system, were to overfly the affected region at the proper time and, of course, the high coverage frequency required would not be available. The space shuttle in sortie mode could conceivably be used for an initial look at such disaster areas, but again it could not provide the coverage frequency required, and the response time (time from recognition of the surveillance requirement to launch) would probably have to be on the order of a few hours to be effective even for an initial survey. Further, the possibility of cloud cover obscuring the affected region must be considered. Nighttime surveillance further complicates the problem. Consequently, it is felt that such disasters will continue to be monitored by aircraft, helicopters, and ground units operating in and over the affected regions as required. Remote areas not immediately accessible to such vehicles generally have little or no population and, therefore, do not pose a serious disaster assessment problem.

Based on the data generation characteristics outlined above, it appears that the processing of agricultural data will dominate any operational earth resources data processing system. While other resource areas are equally large in area extent and thus produce as much data per band as agriculture during a given pass, the turn-around time in general is much longer and consequently a much slower data processing system

would be tolerable. Thus, it is felt that a data processing system sized to accommodate agriculture will be able to accommodate the other major resource areas.

This emphasis on agriculture also appears to be consistent with some estimates that have been made of potential economic benefits accruing from earth resources satellites. One of these estimates is summarized in table 2. The benefits given are for the entire world. Those for the United States alone would be considerably less. For example, the benefit for "Agriculture, Stress" for the United States would be only $4.5 billion, instead of $27.0 billion. While these types of estimates are extremely controversial, they are felt to be indicative of the relative importance of the various disciplines.

Area/Coverage Considerations for Agriculture

While only some 15 percent of the continental United States is in cropland, not all of the crop areas are contiguous or homogeneous. Consequently, the satellite system must cover a larger fraction of the United States to obtain complete crop coverage. This fraction has been estimated at 25 percent here and the computations hereafter are based on that percentage. The data load, of course, varies from track to track because of the non-uniform distribution of cropland as in figure 3, which shows the dominant agricultural areas. It is assumed here that all of the data collected from the entire crop producing area must be processed in a time consistent with the required coverage frequency. Thus, the daily data load is assumed to be uniform. This simplification is not felt to be seriously misleading and should not significantly affect the data handling considerations. Too, all of the crop areas are not necessarily of interest simultaneously because of cropping and climatic variations. However, at certain times during the growing season virtually all areas are producing and, consequently, the system must be sized to handle this worst case situation.

TABLE 2

NET ANNUAL GLOBAL BENEFITS POSSIBLE FROM EARTH RESOURCES TECHNOLOGY[5]

	Net Benefits In Million $
Agriculture, Stress	27,000
Health, Diseases of Man	16,313
Agriculture, Inventory in Yield	11,340
Oceanography, Fishing	1,560
Health, Diseases of Animals	1,350
Natural Disasters	645
Health, Solid Wastes	373
Resource Management, Soil Survey	115
Geography, Mapping	114
Government Operations, Tax Assessments	87
Search and Rescue	57
Geophysics, Location of Fuels and Minerals	42
Health, Water Pollution	19
Forestry	9
Oceanography, Nautical Charting	9
Oceanography, Ship Routing	7
Health, Air Pollution	1
	59,230

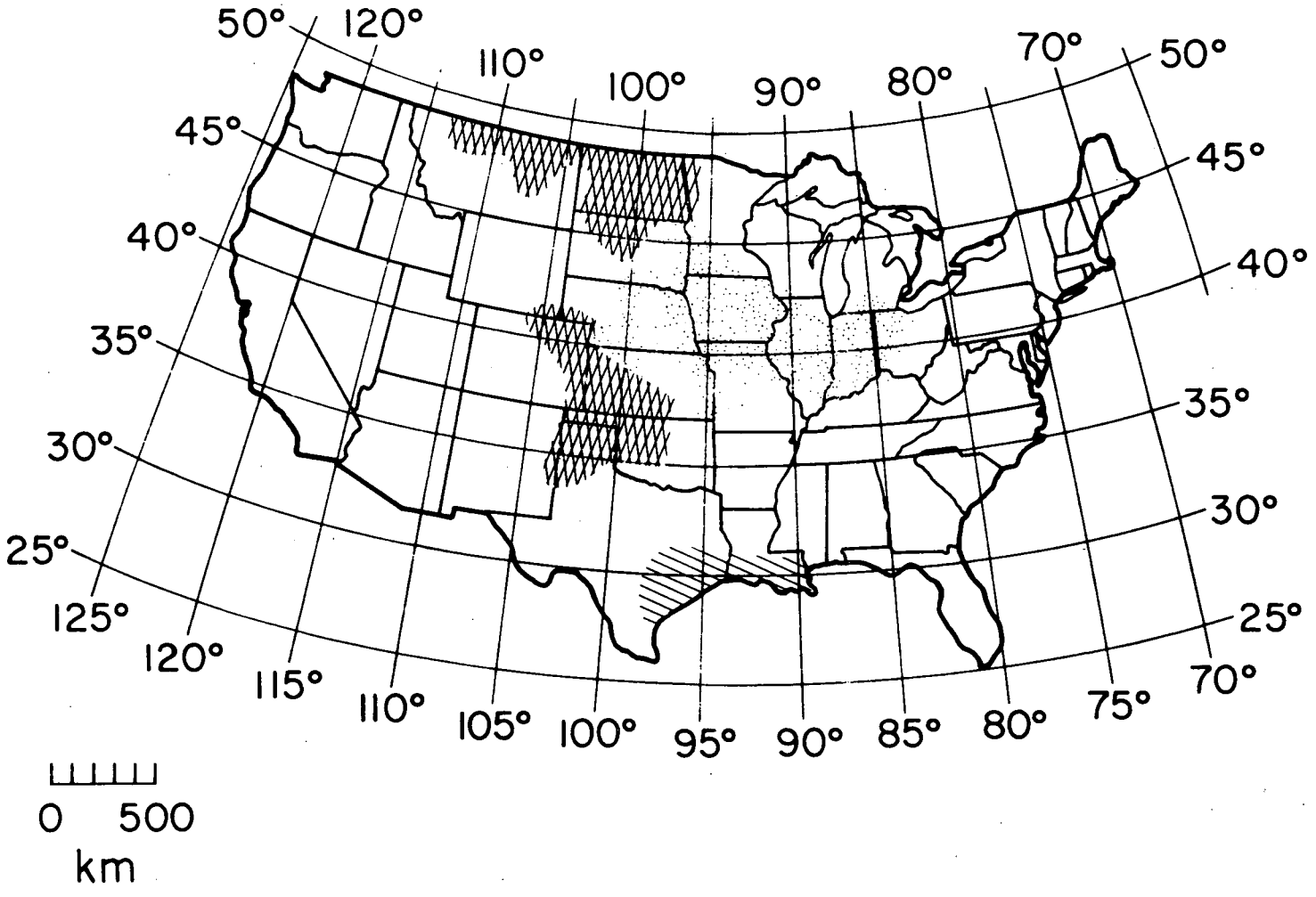

Figure 3. - Crop Inventory and Forecast Coverage.

An operational system demands by its very nature a rather high probability of providing the data required routinely regardless of circumstances such as weather, computer breakdown, etc. Consequently, in designing the space segment for such a system, factors such as cloud cover and onboard equipment reliability must be considered. To provide the coverage frequency specified over the continental United States with high reliability it has been shown (ref. 6) that a near polar orbit system consisting of four satellites is required just to take care of the cloud cover problem. It is assumed that sufficiently reliable satellites can be provided for a suitable operational period so that on-orbit spares are not required.

Spectral/Spatial Resolution

While the evidence appears to show that a sizable number of bands (perhaps 12 or more) will be required to permit selection of the most useful bands for any given crop identification problem, it appears that typically only 3-4 bands, properly chosen, will be required over any one agricultural basin. Figure 4 shows the classification accuracy achieved versus the dimensionality of the selected feature subset for a typical agricultural classification problem. The optimal set of bands required will vary with crop type, time of year, and location; and consequently, the satellite must collect data in more bands than will actually be used at any one time.

The spatial resolution requirements for the system also have a dramatic effect on the data characteristics. While these requirements span a fairly wide range, it appears that the user feels he needs resolution in the 10-20 meter range for the more demanding agricultural applications, such as crop classification. The choice depends, of course, on the field size distribution for the economically important crops. Data in the most recently available census of agriculture (1964) show that only about 8 percent of the total value of agricultural products is produced on farms of less than 200 acres (see figure 5). If all of the

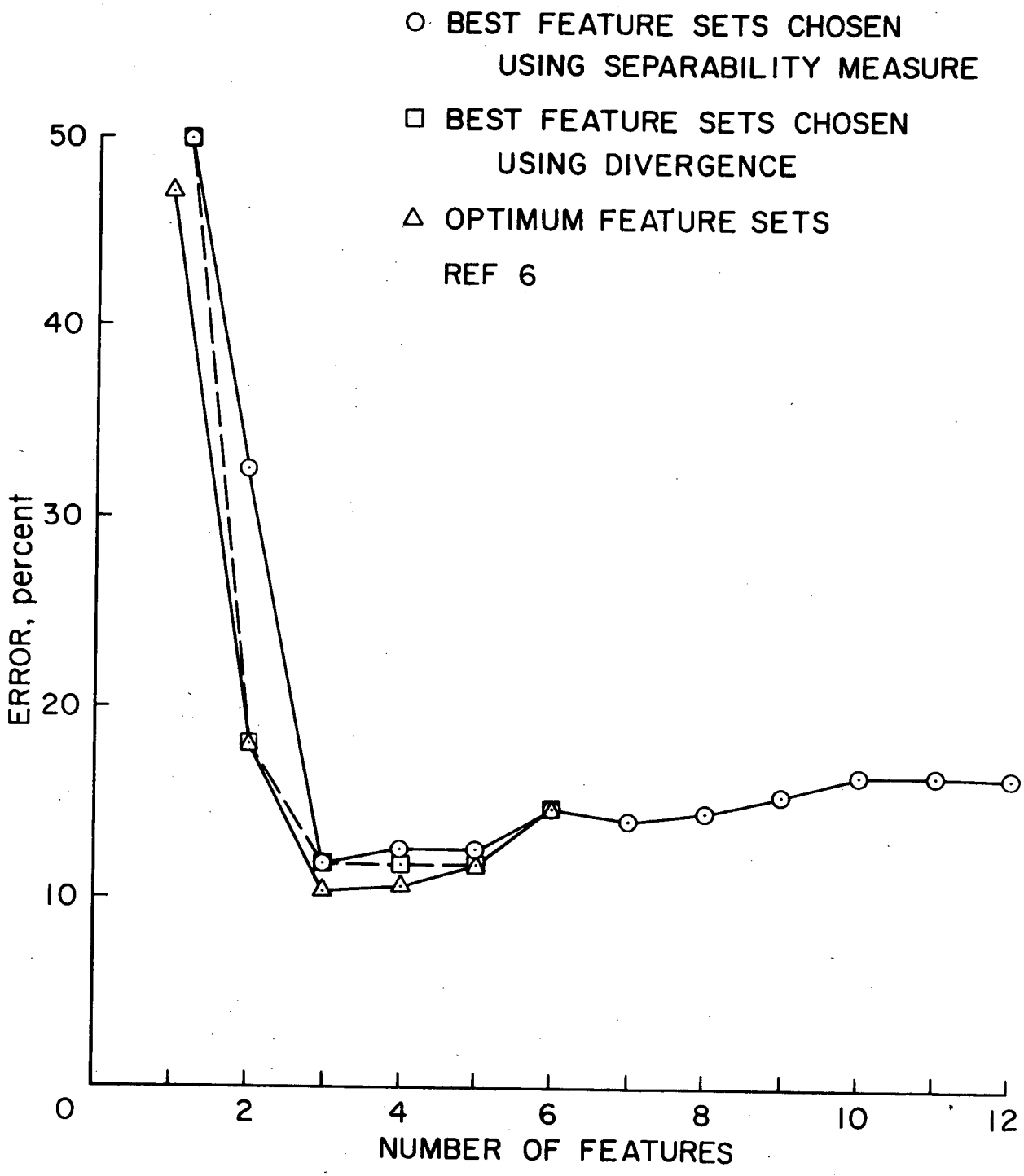

Figure 4. - Percent Classification Error Versus Number of Features.

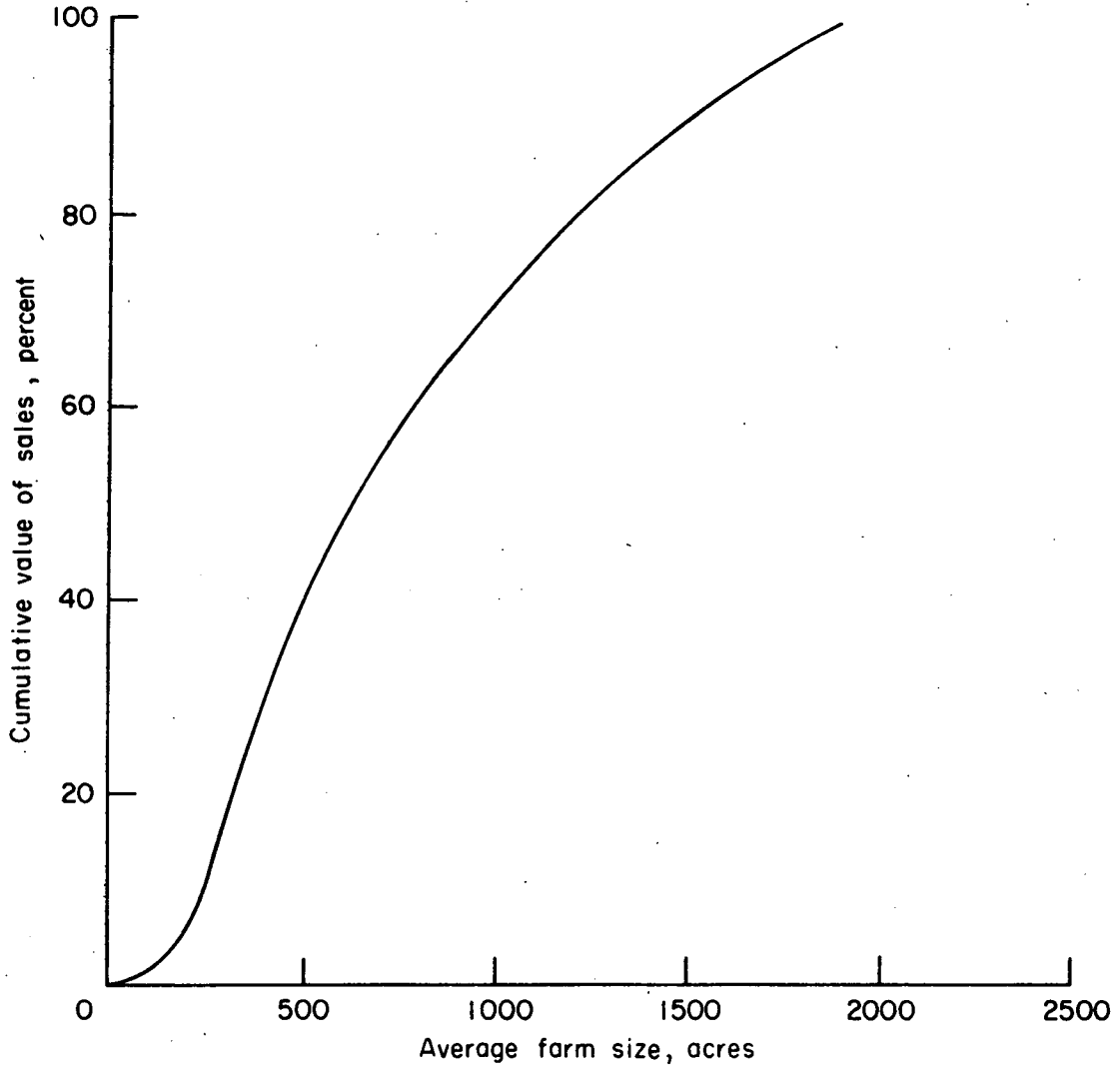

Figure 5. - Annual Value of Sales Versus Farm Size.

data from fields smaller than 200 acres were lost, the effect on the overall result would therefore be relatively small; but as we shall see, even using 200 acres as the threshold criterion, data will not be lost from all fields smaller than 200 acres over a large range of resolution capabilities. Since, in addition to crop type, the amount of the area planted to a given crop is also desirable information, it is important to determine the effect on areal accuracy of resolution element size. Figure 6 shows the areal determination error as a function of pixel size and field size. This figure is based on the assumption that field boundaries can be determined to an accuracy of 1 pixel. While, with proper processing, sub-pixel resolution can probably be achieved when there is adequate contrast, it seems prudent to assume that the error will typically be one full pixel. Assuming a threshold field size of 200 acres, we find that a 10 meter pixel resolution will result in a 5 percent areal error rate. Using the line pair photographic convention and correcting for the resolution loss caused by line scanning (Kell factor), a 10 meter pixel resolution corresponds to about 30 meters resolution, which coincides reasonably well with the stated user requirements. With this figure, fields as small as 100 acres will be observed but with an areal measurement error as large as 20 percent as shown in figure 6.

With the criteria outlined, that is, a threshold field size of 200 acres at a required areal accuracy of ±5 percent, the nominal data acquisition rate for an operational earth resources satellite system is estimated to be 7.5×10^8 bits per second.* This figure assumes twelve spectral channels and an effective ground pixel size of 10 meters.

---

* $Dr = \frac{NSGV}{r^2}$

where N is the number of bands
S is the swath width (185 km)
G is the grey scale (8 bits)
V is spacecraft velocity (7.8 km/sec)
and r is pixel size (.01 km)

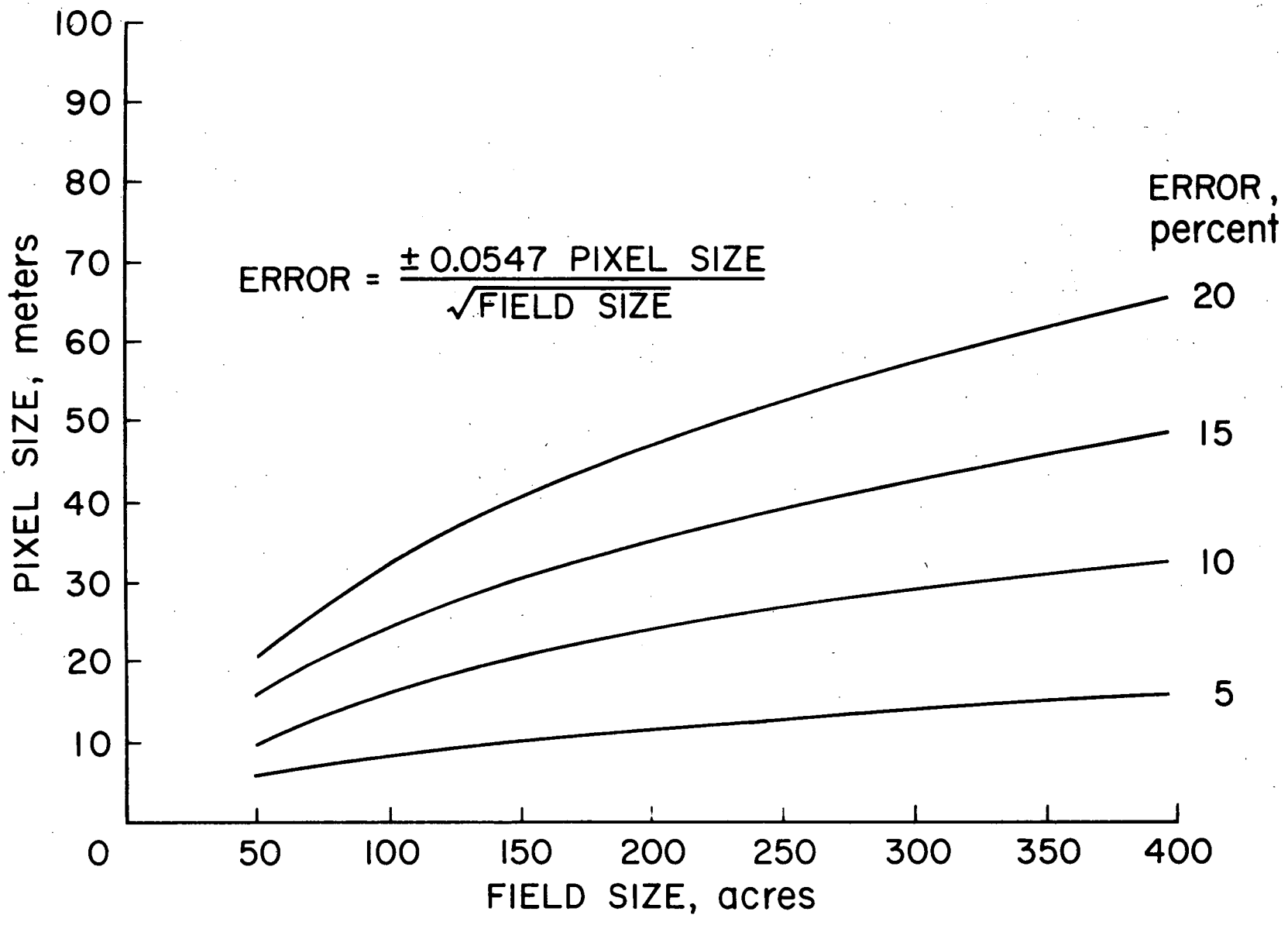

Figure 6. - Error in Area Determination - Pixel Size Versus Field Size.

The data load for a data processing system depends on the useful data collection time for a typical satellite pass and for a specific resource management problem. It is assumed that redundant data collected because of low cloud cover incidence will not be utilized by the system. Since it has been shown (ref. 6) that a four satellite system has a very high probability of providing the required coverage, it is further assumed that effective observations are made over the whole region in the cycle time required. Since the longest track length through this region is 2,000 kilometers, the total data load for this track with the nominal data rate postulated above is 1.92×10^{11} bits*. A 24-hour period is now available to process this data before the next batch will be collected. Consequently, the highest average 24-hour data rate for the system based on coverage of the agricultural areas shown in figure 3 is 2.22×10^{6} bits per second.

We have dwelt on data characteristics because they are a dominant factor in the design of the system. The data rate of approximately 2×10^{11} bits per day that we have just derived will be used throughout the study as the nominal data rate for the system. As we will see, the use of such a large data rate figure will lead to extreme demands for communications link technology, memory technology, and computer technology.

Whether the requirement for data will be as severe as this depends primarily on the area of coverage and the resolution. The area of coverage we have assumed is less than the total agricultural area and so is highly conservative. The requirement of resolution of 10-20 meters is the other remaining source of the high data loads assumed here, but if fields of 200 acres or more are to be observed with high areal accuracy such a figure is justified. (Because of registration errors among the spectral bands, the one pixel error criterion used to derive the 10 m pixel resolution is probably optimistic.)

---

* $\frac{2000 \text{ km} \times 7.5 \times 10^{8} \text{ bits/sec}}{7.8 \text{ km/sec}} = 1.92 \times 10^{11}$ bits

Some users specify better than 10 meters pixel resolution; users cannot be completely satisfied by 10 meters but they would prefer it to a higher figure. There is a class of users (ref. 8) who feel that the quality of 10 meter resolution high altitude aircraft photography is just adequate for their needs, and these should find 10-20 meters acceptable. There is the final class of those who do not need resolution as good as 10 meters--these should at least not object to this fine a resolution. Thus 10-20 meters should be acceptable from a user standpoint.

Referring to figure 2, it is clear that from 50-90 percent of user requirements can be handled at a resolution of 10 meters. Referring to figure 5, it can be seen that roughly half of farm sales are from farms of less than 500 acres. And yet, from figure 6, it can be seen that a resolution equivalent to ERTS (about 60 meters) would give an error in field size determination of 20 percent; a resolution of 10 meters will give an error less than 5 percent. This is another argument in favor of 10-20 meters resolution, since an error in as crude a measure as field size of 20 percent would certainly be of little utility.

Therefore, although the requirement of 10-20 meter resolution creates serious problems in the design of the earth resources ground data handling facility, it appears justifiable on the basis of user requirements.

We have used stated user resolution requirements in the literature as well as resolution requirements for areal accuracy to support the requirement for 10-20 meters resolution. These considerations are not of equal importance. A few words concerning the meaning of the user resolution requirements will indicate that there are uncertainties in them that cannot easily be removed. However, these uncertainties, if removed, would tend to produce even more stringent system requirements, so the argument that the user wishes at least 10-20 meters resolution is still valid. The uncertainties in the user resolution requirements make the areal accuracy argument relatively more important.

The principal uncertainty in curves involving resolution such as figure 2 lies in the definition of the term "resolution" and the fact that not all users queried respond with the same definition in mind. Resolution refers to the linear dimension of one picture element (pixel) at ground level (a pixel is equal to the angular Instantaneous Field of View (IFOV) multiplied by the range). Resolution is an ambiguous performance measure unless measurement parameters such as contrast, phase angle, etc., are clearly defined. In system studies one assumes that the hardware designer will insure that under all anticipated measurement conditions the required resolution will be equaled or exceeded. Even with these caveats the term resolution has had no consistent meaning in earth resources studies performed to date. Different meanings in common use are: 1) the photographic line pair, 2) the pixel, 3) the pixel pair, and 4) the pixel pair corrected for line scanning. This latter definition for TV systems is essentially synonymous with the photographic line pair. For a given value of resolution, each of these definitions places a different requirement on the system. In fact, the last definition requires system performance roughly three times better than the second definition. Generally, the line pair (or pixel pair corrected for line scanning) has been the traditional criterion of the "Principal Investigator" because a contrast change in the imaged field requires at least two adjacent pixels to be perceived. For the type of automated multispectral analysis being considered here, however, processing will probably be performed at the pixel level and, consequently, pixel resolution is both meaningful and convenient. Wherever the term "resolution" is used in this report, therefore, it will refer to pixel resolution.

SYSTEM ALTERNATIVES

A wide variety of system alternatives must be considered for an operational data handling system, ranging from very rudimentary manual processing systems to sophisticated and virtually completely automated systems. The feasibility of each approach depends on the projected operational data flow and user needs. This section of the report will present and discuss a number of these alternative data handling concepts. It is felt that the alternatives considered here probably encompass the likely range of data handling systems.

System Elements

The various system alternatives can be synthesized from a standard set of modules or building blocks. Since the functions to be performed in the various systems are essentially similar, the building blocks change little among alternatives. The principal differences are due to the amount of processing actually done, the function of this processing, and the location of the processing entities. This latter consideration, centralized or decentralized processing, is one of the dominant questions that must be addressed in the overall design of the data handling system.

Before discussing the data handling system concepts, each of the building blocks used will be discussed briefly below. Most items are also discussed in more detail in appropriate sections of the report.

Spacecraft. - The spacecraft consists of the sensors, onboard data processing, communications, and necessary ancillary systems such as attitude control. Of three conceivable levels of spacecraft sophistication, the first would be essentially a collection system with no onboard processing beyond straightforward data formatting; raw data would be collected and retransmitted in essentially unchanged form to a ground station or a relay satellite. The second level would process onboard to minimize data redundancy or eliminate useless data such as cloud cover regions. The third and most ambitious level would essentially process

completely onboard up to and including such sophisticated functions as crop recognition, insect damage assessment, etc.

RF Data Link. - This is the communications link from the spacecraft to the ground reception facility, either direct or via a data relay satellite. Since this paper assumes that data will be processed only for the continental United States, a central location such as Sioux Falls, South Dakota, could collect all the U.S. data produced by the satellite. Stations outside, or at several points within the United States, require high speed ground data links to the processing facility; and, in the case of out of country stations, onboard recording capability to retain substantial volumes of data. A dramatically different alternative would transmit information from the satellite directly to the user like the automatic picture-taking system (APT) of the early meteorological satellites, or the rebroadcast mode of the SMS/GOES System. As we shall see, a data relay satellite appears to be the most likely alternative in view of the extremely high data rates to be handled.

Aircraft Platform. - Certain ancillary measurements may be required from aircraft underflying the satellite acquisition platform. The aircraft platform might even be the prime gathering mechanism. The timeliness constraints on most user needs make real time processing on an aircraft platform unnecessary; rather more likely, the data would be collected and delivered as hard copy (photographs and tape recordings) to the data processing station(s). In the multistage approach to be discussed later, the aircraft plays an essential system role.

Ground Truth Acquisition. - For most automatic processing schemes, some form of ground truth data must be provided as a function of geography and time. It is assumed that generally ground truth is a low data rate function that can be adequately handled over existing telephone lines to the processing facilities, although in such uses as a multistage sampling plan, it is possible that these data might be relayed through a data relay satellite via a remote terminal.

Preprocessing. - Preprocessing relates the collected data to the continental United States map base; geometrical corrections and platform

attitude corrections to the raw data are required before map matching. Preprocessing here includes all geometrical and look angle corrections and also some ancillary types of correction like photometric correction.

Processing. - Processing includes converting preprocessed information into material directly useful to the ultimate user. For example, in agriculture, it includes crop recognition and insect infestation detection.

Commercial Processing. - Commercial processing would serve the same function but would be performed by entrepreneurs for one or many user communities; many commercial processors might be active in converting raw or preprocessed earth resources data into suitable user products.

Remote Terminals. - Remote terminals would permit the user to obtain his information by interacting with a data base remote to his location. These terminals would access any or all of the system data and would use standard computational routines available in some central facility to process the user's data, including provisions for formatting directly useful output information.

Users. - Users include federal, state, and local government activities with a direct and legitimate need for earth resources data and the multiplicity of commercial interests that can use such data for more effective operation of their own or clients' businesses.

System Concepts

As mentioned above, a wide variety of system concepts can be postulated, primarily depending on how much processing is done before the data are delivered to the user, and whether the system is centralized or decentralized.

The principal functions performed on the data are acquisition, preprocessing, processing, use, and transmission; all but the latter are shown in the accompanying diagrams for each concept. The functions indicated may take place onboard the spacecraft, at a central facility,

at a user facility, or at a commercial processor facility. Although a large number of combinations are possible, only eight of the more interesting future discrete possibilities are examined. The concepts are diagrammed in figures 7 through 10.

Concept A. - In Concept A, data are acquired at the satellite, transmitted to a central facility where they are preprocessed and processed, and from there they are transmitted to the user facilities where they are analyzed. In this concept, the data are disseminated directly to the users, probably by hard copy according to some prearranged scheme.

Concept B. - In Concept B, all of the work related to map matching is done at the central preprocessing facility, the data are disseminated directly to users who do their own processing, either manually or with local computers.

Concept C. - This concept is similar to B above, except the preprocessed data are delivered directly to users and also to commercial processors who serve an established clientele with predefined processing services. Generally, one would expect that fairly sophisticated users would obtain the data directly from the preprocessing facility and users without the capability for sophisticated data processing would purchase the services of the commercial processor.

Concept D. - In Concept D, the system is the same as in Concept E except that the data would not be disseminated directly to the user according to some *a priori* schedule of requirements, but rather the user would have access to the raw data base through his own remote terminal. This approach limits the data output to actual demand, since the user must take definite action to receive data desired, rather than automatically receiving it on some distribution basis. There would be a requirement for a large scale accessible memory in this concept. The concept is somewhat analogous to the TELOPS concept (ref. 4) recently proposed for scientific data acquisition at Goddard Space Flight Center. The TELOPS concept calls for one year mass storage of scientific data, but the amount of data to be stored would be much greater for earth resources data, so that the length of storage would have to be decreased or a larger memory employed.

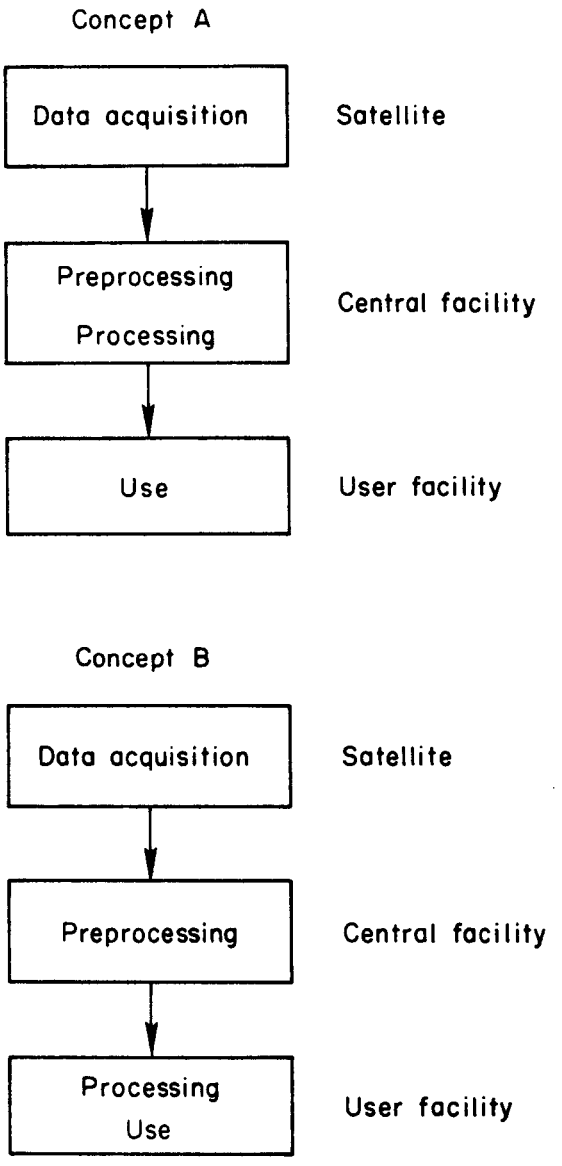

Figure 7. - Advanced Data Handling Concepts - Concepts A & B.

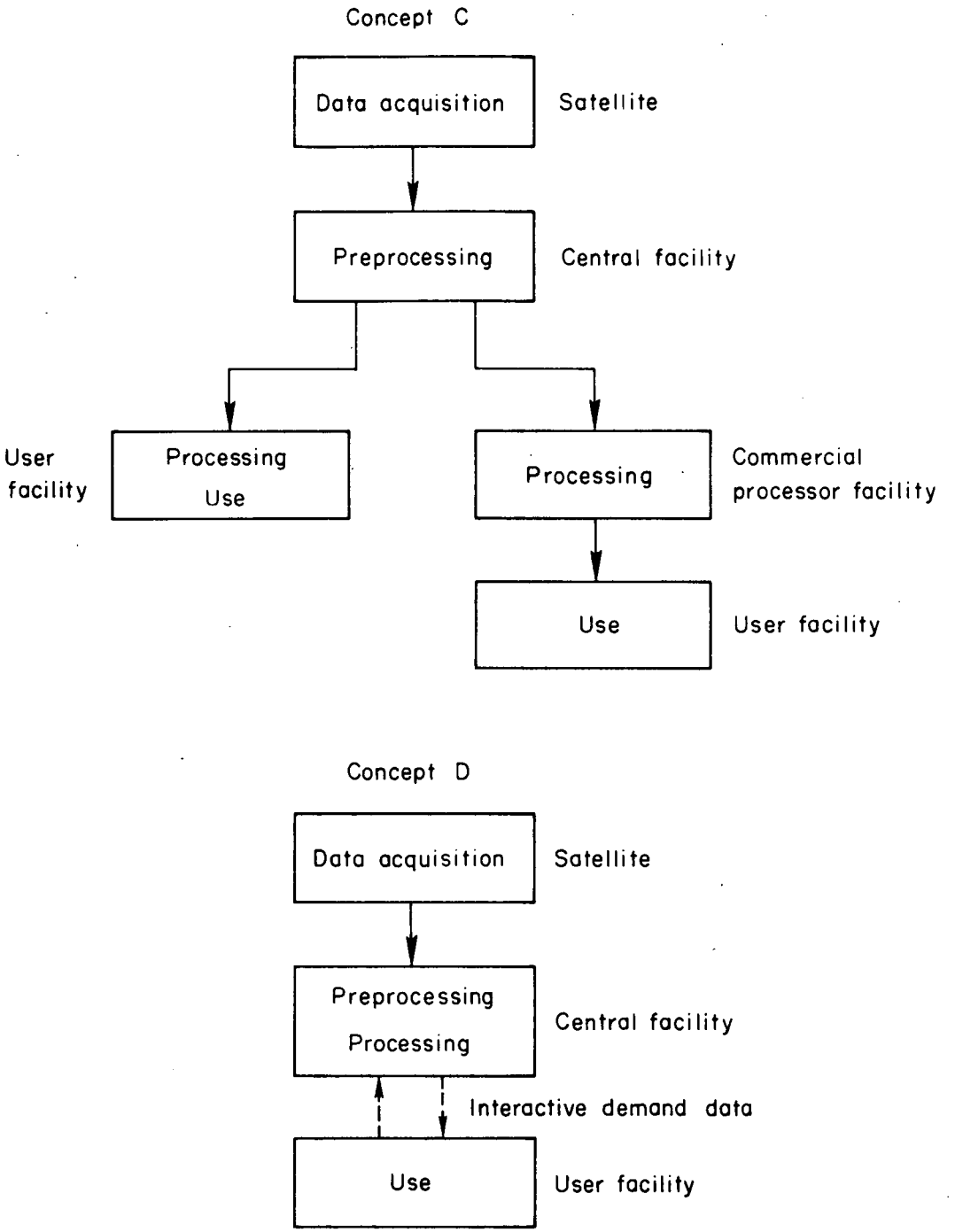

Figure 8. - Advanced Data Handling Concepts - Concepts C & D.

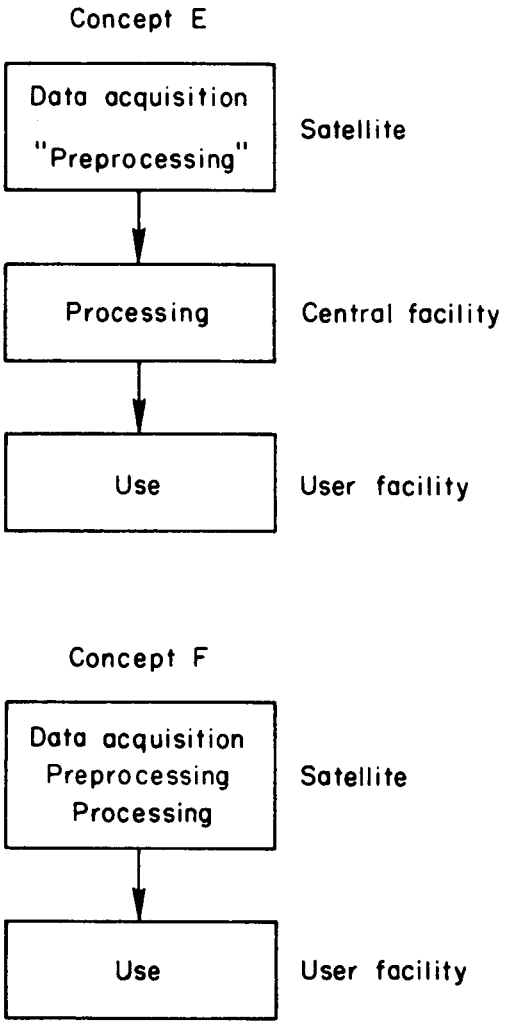

Figure 9. - Advanced Data Handling Concepts - Concepts E & F.

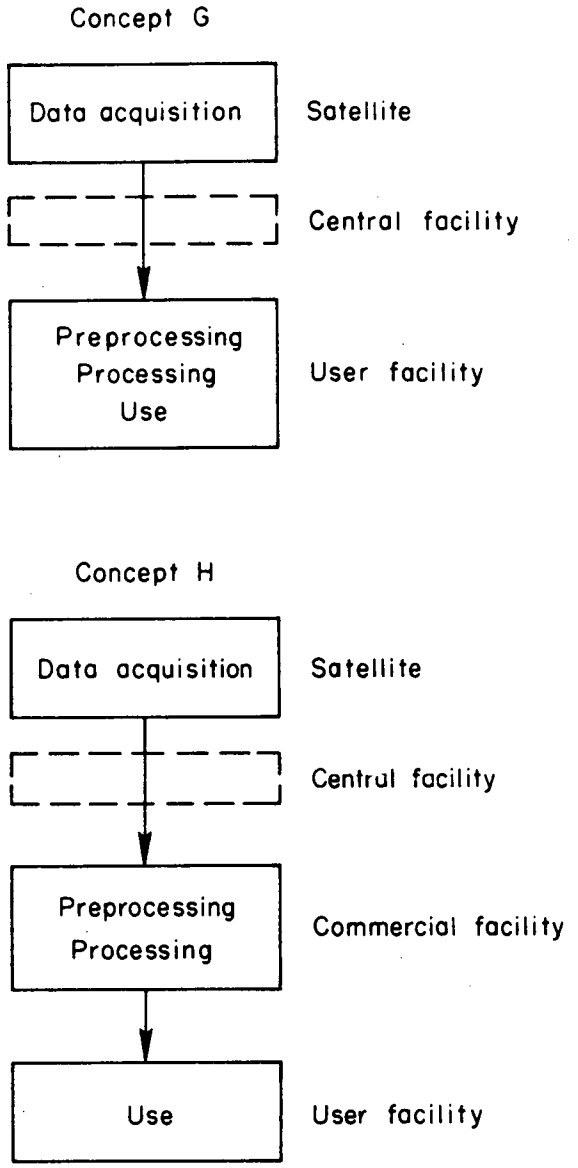

Figure 10. - Advanced Data Handling Concepts - Concepts G & H.

Concept E. - In Concept E, the processing is still done in a central facility but certain types of preprocessing may be done in the satellite. Such a concept would be likely to evolve only in the distant future and would possibly involve optical preprocessing that would result in a decrease in the total amount of data transmitted.

Concept F. - Like Concept E, Concept F involves performing central functions onboard the vehicle. Here certain processing functions as well as preprocessing functions occur onboard that would reduce the data flow to the ground. Both concepts, E and F, should probably be considered jointly since the data compression functions that would be performed onboard would likely be a mixture of preprocessing and processing. For example, data in a particular area might be preprocessed sufficiently to permit processing and then processed sufficiently using interclass divergence techniques to determine an optimal set of channels from those available to reduce the amount of data transmitted to the ground facility.

Concept G. - In Concept G, the raw data acquired by the satellite are transmitted directly to the user without being preprocessed or processed in a central facility. The data might be transmitted directly from the satellite to the user facility or it might be received at a central facility and routed to the user facilities via landlines or as mailed tapes. Mailed tapes are likely to be too slow for many users of an operational system. The landline charges are likely to be very high for a system that routinely retransmits the output of an advanced earth resources satellite directly to all users. A concept such as the SMS/GOES System might be employed, in which data are acquired from the satellite, very crude preprocessing is employed at the receiving facility, and the data are transmitted back to the satellite where it is essentially broadcasted to the decentralized user community.

Concept H. - This approach is essentially the same as Concept G except that the data transmission is either direct to the ultimate user or to a commercial processor who then makes the information available to user clients.

Evolution of Ultimate Concept. - Not all of the systems just discussed are equally likely to evolve. The system that ultimately evolves will depend on detailed user requirements and preferences that will only later become evident and upon the interplay between the technical implications of these requirements and available technology. For example, a system would be highly desirable if data could be stored at the central facility and transmitted only on demand to interactive remote terminals; but the storage requirements for such a system might exceed capabilities, so that such a desirable concept may have to await the advent of cheap bulk storage on the order of 10^{13} to nearly 10^{14} bits, possibly as late as the 1980's. Similarly, it would be desirable to perform sufficient preprocessing and processing onboard the spacecraft to reduce the total data transmission bandwidth to acceptable limits; but preprocessing does not significantly reduce the data load to help in this regard and processing has been shown in this paper to present a critical load even for ground-based computers, so the bulk processing will probably not be done onboard the spacecraft. One approach would be to do sufficient preprocessing to permit limited analysis of the data to determine the optimal subset of multispectral scanner channels to transmit to the ground. This might reduce the bandwidth by a factor of 3X or 4X. It also seems fairly clear that some provision must be made, just as in the WEFAX, or Weather Facsimile system, to supply data to commercial processors who process data for user clients; this too should be a part of any operational system.

Figure 11 represents an interim concept in the 1980-90 era, in which some of these evolutionary trends have had an effect on system design. A limited amount of preprocessing and perhaps such processing steps as interclass divergence testing might be done onboard the satellite. The bulk of the preprocessing would still be done in a central facility. Processing would be provided for those who could use the standard product, and the standard product would still be distributed according to an α *priori* distribution schedule to such terminals, whether commercial processors or users.

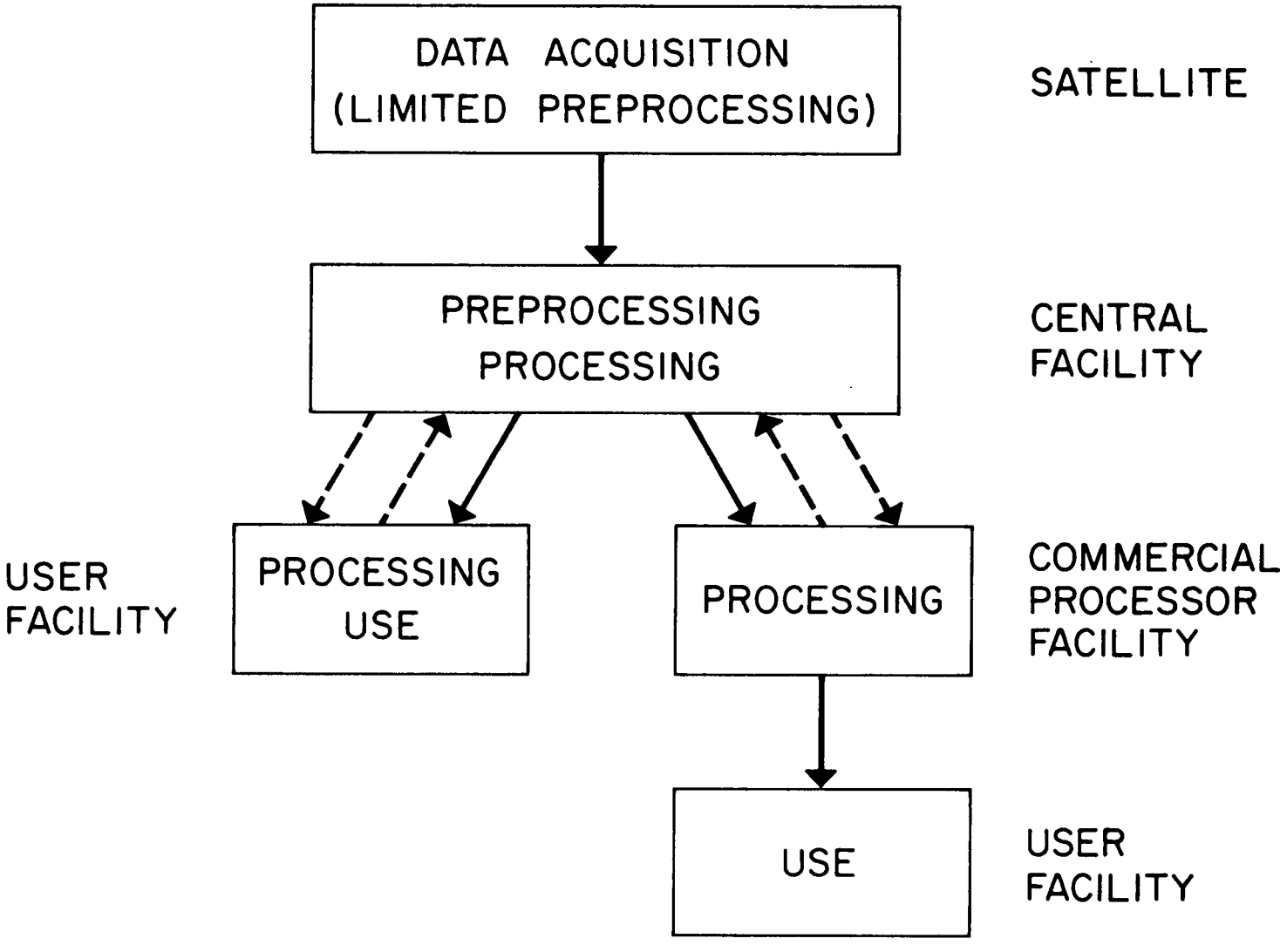

Figure 11. - Advanced Data Handling Concepts - Interim Concept J.

It is conceivable that by the next generation of computer memories it would be possible to have bulk storage capable of storing several months of unprocessed data which could then be processed on request by the user. Alternatively, the data could be processed, which entails something on the order of a factor of 20 reduction in storage in the case of multispectral recognition processing, and then made available to users directly on receipt of a request for data for a specified region. Tapes might still be used, but as in the case of NASA scientific data, the bulk memory version would probably prove more desirable and would probably be cheaper in terms of staffing requirements. Processing the data in anticipation of demand would involve the use of a larger, more advanced computer, since the volume of data to be processed would be larger; but the data would be available for immediate response to user requests and a smaller bulk storage unit would be required. It thus seems likely that the precomputation approach would be used.

A standard series of reports would be sent out, specifying changes that had occurred since the last run, identifying blights and infestations that had been spotted by the processing routine, summarizing areas planted to various specified crops, and so forth. A report generating program would prepare a tailored report for each user based on a standard format.

Each user would be provided with one or more remote terminals from which he would be able to request detailed data on any area which he would perhaps specify with a CRT display and a light pen. He could perform many types of interaction that had been provided for by using his own data base which would be centrally located at the processing facility. He could request hard copy which would then be sent to him by mailed tape. In addition, it would probably be desirable to provide for transmission of digital data which the user would then process on his own computer using special routines pertinent to his special uses that would not be provided at the central facility. The cost of dedicated landlines versus the cost of providing the additional relay capability on the synchronous relay satellite would have to be examined, but the satellite relay does appear promising for this purpose.

System Implementation

Various hardware and software approaches can be used to implement any one of the systems outlined above. While some of these decisions will probably have little effect on overall system cost or effectiveness, those related to machine processing and output format will have a major effect on overall system cost/performance. They have strongly influenced the above account of the probable evolution of the ultimate operational system.

One of the critical choices affecting the system is the type of classification algorithm used for the automatic processing function, e.g., classification of crops, recognition of disease infestation, etc. These classification schemes divide the feature space into some small, predetermined number of pattern categories. Various decision rules can be used, generally those are used that discriminate well and that give the greatest separability among the classes to be identified. Supervised methods operate on the basis of *a priori* information obtained from the training set while the unsupervised approaches require no such advance knowledge. Either deterministic or statistical methods may be used, although statistical methods are far preferable.

In the unsupervised methods for pattern classifications (see ref. 6), no parameter of the training set is evaluated *a priori*, but a systematic error correction scheme from the training set is used to update a set of weighting functions for each new determination.

The relative complexity of the processing algorithms employed has a strong effect on the type of computer required, the degree of centralization of the processing function that will be feasible, and the degree of interaction between the user and the computing facility. Even the simplest useful algorithm may require processing capability beyond that available onboard the spacecraft for many years in the future. A later section of this report will discuss classification algorithms more fully.

The hardware approach also has an important effect on system complexity, cost, and operating time. The three basic hardware implementations are digital, analog, and hybrid. The digital approach can be further subdivided into general purpose or hard-wired. The general-purpose digital computer offers the greatest flexibility, permits the use of new or improved algorithms, large volumes of accessible stored data, and various types of input/output capability. The cost of this flexibility is in terms of computational speed. On the other hand, a hard-wired digital computer can achieve significantly higher throughput than the general purpose machine, but with a correspondingly great loss in flexibility. For the classification algorithms considered in this report, the analog approach appears to have the highest throughput potential, but the training problem appears to be much more difficult than with the digital approach. The third alternative, the hybrid computer, which is a combination of a general-purpose digital machine and an analog computer is a compromise that appears to minimize the training problems while still achieving the basically high throughput of the analog approach.

One of the key problems affecting all of the alternative systems discussed earlier is the nature of the output display format. The data must ultimately be converted to a form that conveys data to the user in readily understandable terms; and the display must not represent a significant bottleneck to processor throughput. In addition to transient displays, there is a requirement for hard copy output, both in color and black-and-white, and perhaps in processed formats using standard printers or other similar machines. All of these alternatives must be considered and related to the user and his ultimate information requirements before any decisions can be made.

Another problem that affects the system design is the potential requirement for substantial buffering of data to stabilize the fluctuating data load produced by the satellite so that the computer may operate at a relatively steady rate. Each day produces a relatively brief pulse of very wideband data stream from the satellite. Since the computing

center will be operating near the state-of-the-art in computers, it will be infeasible to process this data as it is acquired; it must be stored so that the computer can process it in the interim period between data peaks.

The processing rate may be reduced by utilizing such a buffer to store a pass of data and then allow the results to be computed during the remainder of the day while the satellite is not producing useful information on the area of interest.

From an orbit with a repetition period of 18 days, the pass time over an area 1,850 km long is 237 seconds. Then 1.85×10^9 elements per day are generated and all in this brief period. If a buffer stores that data, it would then provide data to the processor at the rate of about 1.85×10^4 elements per second. For a digital computer using an algorithm that performs 10,000 operations per pixel for recognition and classification, this figure would correspond to 1.85×10^8 computer operations per second. As we will show later in this paper, this is an exceptionally stringent digital computer requirement, but buffering has at least made it possible to consider a digital computer for this type of operation. As we will also show later, the characteristics of such a buffer are only slightly beyond the present state-of-the-art, and probably well within the 1980 state-of-the-art.

If the satellite generates $.8 \times 10^7$ elements per second, and if we further assume multispectral data in 12 bands with 8 grey levels coding, then the input rate to the buffer would be about $.8 \times 10^9$ bits per second. This is approximately the 1972 state-of-the-art for modulated laser transmission and would require the buffer to store a total of 185×10^9 elements for each day's processing load. This is a fairly large amount of data to store, but could be handled using laser optical memories slightly more capacious than those now on the market. The readout rate from the buffer would be 1.85×10^6 bits per second to the processor, which would not give serious problems.

The following sections of the paper will consider the various alternatives outlined above and show the implications of each.

PREPROCESSING REQUIREMENTS

Preprocessing of earth resources data as considered in this study consists of all processing prior to and in preparation for classification processing. This includes correction for instrument and measurement errors and any other preprocessing needed to suitably format the output for some process which handles the data as a whole. That is, the "raw sensor data" must be converted to computer readable data. Preprocessing requirements include the correction of errors caused by geometric distortions, radiometric distortions, scan linearity, and spacecraft dynamics. Other preprocessing requirements may include registration and scaling, correction for illumination or cloud conditions, annotation, enhancement, dropout compensation, and reseau removal.

Geometric errors result from spacecraft platform instabilities which cause errors due to translation, rotation, and skew. Since the multispectral scanner image is built by the scanner generating a line at a time, any spacecraft roll, pitch, or yaw will cause map distortion. Figure 12 illustrates the effects of pitch, roll, or yaw on the output image. These error rates may be measured onboard the spacecraft and transmitted with other sensor data for use in computing the image correction transformations. Other geometric errors may be caused by earth curvature, earth rotation, and variations in spacecraft altitude. In addition, geometric corrections are necessary because of differences in the geometry of the image and conventional map coordinates. Orbital ephemeris data, ground control points, or a combination of the two provide the information needed to define sensor location and orientation.

Radiometric distortions arise because of sensor non-linearities. Radiometric errors are also due to differences between the multiple detectors in each of the spectral bands. The correction for this type

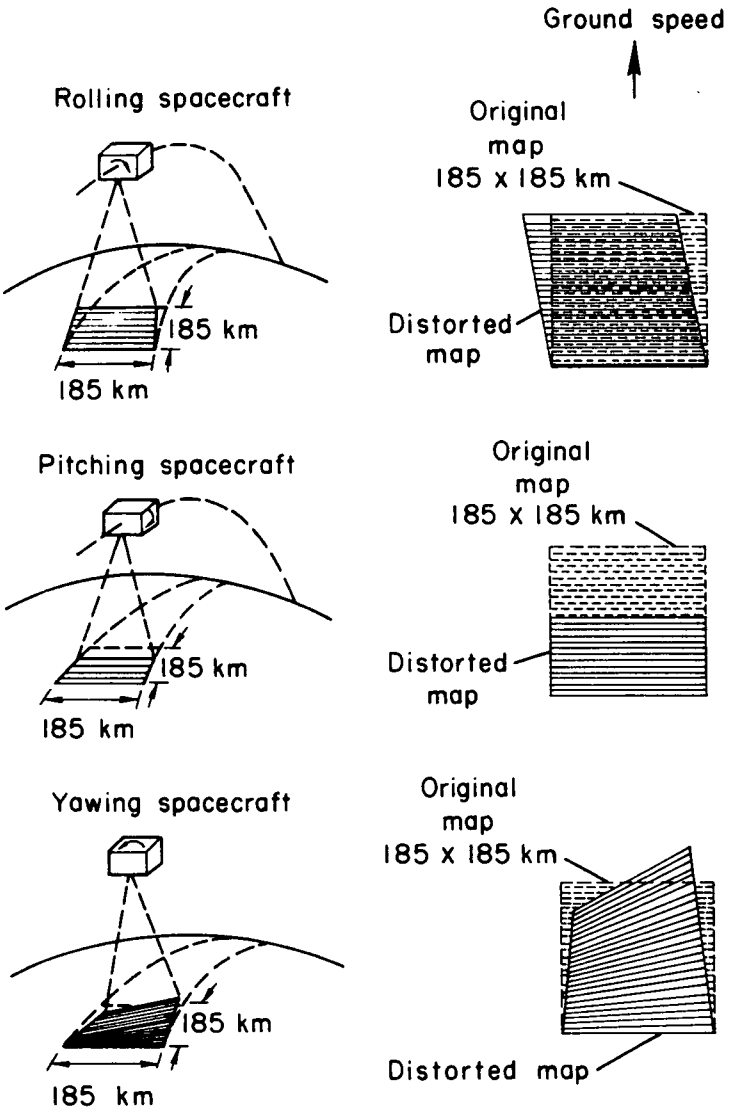

Figure 12. - Effect of Roll, Pitch, and Yaw on Output Image.

of error is provided by calibration data. In addition, correction must be made for scan non-linearity; the correction is the same for each of the spectral bands.

Dropout errors caused by the absence of legitimate data may be detected by testing for unexpected intensity values and corrected by generating values which are the average of neighboring resolution elements. Reseau removal consists of replacing the picture elements which compose the reseau pattern with an average of adjacent pixels. Annotation is needed to identify each image and give the user all data necessary for correct interpretation of the image. The annotation includes registration marks, tick marks, map coordinates, date and time, and sun azimuth and elevation. Annotations of spacecraft orbit number, altitude, and heading may also be provided.

Preprocessing Concepts

Preprocessing can be accomplished using digital, optical, analog, or hybrid (digital/analog) techniques. In the digital method, a large scale digital computer capable of performing geometric and radiometric manipulation of digitized image points would be used. With automatic reseau measurement and ground control point matching, an image transformation may be computed to any degree of accuracy necessary to resolve the geometric errors. The computer then uses the transformation equations to determine the input location on the desired map image. By locating the four input image samples which surround the desired point, the proper output density for the desired position can be computed by linear interpolation of the surrounding image points. The radiometric corrections required by sensor calibration data can be made during this last step.

Besides the obvious cost and complexity of such a machine, there are also several specific drawbacks to the approach. First, rapid automatic reseau measurement and ground control point matching is difficult.

The digital computer must be able to correlate or otherwise identify and locate reference reseaus or control points within the digitized image to accomplish this task. Manual intervention would frequently be required when the computer is unable to find the desired point. Such manual intervention, using an interactive display, could dramatically increase total system throughput time. Unfortunately, selecting a proper balance between manual and automatic operations is complicated by the fact that the position, rotation, scale, and skew of the desired image point are not well-known *a priori* which means that an automatic search may not be successful in most cases.

Experiments with optical processing indicate that these techniques have the potential of very high processing speed and throughput capability. In addition, they have resolution and compilation speed advantages. Unfortunately, these methods lack a straightforward means of implementing spatially variant radiometric corrections, and the parallel processing, typical of the optical approach, is fundamentally incompatible with the serial conversion necessary to provide an image digitizing capability. Despite these drawbacks, the approach must be considered very attractive and future developments may make it competitive with present techniques.

Analog methods are also capable of high speed and large throughput and the state-of-the-art is well developed. However, this approach lacks the flexibility and the bookkeeping capability needed for control, execution, and formatting of the large amount of data involved.

Hybrid image processing consists of digital control of a high speed analog process. In this approach, the basic processing of the video data is performed by analog devices, but control of the system is maintained through the use of digital computer techniques. This method combines the accuracy, computational capability, and the ease of data storage of the digital computer with a high throughput capability of analog methods. The digital capability of the control computer is used to compute the transformation equations required to correct the image to the desired

output system. The measurement of reseau and ground control joints can be made automatically using electronic image-scanning and correlation methods. At the same time, radiometric corrections may be added at a video rate and made available for analog to digital conversion. This latter approach (hybrid) is the one used in the NASA Data Processing Facility (NDPF) in support of ERTS A and B.

Since most of the functions ultimately required of a preprocessing facility are now being performed by the NDPF, this facility will be considered as a prototype preprocessing facility for a future operational system. It is expected that future preprocessing requirements will at least include the functions performed by the NDPF. For this reason, the NDPF is described in the next section.

NASA Data Processing Facility (NDPF)

The NDPF is responsible for processing, distributing, and storing all sensor data acquired and relayed by the ERTS spacecraft. In support of this function, the NDPF accepts payload video tapes and associated data derived from telemetry and produces bulk-processed images, precision processed imagery, and digital image data. This image processing is accomplished by three basic subsystems; the bulk, precision, and special processing subsystems.

Bulk processed output provides radiometrically corrected imagery and a limited amount of geometric correction at a high throughput rate and is sufficient for most users. Precision processed output provides a higher degree of geometric correction intended to reduce geographic location error for selected image data. Special processing provides user data in a format compatible with computer processing and is used primarily by researchers performing computer analysis of the data. The functions of the three processing subsystems are summarized below.

Bulk Processing. - The bulk processing subsystem accepts videotaped image data from both the Return Beam Vidicon (RBV) and the Multispectral Scanner (MSS) and produces corrected and annotated 55 mm latent images on 70 mm film. In addition, bulk processing produces a high density digital tape of either digitized RBV or reformatted MSS data for later use in special processing. The subsystem also has the capability to generate registered RBV color composites from the three black-and-white RBV images.

Additional functions of the bulk processing subsystem are to accept satellite housekeeping data for annotation of the film image, to perform geometric corrections using the pointing error data caused by spacecraft instabilities, to correct for internal RBV errors, and to perform radiometric corrections using MSS calibration data. The latent images produced are developed and if useful (i.e., not cloud covered) are enlarged, printed, and distributed to ERTS users.

The implementation of bulk processing consists of input video tape recorders for MSS and RBV data, a high resolution film recorder and a high density digital tape unit, all of which are controlled by and interfaced to a process control computer. This computer calculates the first-order geometric corrections which are applied to the video data simultaneously with image recording. The correction coefficients are stored in the computer and used to position the writing beam in the high resolution film recorder. It should be noted that the image data never enter the control computer, but remain in their original analog form. In addition, preprogrammed corrections for sensor, transmission, and recording errors are made and annotations added to the corrected images.

Precision Processing. - The precision processing subsystem accepts selected scenes produced by bulk processing and produces images on a nine and one-half inch format. Geometric errors are measured via automatic correlation with reseau marks and recognizable ground objects in the image in order to remove geometric distortions down to one-half of a picture element and to perform precision location and scaling of the corrected video relative to map coordinates. The measurement of reseau

and ground control points is performed by the viewer-scanner under the control of a processing computer which calculates the transformation required to geometrically correct the image. The video processor, under computer control, provides the radiometric corrections.

The geometric and radiometric corrections are then applied to the scan shaping and video signal input to the film recorder. The corrected image data are also recorded on high speed digital tape for later conversion to computer-compatible tapes. The output precision processed film is annotated and provides a very precise map image of the scene.

The use of ground control points to correct positional errors in MSS and RBV images is very important to the accuracy in positioning ERTS images with respect to the earth's surface. These ground control points have been precisely determined by the United States Geological Survey and stored in the computer program. The computer drives an optical scanner to search for the control point on the 70 mm film image. This is repeated for from six to nine of the control points on that frame. From these measurements, the control computer calculates the appropriate correction coefficients for the geometric transformation. The resulting correction is much more accurate than that which could be achieved using satellite data alone.

Special Processing. - Special processing provides the function of converting digitized image data to computer-compatible tape. These tapes are made available to investigators who plan to use additional computer processing for research purposes. The image data are read from high density digital tape, reformatted and edited into selected sub-frames. The image data are then corrected and written onto computer-compatible tape in industry standard format.

NDPF Throughput. - The present NDPF system is sized for a throughput of about 118 scenes per day from both the RBV (3 channels) and the MSS (4 channels). For the seven channels together, then, this amounts to 926 spectral images per day or almost 10,000 bulk film images per week. In addition to this quantity of bulk processed data, 5 percent of the output may be precision processed depending on specific user

requests, and another 5 percent may be processed to computer-compatible tape format.

Future Preprocessing Requirements

It is probable that an advanced data handling system would use a hybrid system for preprocessing presuming that the same techniques that are used in the NDPF are applicable for advanced systems, and that the preprocessing and classification processing would be separated in function. The classification processing, in this situation, would presumably evolve toward the use of a digital system as user requirements become more clearly defined. On the other hand, in the case of a very large digital system used for classification processing, the preprocessing requirements may be so small in comparison that this function may be easily absorbed by the large digital system. It should be noted in this connection that the requirements for preprocessing in an advanced system are certainly not as severe as that for classification processing because of the more sophisticated and more diversified types of processing for many types of output and variety of user.

Assuming the use of hybrid system for data preprocessing, similar to the present NDPF system, it is possible for the control computer to process any number of images per scene by handling all images for each scene in parallel analog channels. The control computer can easily apply geometric corrections for distortion, vehicle motion, and map projection to each analog channel since the scanner uses common optics for each spectral band and corrections only have to be computed once. That is, the same correction is applied to each channel. Of course, because of the higher resolution required, more ground control points will be needed for higher order correction equations, but the increase in computer capability required can easily be satisfied by scaling up the control computer from the present NDPF system which uses an XDS Sigma 3.

Whereas the requirement for scaling up of the digital part of the preprocessing system to an advanced system is not severe, scaling up the analog part of the system is more likely to be severe. For example, scaling from 100 m to 10 m resolution for the same film image throughput rate would require writing with film recorders 100 times the bit rate and/or analog bandwidth, i.e., ten times the bandwidth in each scan and ten times the scan rate. If this amount of speed increase is not feasible mechanically or electronically, it may be possible to add parallel units.

The previous discussion assumes that there will in fact be film output from the preprocessing system, but it is not clear whether or not film output for preprocessed data is required at all. (Certainly, film output will be required after the data has had additional processing, but at such a point the amount of data will have been reduced in magnitude.) However, there may be a requirement for the retention of preprocessed output in film form for archival storage rather than in the form of digital data in mass storage. The use of film storage of preprocessed data may greatly reduce the requirement for mass digital data storage for more than a nominal storage time of, say, one week to one month since if it is necessary to reprocess the data using a different classification algorithm, digital data may be easily regenerated from the film record. In some cases, it may be more economical to regenerate the desired data from film by means of a film reader than to pay the storage costs of digital data. In any event, it is envisioned that whether or not there is film output, there is a necessity for direct analog to digital conversion and transfer to a working storage system that has common access with the classification processor.

Advanced Preprocessing System. - A concept for an advanced preprocessing system is shown in figure 13. Real time or recorded MSS data are demultiplexed and enter the system in multiple analog channels under control of the digital processor. In addition, ephemeris, housekeeping, and calibration data enter the control computer directly. These data, together with the ground point data stored in the computer, are

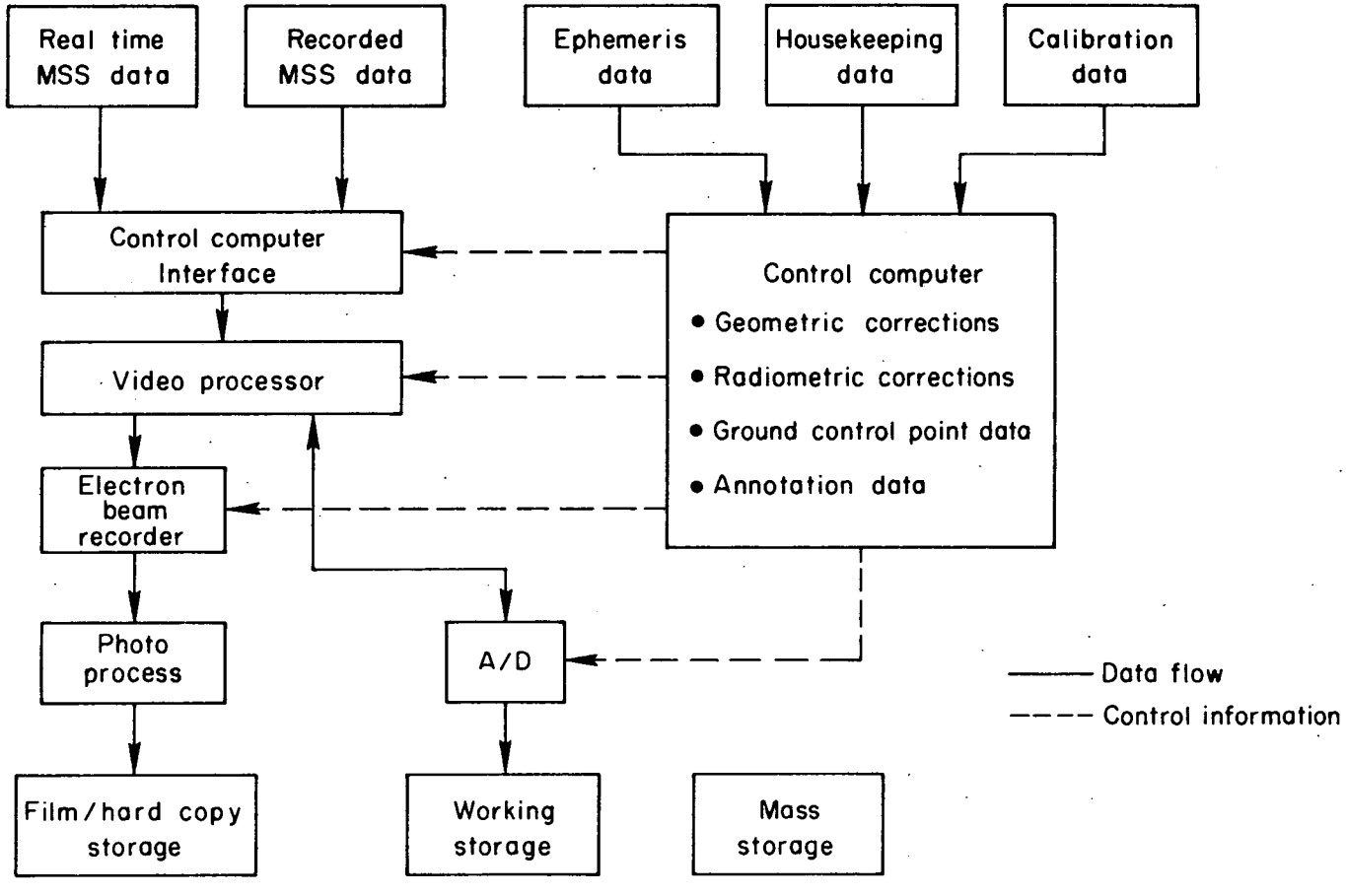

Figure 13. - Advanced Preprocessing System.

used by the control computer to perform digital calculations for the geometric transformation equations. The geometric corrections, together with the radiometric corrections, are then used to control the video processor scan and sweep whose output drive the electron beam recorder.

Annotation data supplied by the control computer are also applied to the film recorder and the output is photo-processed and stored as hard copy film image data. Simultaneously, output from the video processor may be converted from analog to digital data, cataloged, and placed in a working and/or mass storage medium which is accessible by the classification processor.

PROCESSING REQUIREMENTS

In this paper, processing of earth resources data consists of all operations for classifying the preprocessed data. For example, data obtained over cropland would be classified into specific crop types and, depending on the sophistication of the system, crop vigor, status, disease infestation, etc., would also be determined. The methods used to accomplish this task range from conventional manual photo-interpretation to completely automated pattern recognition systems using computers. This section of the paper will concentrate on these latter methods.

Automatic pattern recognition is still a largely unsolved problem in earth resources and yet it is essential to employ it since manual methods will be inadequate to handle the future information flow rates postulated in this paper. Most authors have relegated such methods for automatically analyzing real time data flow to the far future. However, there has been sufficient recent success with automatic pattern recognition to make the use of pattern recognition in earth resources appear encouraging.

This section will focus on these new methods, discuss the application of pattern recognition to earth resources problems, outline several of the promising algorithms, make some preliminary remarks and comparisons

of the methods regarding their potential use in future operational systems, and discuss ways that the various techniques may be combined.

Application of Pattern Recognition to the Earth Resources Problem

The basic problem for earth resources pattern recognition is to convert a large quantity of multispectral scanner data into a recognition map in which each picture element is identified as a specific substance, such as corn, bare soil, blighted wheat, insect infested tobacco, etc.

The basic idea is to collect a set of information that describes the objects to be identified in various spectral bands or features and to process this information to provide a recognition map in which the various objects have been identified and labeled. The raw multispectral information makes very little sense until it has been processed. Processing makes use of the fact that each object will have a different "signature," or characteristic set of measurements in the various features that uniquely determine it.

Figure 14 is an example of a multispectral view of a residential area. In this example, there are 100 picture elements, each containing three numbers, one for each of three spectral bands or "colors." Typically, from three to twenty-four bands may be used. Each number signifies the brightness of the resolution element as viewed through a particular window in the spectrum, and in this case brightness has been placed on a scale of 0-9. In this representation, the spectral signature of a known material would appear as a 3-vector and the spectral signature of any unknown material could be compared with it, vector components by vector component to determine whether they were similar. The uniqueness of spectral signatures is of key significance.

Figure 15 gives a spectral signature for a "substance X." We assume that a resolution element contains nothing but substance X and that a measurement has been taken at each band of wavelengths, using

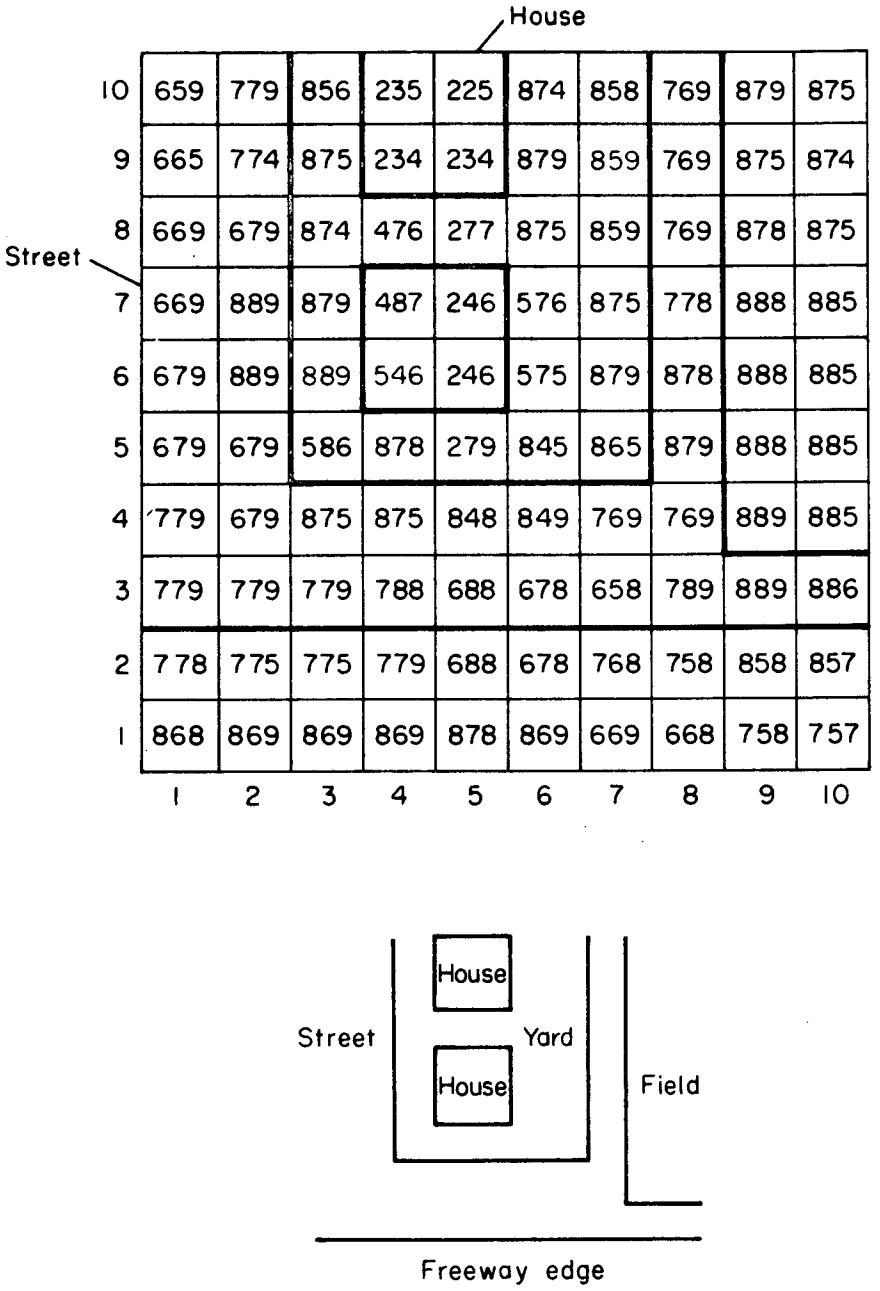

Figure 14. - Multispectral View of a Residential Area.

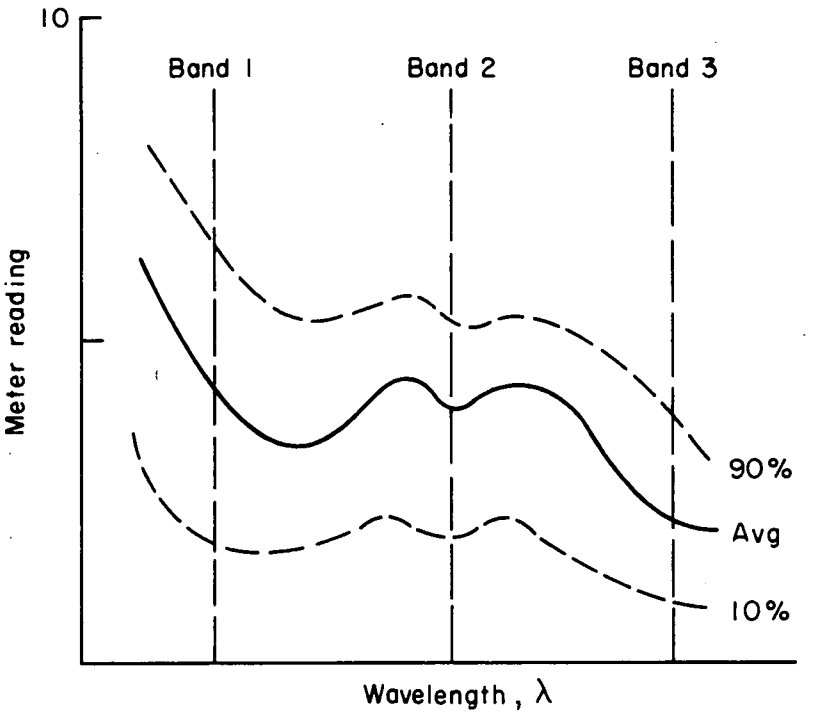

Figure 15. - Hypothetical Spectral Signature.

some appropriate instrument which measures the response at each given band of wavelengths. If many such resolution elements, each containing the substance X, were scanned, the results would not be identical each time, owing to local condition variations, illumination angle, etc., but broad "confidence" bands could be drawn.

The three readings at different wavelengths can be interpreted as a three-vector which also is a signature of substance X but in a different space and may also be plotted as in figure 16. In this space, the random variation due to scanning large numbers of separate instances of substance X produces an ellipsoidal probability density; for N, more than three wavelength bands, the space would be an N-space and the probability density would be a hyper-ellipsoidal one.

Pattern recognition methods assume that each type of material scanned has a distinctive spectral signature. Sets of possible ground truth signatures may be compared with the signature of the unknown substance, by automatic comparison techniques, to identify the unknown substance.

The general method described above could be called spatial/spectral pattern recognition. That is, the recognition is performed in near real time using current ground truth. Another alternative that will be discussed in more detail later uses temporal data in addition to spatial/spectral information.

Regardless of the specific approach used, the basic problem is to take multispectral data from a single set of measurements or a series of measurement sets over time together with *a priori* knowledge (ground truth) to classify unknown signatures.

In the material that follows, the various techniques under development to accomplish this task will be discussed as well as the problems and advantages inherent in each approach.

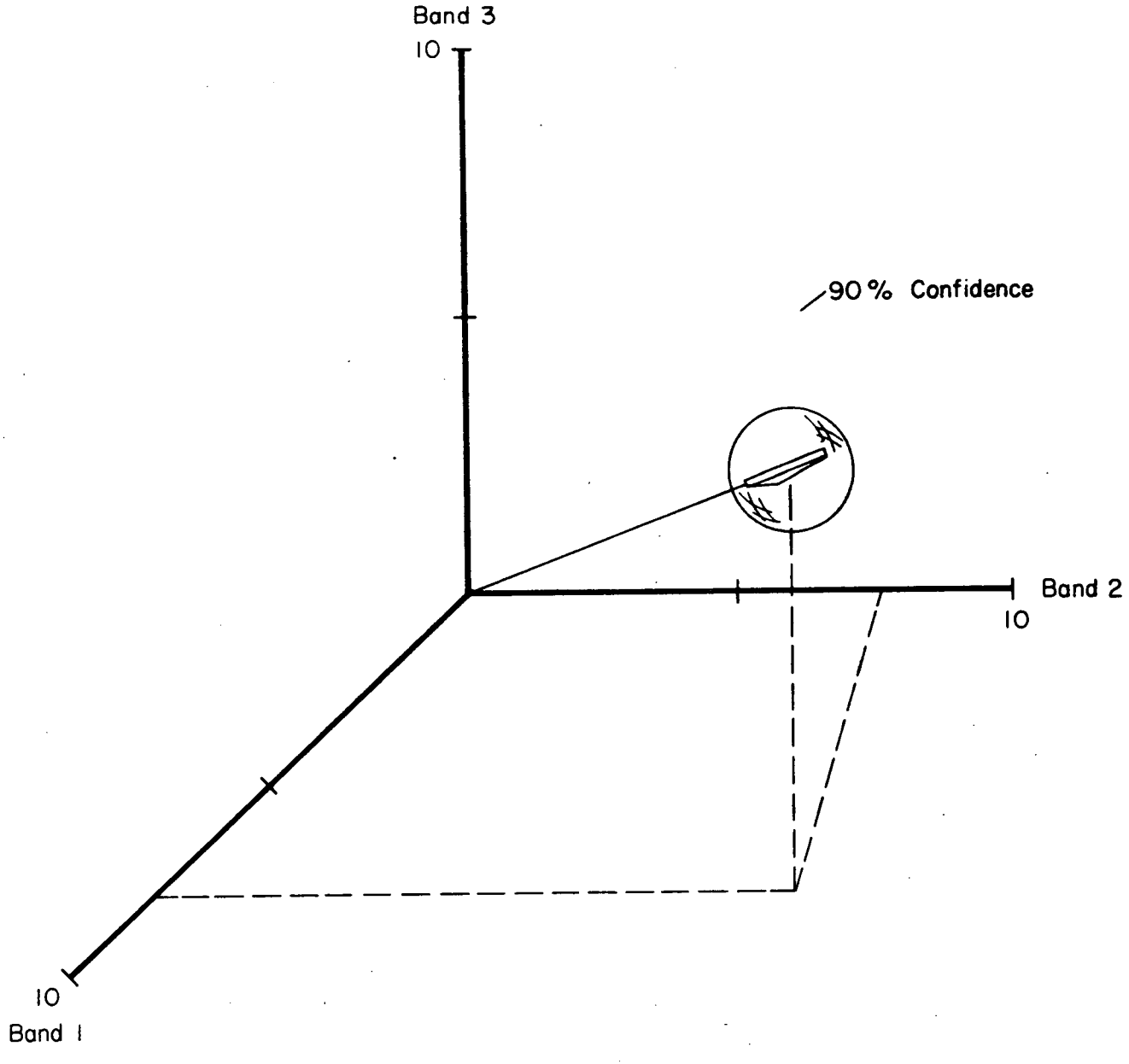

Figure 16. - Hypothetical Spectral Signature in 3-Dimensional Feature Space.

Clustering Algorithm

The clustering approach takes the data points and develops proximity relationships between them in multispectral space that can be used to form natural clusters of similar objects. Once these clusters are formed, the cluster may be identified as a single substance and the results may be mapped back into Euclidean space as a recognition map.

Figure 17 shows the measurement vectors of figure 14 in two dimensions and rounded to the nearest digit on a scale of ten for simplicity. The street, for example, cannot be characterized by a single vector but consists of vectors (6,6), (6,7), (7,6), (7,7), and (8,7). Thus, the locus of the vectors that describe "street" represent something like an ellipse in two-dimensional measurement space. Similarly, "field" has points in (8,7) and (8,8), and "yard" has points in (8,5), (8,6), and (8,7). Using this for ground truth, it would be possible immediately to identify an unknown point (6,6) as part of a street, a point (8,8) as part of a field, and a point (8,5) as a yard. But so far it would not be possible to classify an unknown measurement of (8,7) unambiguously into one of these three classes since it falls in an overlap of several classes. Additional dimensions might resolve the ambiguity. A clustering algorithm (figure 18) would arbitrarily resolve the ambiguity by classifying the unknown into the closest of the three classes the algorithm happened to be building a cluster for at the time the element was scanned.

The method works as follows. A multispectral view such as figure 14 is scanned to determine which readings occur with the greatest probability. The reading (8,7,5) occurs most frequently, probably "grass." Now all neighboring readings such as (8,8,5), (8,7,4), and (7,6,5) are scanned to determine their probability of occurrence. Those with a probability of occurrence above a certain threshold value epsilon are annexed to the cluster and the process is continued until all "nearby" readings above epsilon in probability have been annexed, forming a

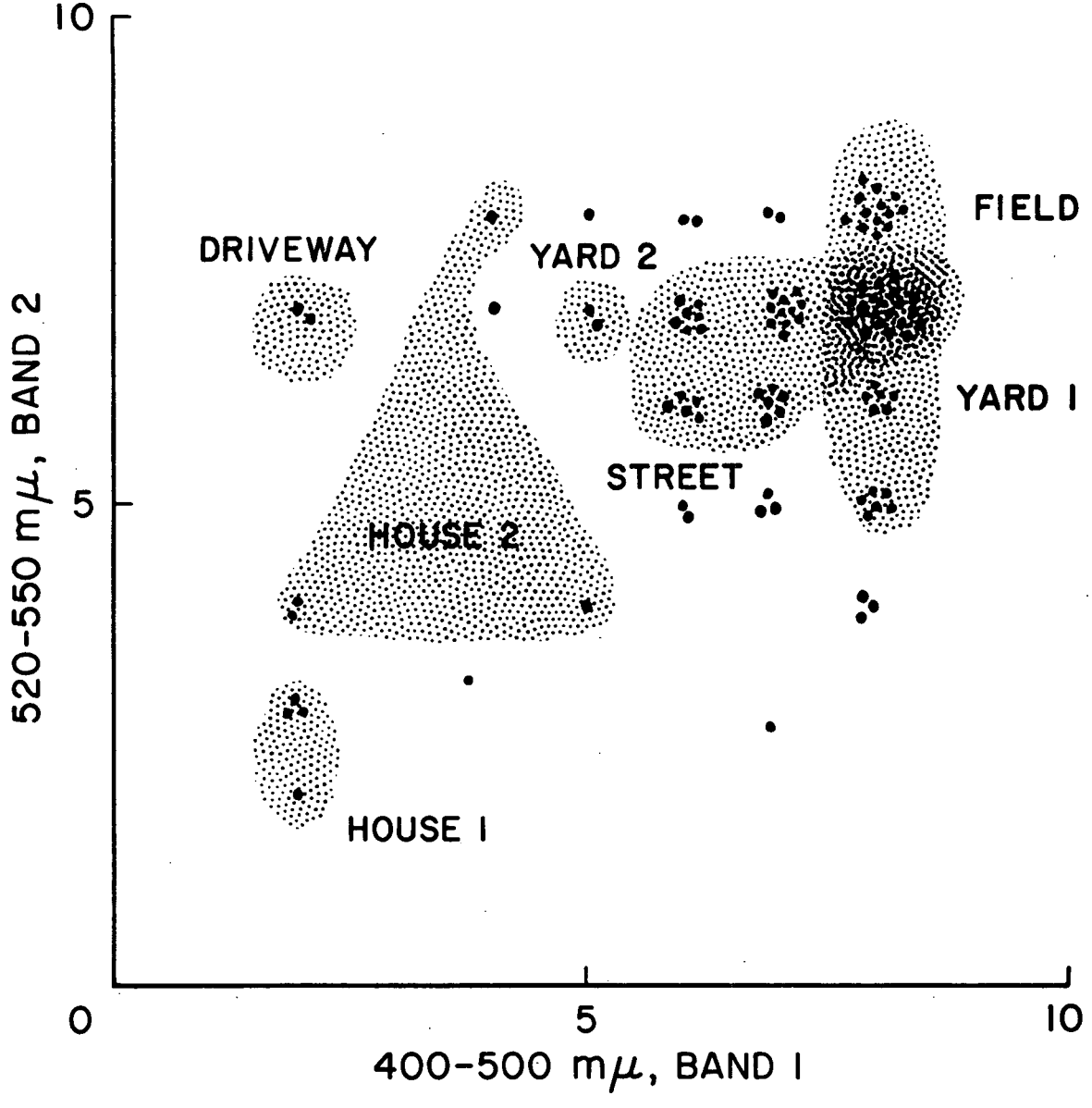

Figure 17. - Clustering of Residential Scene in 2-Dimensional Feature Space.

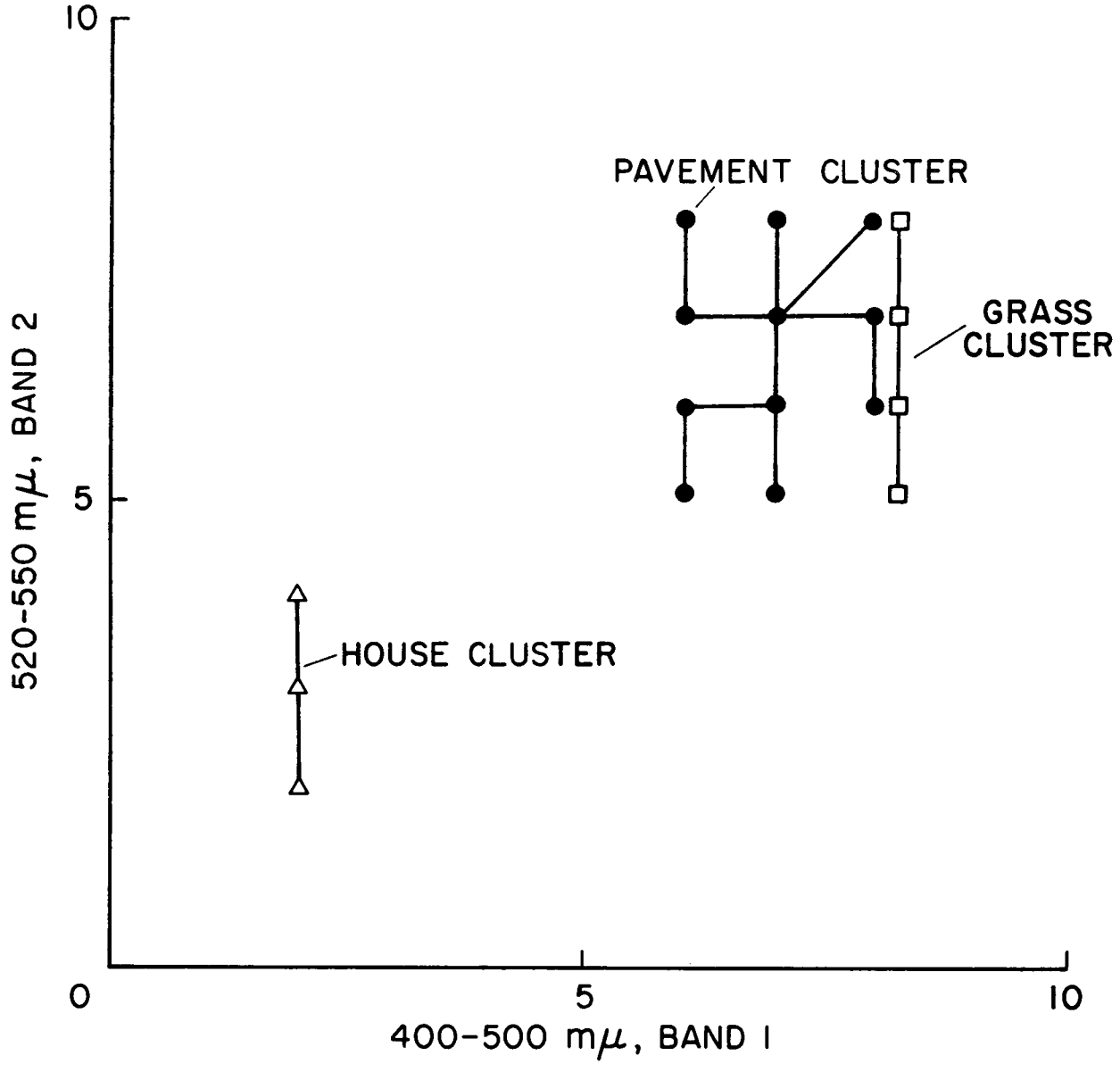

Figure 18. - Clustering Algorithm Results in 2-Dimensional Feature Space.

single complete cluster. A list of probabilities of readings, in descending order, is used in the algorithm.

The next most probable reading not yet assigned to a cluster is (7,7,9) which occurs in 6 percent of the cases. A cluster is progressively built on this initial reading which turns out to be Cluster A, "pavement." Another cluster of "roofs" is built and there remain a number of readings that were missed in the cluster building. For example, (8,5,8) and (8,5,9) are probably members of the "grass" cluster missed because of the peculiarities of the algorithm. Such unidentified elements are generally assigned to the cluster nearest to them.

The clustering algorithms can work not only in feature space but in ordinary Euclidean space so that, for example, the (2,7,7) and (2,7,9) readings that were missed by clustering in feature space would be picked up as "roofs" when the clustering was completed in Euclidean space, because they are spatially adjacent to the "roof" cluster.

The computational cost of the clustering type of algorithm is heavily dependent on the number of vectors that are to be simultaneously classified. If K N-vectors are to be simultaneously, classified, a preliminary ordering of the vectors requires a time of:

$$(N \log_2 K)\ \alpha \text{ seconds.}$$

An additional reranking of these vectors will cost not more than this same time again making a total time cost for ordering of:

$$(2N \log_2 K)\ \alpha \text{ seconds.}$$

To cluster the vectors will require an additional time of:

$$K!\ N\ (\alpha+\mu) + \mu \text{ seconds,}$$

where α is the computer add or compare time (assumed to be the same) and μ is the computer multiply time. It should be noted that the cost

increases extremely rapidly as the number of simultaneously compared vectors increases, due to the K! term. The total time will be:

$$(2N \log_2 K)\ \alpha + K!\ N\ (\alpha + \mu) + \mu \text{ seconds.}$$

One advantage of the clustering algorithm is that it is not necessary to make assumptions concerning the distribution of readings in feature space--they may be of any shape as long as they are connected. The algorithms are generally more sophisticated than the form just presented; the above being presented merely to illustrate the process. One of the disadvantages is that a number of passes must be taken through the data and a relatively large block of data must be ready for processing in the computer memory at one time for the algorithm to work. If the probability distributions in feature space actually are long, stringy, and complex, a clustering algorithm may be essential, at least as an adjunct to the main method.

Clustering is inherently a slow digital technique that would not be amenable to analog computation. Therefore, it is unlikely to be implemented for real time analysis in the near future until faster digital computers become available. Clustering will appear most likely in combination with the methods to be described below rather than as an independent technique. The importance of clustering is that it can classify materials without using ground truth until after the clustering has been done. This makes it a potential candidate for an unsupervised algorithm.

Likelihood Ratio Algorithm

The next method to be described is based on likelihood ratios. It is more subtle, and also more satisfactory than the method just described.

In the likelihood ratio approach, a multidimensional Gaussian decision rule is used to assign a given measurement into the class that the ratio shows it is "maximally likely" to belong to, considering the

relative densities of the several ground truth probability distributions at the point in measurement space that the measurement occurs. Thus, the measurement space is divided into regions and the unknown measurement is assigned to the classification corresponding to the region in which it occurs. Unlike the clustering method, this method would require a special way of handling ties if they occurred; fortunately ties occur seldom. In general, the more dimensions that are used, the better for recognition performance; but there are practical limits to the number of dimensions used.

In figure 18, a rather complex pavement cluster was identified by the clustering algorithm, which was able to distinguish it from a rather nearby "grass" cluster and a more easily distinguished "roof" cluster. With the parameters used for the illustration, the algorithm was not able to distinguish between "yards" and the large field in the right of the scene. Figure 19 shows how the same situation would be handled by the maximum likelihood method. The ellipses shown on the figure represent the standard deviation of the two-dimensional Gaussian elliptical distribution for "field," "roof," "street," and "yard." The dotted lines divide the space into sectors, so that every unknown reading that falls into the sector for "roofs" is classified as a roof, for example.

The multi-variate Gaussian density function is given by:

$$p(X/W_i) = \frac{1}{(2\pi)^{\frac{1}{2}}} \cdot \frac{1}{|K_i|^{\frac{1}{2}}} \cdot \exp\left\{-\frac{1}{2}(X-M_i)\ K_i^{-1}\ (X-M_i)\right\} \quad i=1,m$$

where $P(X/W_i)$ is the density of the i-th class of material at the point in N-dimensional space represented by the N-vector X; K_i is the covariance matrix for the i-th class of material; M_i is the vector of means of the i-th class of material; and m is the number of classes of material under consideration. Thus, it can be seen that once M_i and K_i are known, the distribution for the i-th material is uniquely specified.

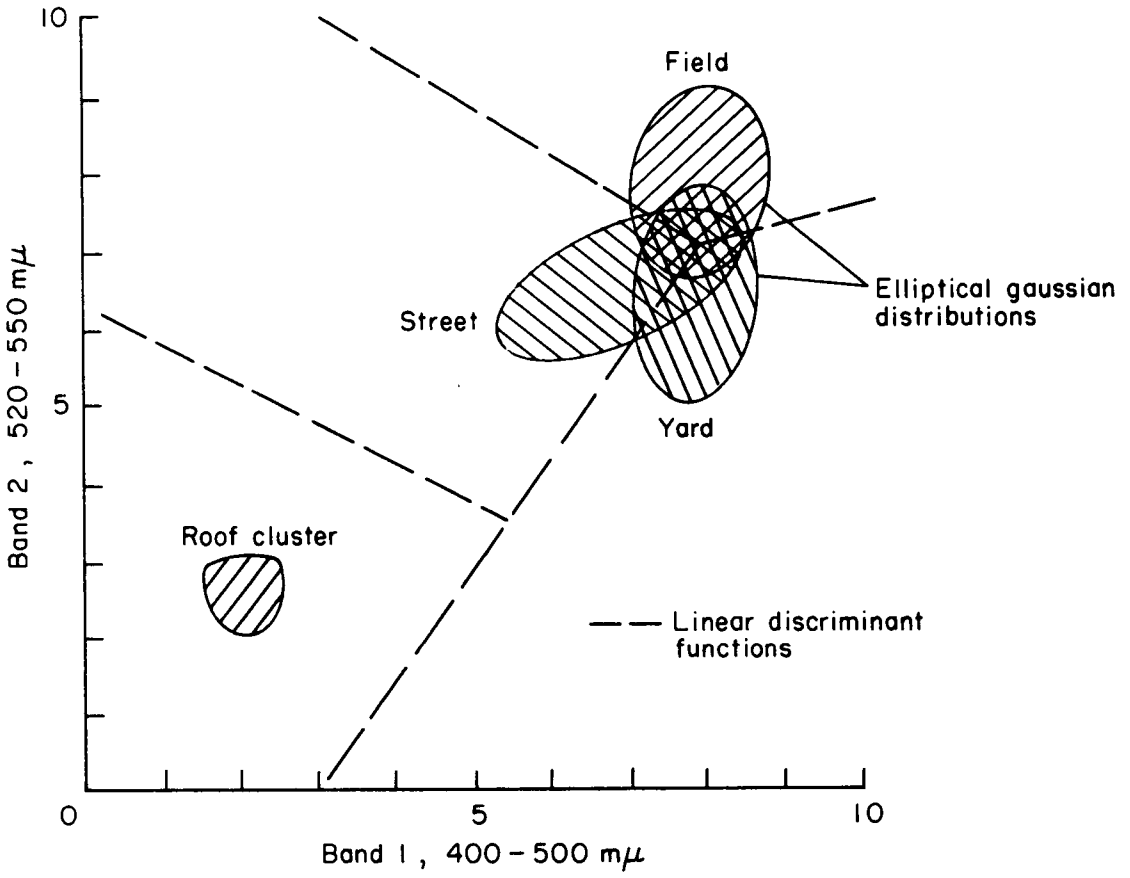

Figure 19. - Maximum Likelihood Ratio Algorithm.

A sample of ground truth data points may be used to estimate M_i and K_i to determine the distribution. The greater the number of sample points available, the smaller will be the variances in the terms of K_i, so that subsequent estimation tasks may be done with less ambiguity.

The process of collecting such ground truth data, estimating the means and covariances, and inserting the resulting values into the computer algorithm so that they may be used to classify unknown data is generally known as "training" the algorithm or the processor. It is an expensive and time consuming process at the present state of development of the algorithms. In the case of an analog processor it requires that a training tape be played through while adjusting the potentiometers until they reflect the covariance matrix of the selected class of the training set, and repeating the process for each class of the training set. This training process will be discussed in more depth in a later section.

One of the simplest decisions to be made once the training set means and covariances are available is whether an unknown material whose measurement space vector is X belongs to a specified class of materials or not. For this, a likelihood ratio criterion is used, that can be explained as follows.

The probability that an unknown point X belongs to material j as opposed to materials 1,2...(j-1), (j+1)...m is:

$$P_j(X) = \frac{f(X;\ M_j,\ K_j)}{\Sigma f(X;\ M_i,\ K_i)}$$

where $f(X;M_iK_i)$ is the density of material i at point X in the multidimensional space. In other words, at any given point in space there is an accumulation of density, some contributed by one material's density function and some by another material's density function; the probability that the point belongs to a specified material is proportionate to the

relative contribution to the total density that is made by the specified material's density function at the given point in space. One method of classifying the point into one material or another is to select the material having the highest probability, or

$$P_j(X) > P_i(X) \text{ all } i \neq j$$

On the other hand, the likelihood that an unknown point belongs to material j rather than i, for $i \neq j$, is given by the probability ratio (likelihood ratio):

$$L_j = \frac{P_j(X)}{P_i(X)}$$

Analog and Hybrid Computation of Likelihood Ratio Algorithm. - The SPARC analog computer at Willow Run Laboratories, University of Michigan, is designed to compute this likelihood ratio and decide that an unknown material with measure space vector X belongs to class j if the likelihood ratio for class j at point X exceeds a certain threshold value k.

In the work at Willow Run, the value of the parameter k is set at k = 1, which is equivalent to assigning the unknown to the class that has at least 50 percent of the density at the point X. This threshold would be rather restrictive for small target areas but it appears to work for typical agricultural areas. Controlling the magnitude of the threshold value k gives some control over the relative proportion of correct detections to false alarms.

In the above discussion we have assumed that it is known that each of the materials i = 1,...m is present in the area. The algorithm used on the SPARC computer actually weights each density according to the *a priori* probability that the given material is in the area of search; this weighting formula is not given here because the processor is usually operated with all of the weights equal, which is again equivalent to our assumptions above.

The procedure for using the SPARC computer is interesting and illustrates several important points concerning the whole problem so it will be briefly described at this point. A 12-channel tape is made using a multispectral scanner. A tape containing a suitable ground truth area is run through the computer. The ground truth area is masked off so that only signatures from the selected material are entering the computer. The computer contains circuits that generate a multivariate Gaussian density value for a given input vector. The input vector X is transformed to a vector Y through a transformation that leaves the components of Y uncorrelated and with a variance of unity. This allows the coefficients corresponding to the variances and covariances to be adjusted into the set of potentiometers by a recursive technique. The areas of the elliptical group of readings are centered on the y_1 and y_2 axes of a scope to represent the probability distribution of y_1 and y_2. Then a potentiometer is adjusted until the probability density axes correspond to the scope's vertical and horizontal axes so that no correlation exists between y_1 and y_2. Next a potentiometer is adjusted until the bivariate density is circular, so that the variance is unity. Thus, three of the potentiometers have been adjusted.

Next, the scope is connected to y_1 and y_3 and the correlation between y_1 and y_3 is nullified; next it is connected to y_2 and y_3 and their correlation is nullified. Next the potentiometer corresponding to $x_1 - X_1$ is adjusted so that the density is circular making the variance equal to unity. This procedure is repeated until all N components have been "stored" in the machine in the set of potentiometers.

Notice that only one potentiometer is set for the first component to adjust the variance to unity, but two settings are required for the second component, three for the third, etc., making a total of $N(N+1)/2$ potentiometers to set for the N components. This is the same as the number of nonredundant terms in the covariance matrix, which is more than coincidental, since the potentiometer settings are in 1:1 correspondence with the covariance terms.

The number of potentiometers which must be set to store a twelve-dimensional ground truth signature is therefore 12 x 13/2 = 78. Even allowing only one minute per adjustment, this process would require over an hour per ground truth setting, or a full working day to set the potentiometers for a reasonably large group of potential targets.

To reduce this problem to tractable form, hybrid computers have been designed that compute the potentiometer settings digitally and set the potentiometers automatically. It should not be necessary to go through the nulling process on a potentiometer-by-potentiometer basis using this approach since the digital computer can compute the settings directly through a matrix transformation.

It might seem reasonable to assume that once the setting for "wheat" has been determined, the potentiometer values could be looked up, set in, and the process of identification could begin. This ideal situation apparently does not obtain. The ground truth must be taken under the same conditions as the unknown readings to be identified, as nearly as possible at the same time, from the same altitude, and even from the same angle. That is, the readings of a ground truth patch at the nadir cannot in general be effectively used to identify unknown materials at a wide angle from the vertical. Some work has been done to theoretically extend the signatures to apply over a much wider area, but the problem is still unsolved. Spacecraft, of course, view large areas with relatively little angle variation, so that this may be less of a restriction for spacecraft than for aircraft.

The economic implications of this situation are that a great deal of expensive ground truth information will be required unless the problem of signature extension is solved. Figure 20 illustrates the situation schematically. If essentially 100 percent recognition can be obtained with a ground truth plot every 1/100th of a square mile, and something like 10 or 20 percent recognition can be obtained with essentially no ground truth, then a decision will have to be made as to how much the investigator is willing to allow recognition performance to deteriorate in order to save the cost of extensive ground truth. To make such a

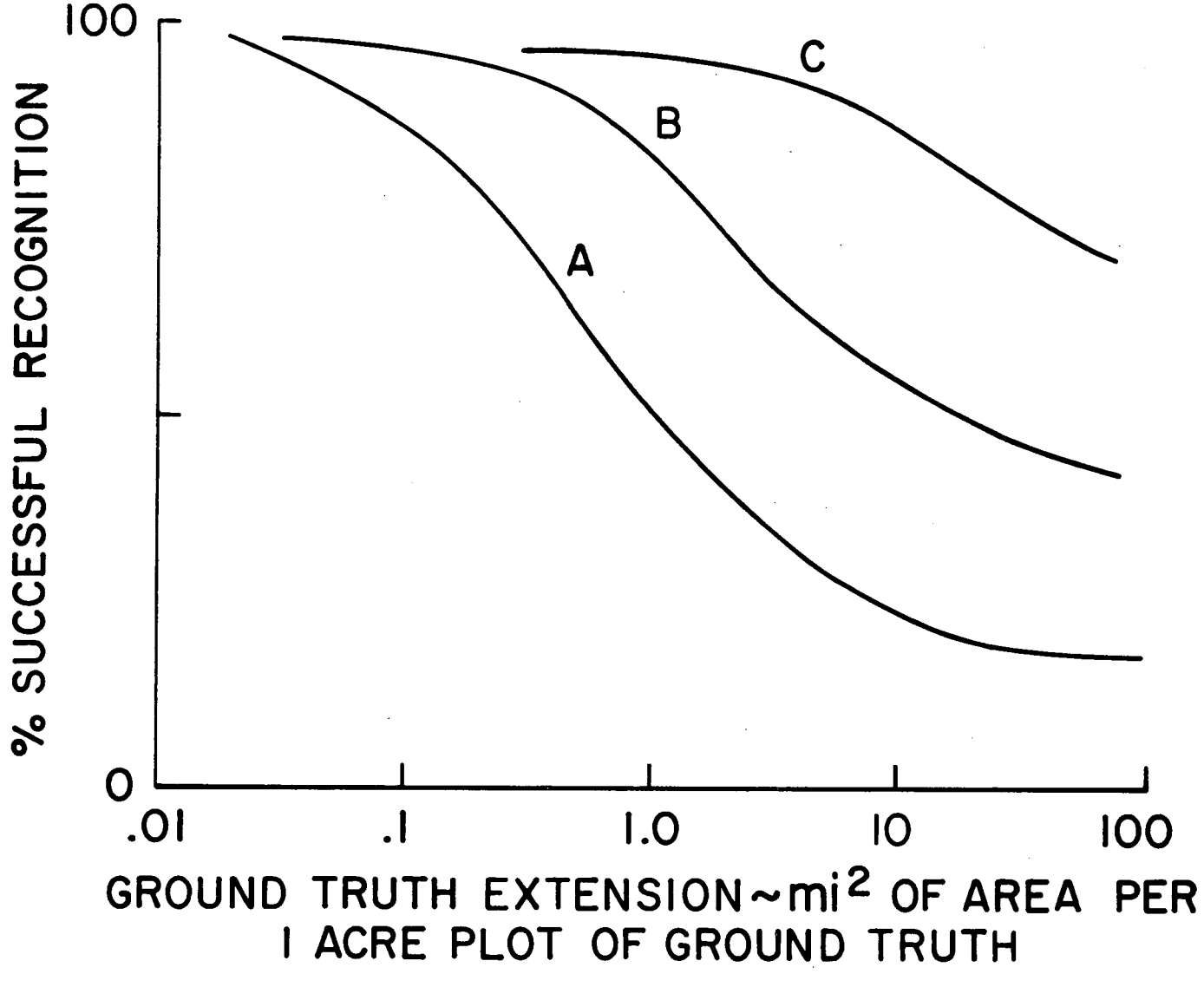

Figure 20. - Recognition Success as a Function of Ground Truth

decision will require much greater knowledge of the exact nature of this tradeoff curve than is now available. In addition, further work on the theoretical basis of signature extension is needed to offer the potentiality of moving from the situation of Curve A to that of Curve B, or even Curve C. The cost of ground truth is not only the expense of providing and maintaining ground truth sites, but of the essentially fixed costs of having to manually train the processor frequently. This tradeoff is of fundamental importance in the design of a future operational system and yet the state of knowledge concerning the details of construction of such a tradeoff is essentially nil.

Returning to the operation of the SPARC computer, we have just seen how the training of the processor takes place. Now a tape of the scene to be identified is run through the computer. The computer computes the likelihood ratio for each material that it has been programmed to look for at the point in twelve-dimensional space represented by the unknown reading. If the run is set to look for "wheat," for example, a dot is placed on a strip of 70mm film in the spot corresponding to the location of the resolution element whose signal is being analyzed if the likelihood ratio test for "wheat" is passed by the unknown material, or no dot if the test is not passed. The scanning then continues, printing a dot at each point judged by the processor to contain "wheat." When the scanning is completed the strip of film will contain essentially a map of the areas containing wheat.

A separate film must now be made for each target material, after which the resultant separate strips are combined by a mechanical process to achieve false color maps representing the various target substances. It does not seem unreasonable, in the future, to design an appropriate optical system to provide such false color maps directly.

Digital Calculation of Likelihood Ratio. - The progress described above for the analog computer has the merit of being probably the simplest technique available. It is worthwhile to discuss its processing time requirements if implemented on a digital computer.

Marshall and Kriegler (ref. 8) state that the present SPARC analog computer used at Michigan is capable of decision rates on the order of 10^5 decisions per second. Since the analog computer technology involved is circa early 1960's, it would not appear difficult to advance the capability sufficiently by 1985 to handle the expected load easily. (Analog computer throughput rates will be discussed at greater length in another section of this report.) Since we shall see that the most highly advanced digital computer of 1985, costing in the tens of millions of dollars, is barely adequate to handle the anticipated load even after it is critically reduced, it is evident that properly configured analog computers will form a part of any early operational system. Because of the highly complex and time-consuming set-up procedure required for a purely analog computer, the 1985 operational system will probably have to be a digital-analog hybrid computer.

Marshall and Kriegler estimate the load on the future system as 5×10^9 elements per day, which is very comparable to the estimate used here.

The covariance matrix must next be obtained for each target material to be considered. In the analog version, this was accomplished by such subjective steps as "eyeballing" a scope to center on axes, orienting an ellipse along a pair of axes, and shrinking an ellipse into a circle. Although all these tasks are highly subjective and approximate, they need not be done too accurately. On a digital computer the equivalent process would be to make a statistical estimate of the mean vector and the covariance matrix from a set of data points. Each covariance that must be estimated from the data should involve the calculation of a covariance term of the form:

$$T_{jk} = \frac{1}{n} \sum_{i=1} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)$$

where n is the number of samples, x_{ij} is the i-th sample used in estimating the j-th component and $\bar{X}_j$ is the sample mean of the j-th component. To compute this formula will require two subtractions, a multiplication,

and an addition for each term resulting in a total computational cost of:

$$\text{cost} = \frac{N(N+1)}{2} \cdot (3\alpha + \mu)n,$$

where N is the dimension of the covariance matrix, α is the computer add time, and μ is the computer multiply time.

If a sample size of about 100 points is selected, and the ratio between multiply time and addition time of a typical computer is taken as 5, then the cost can be written:

$$\text{cost} = \frac{N(N+1)}{2} \cdot (800\alpha) = 400N(N+1)\alpha$$

If the add time on the 1980 era computer is taken as 10^8 adds per second,* the total time required to estimate the covariance matrix for N = 12 would be 6.24×10^{-4} seconds (6.24×10^4 addition equivalents). Since only one such estimation is required per ground truth vector, this operation should not dominate the computation.

It should be noticed that the technique used here was to estimate the covariance matrix directly, not to recursively set the coefficients as would be done to manually set the potentiometers for an analog computer. If the aim of the calculation is to set such potentiometers, as it would be in a hybrid computer, then an additional step must be taken to transform the covariance matrix terms into the corresponding potentiometer terms. This would require the equivalent of a matrix multiplication, and would require $6N^2$ additions for a cost of $6N^2\alpha$ in computation time per term. For N = 12 dimensions, this would be $.086 \times 10^4$ seconds.

---

* The Illiac IV, one of the fastest modern computers, has an add time of 325 nanoseconds, or about 100 times slower than assumed here.

Table Look-up Approach

The recognition process has been shown to consist of two parts--a means of describing the distribution of a vector representing a substance in measurement space and a measure of nearness that can be used to compare the vector representations of substances. For example, the description of a distribution as a multivariate Gaussian and the use of likelihood ratio tests to describe the distribution and to test for nearness, respectively, has been extensively used. Eppler, Helmke, and Evans (ref. 9) describe a technique in which the problem is suitably reduced so that the distribution of each material is stored in a computer image of the measurement space and the measure of nearness is a table look-up that requires only a computer "fetch" operation instead of a lengthy calculation involving more time-consuming computer operations such as additions and multiplications. The method thus trades computer storage for speed. For a typical four channel problem with nine materials the likelihood ratio algorithm needs to store only nine covariance matrices of ten terms each, or 180 computer half-words compared to 28,561 for the table look-up method; Eppler, *et al.*, state that table look-up took 0.066 seconds to classify a 222 sample line compared to 2.0 seconds for the likelihood ratio method, or slightly more than thirty times faster. The accuracy was comparable:

	Table Look-Up	Purdue
Correct	92.4 %	93.1 %
Undecided	3.2 %	0.7 %
Incorrect	4.4 %	6.2 %

The table look-up method uses such large amounts of computer storage that it is limited to problems having only a few channels, three or four being the practical maximum. If the channel readings are quantized into 256 gray levels and there are twelve channels, the raw memory requirement, if no storage tricks were used, would be $256^{12} = 8 \times 10^{28}$ addresses which would exceed the size of the largest conceivable computer.

Fortunately for the method, the density of useful data in any one dimension is only about 10 percent--that is, all but 10 percent of the values along a given dimension are usually zeroes. Thus, a clever storage scheme would require 10^{12} fewer addresses than otherwise. Even this is not enough to reduce the memory requirement of the twelve channel case to practical size, however, since 8×10^{16} storage locations would still be required. Again, fortunately, the work at Purdue and Willow Run has shown that three to five optimally selected dimensions are sufficient so that the method can get by with approximately $\frac{256^4}{10}$ storage locations by using the 10 percent packing factor as well as the reduction to four channels. In addition, a technique involving "pointer" functions is used which permits reducing the storage still further by the equivalent of quantizing more grossly than at 256 levels--the area of interest in each dimension, instead of being described by 24 to 36 gray levels, is quantized to twelve levels, thus saving an additional factor of 2^4 to 3^4 storage locations. For this method the computer time goes up approximately linearly with the number of dimensions, m, but the memory requirement goes up as k^m, the total memory requirement being:

$$\text{memory} = 256\ mN + 13^N \text{ half words}$$

where N is the number of channels and m is the number of materials. The memory for four channels is 28,561 + 9,216 = 37,777 which for an additional channel would be 373,000 half-words. The speed is about 3×10^{-4} seconds per decision for four channels and nine materials, or approximately:

$$\text{time} = 5.5 \times 10^{-3}\ N \sqrt{m}$$

The square root term in m arises from the iterative search for the correct material in the algorithm, starting with the initial hypothesis that the material has remained the same from the previous classification. This expression is only approximate, and with nine materials the value found in practice was 2.3.

The authors state that the advantages of the algorithm, including its speed ". . . may make it possible to use an onboard computer to

perform the classification function in flight." According to the results we have presented so far this algorithm represents an increase in speed almost equivalent to a decade of improvement in computer speed, but it is still not fast enough to process the computing load onboard an operational earth resources satellite of the 1980's in real time. In fact, it is about halfway in speed between the Purdue algorithm and the Willow Run analog algorithm. The comparative speeds are as follows:

Algorithm	Speed-Decisions/sec
Purdue	100
Table Look-up	3,000
Willow Run	100,000

The table look-up algorithm, for all its memory limitations, should thus be able to perform as well on a reduced, but still interesting problem, as the Willow Run analog algorithm if table look up is implemented on a computer thirty times faster than at present. As we have indicated it may not be unreasonable to expect such increases in speed by the mid 1980's. In addition, there is some experience to indicate that at least one or more orders of magnitude increase in speed may be achievable through "hard-wiring" the digital circuits. Thus, the table look-up algorithm looks like a contender for operational systems of the 1980's. In particular, it is worth special attention as a potential quick-look algorithm to permit rapid, but unrefined processing results to be viewed in advance, before receiving the detailed processing results.

In addition to the above considerations, it should be noted that the table look-up algorithm makes no assumption concerning the statistical distribution of the measurement space readings, an advantage which it shares with clustering algorithms and to a lesser extent with nonparametric algorithms.

Sequential Recognition Techniques

The techniques we have been discussing have all been formulated so they can be automated on some kind of computer. Sequential techniques have shown themselves to be successful in manual form but have not been automated. We will discuss such techniques and how they may be adapted for computation, as well as how they might be combined with the other algorithms that we have been discussing.

The method relies on the empirical fact that crop planting follows a well-defined sequential pattern, so that a relatively small amount of image data on each of several dates throughout the year may be used, together with knowledge of the crop calendar, to make accurate estimates of what crop is present at a given time (ref. 10). The use of a piece of land for wheat at one time of year virtually assures that the crop grown on the same piece of land the next year will be alfalfa or sugar beets in the Phoenix area, for example. Similarly, a field used for alfalfa one year will be used again for alfalfa the next year in three out of four cases. Table 3 presents a Markovian model of crop rotation based on estimates in the literature together with some modest hypothetical assumptions. The data in parentheses are hypothetical data used to complete the matrix where no data were available in the literature.

If the crop grown on a piece of land is barley in the current season, then the field next year will contain alfalfa with probability of 40 percent, barley again 17 percent, sugar beets 14 percent, cotton 11 percent, wheat 9 percent, and pasture 9 percent. It would be possible, in fact, actually to simulate the crop rotation cycle of an agricultural basin using these Markovian transition probabilities. Non-agricultural uses to which fields are converted with a given probability are also shown.

If multispectral pattern recognition techniques have been used and it has been possible to identify a field as "either sugar beets or alfalfa" it is possible to refer to the previous year's identification

TABLE 3 TRANSITION MATRIX FOR CROP ROTATION

	Barley	Wheat	Alfalfa	Corn	Sugar Beets	Pasture	Other
Barley	.17	.09	.40	.11	.14	.09	
Wheat			.50		.50		
Alfalfa	.06	.04	.72	.06	(.03)	(.03)	.06
Corn	.26	.13		.55			.06
Sugar Beets	.80	.20					
Pasture	(.06)	(.04)	(.70)	(.06)	(.04)	(.04)	(.06)
Other	(.06)	(.04)	(.70)	(.06)	(.04)	(.04)	(.06)

Note: Numbers in parenthesis are hypothetical data.

of the field to make an improved estimate of the relative likelihood of sugar beets or alfalfa. If the field previously contained alfalfa, then it is 0.72 against 0.03 that it is again alfalfa, other things being equal. If the field previously contained barley, however, then it is 0.40 against 0.14 that the field now contains alfalfa and not sugar beets. This clearly indicates an opportunity to implement the crop cycle model as a decision aid in crop recognition algorithms. For an example, in a Bayesian (unsupervised) model, the conditional knowledge of the previous year's crop could be used to determine the *a priori* probability of each crop's being present and thereby modify the recognition algorithm. Most of the methods we have discussed utilize *a priori* probabilities of occurrences of crop types to modify the likelihood ratio test, and even the table look-up algorithm uses *a priori* ranking of tables in the look-up procedure. This should permit the recognition accuracy of the methods to be increased, or possibly to increase the speed. To do so would require that data from previous observations be stored for ready access in a computer memory or perhaps in some form such as the Ampex Video-file. Given that this requirement could be met there are two ways that the data could be used; one to modify the multispectral recognition algorithms by using *a priori* data, and the other an entirely new algorithm that automates the sequential manual techniques and employs multispectral recognition as part of the algorithm.

Since sequential methods are reported to achieve 80-90 percent recognition essentially without employing any multispectral data, and since multispectral algorithms achieve similar accuracy without employing any sequential data, it would seem reasonable to expect that excellent results might be obtained by combining features of both types of algorithm. Before examining how this might be done, let us consider the information storage requirements of such an algorithm.

If we assume an area of 3×10^6 km^2 and a resolution cell of 10 x 10 meters, there would be 3×10^{10} resolution elements. To store ten samples of data for each element 3×10^{11} units would be required.

How much data would have to be stored for each unit? Assume that one multispectral reading has been accumulated for each resolution element for each season, and these data are to be recorded. Since each reading is a sample of one, not a distribution, it would not be necessary to store the entire covariance matrix, but only the components of the single vector reading. Assume quantization to 256 gray levels each of twelve channels--then $\log_2(256) \times 12 = 96$ bits of information are needed for each of the 3.0×10^{11} units, or about 3.0×10^{13} bits of information in total. Optical memories of this capacity are available and cost approximately $4 million, making the implementation of this particular scheme rather expensive.

We have thus shown that it is probably too expensive to store multi-date multispectral information on a pixel by pixel basis. But in an area of 3×10^6 km^2 there would be far fewer *fields* than pixels, so it would be less expensive to store information on a field-by-field basis. Assume that there are 10^7 fields in the area. This would correspond to a field size of about 65 acres. Referring to figure 5, it can be seen that this assumption covers nearly all cases since the average field is nearly ten times larger. Then we save a factor of 3.0×10^3 in the number of stored data points. The address would be more complicated in the case of storing fields and so the savings would be reduced somewhat, but about 10^{10} bits of storage would probably be required for this scheme.

Let us consider how the method might work. Refer to figure 21. A multispectral test is performed on the March 8 observation and it is tentatively decided on the basis of this test that the field belongs to the barley-sugar beet-alfalfa-rye group rather than the alfalfa-rye-lettuce group or bare soil. The April 23 data is next studied and subjected to a new multispectral test, perhaps using different channels than the previous test and sugar beets are ruled out. Then a multispectral test is performed on the May 21 data and it is tentatively decided that barley can be ruled out but it is decided to check the data for August 5, at which time the multispectral test confirms with high probability that

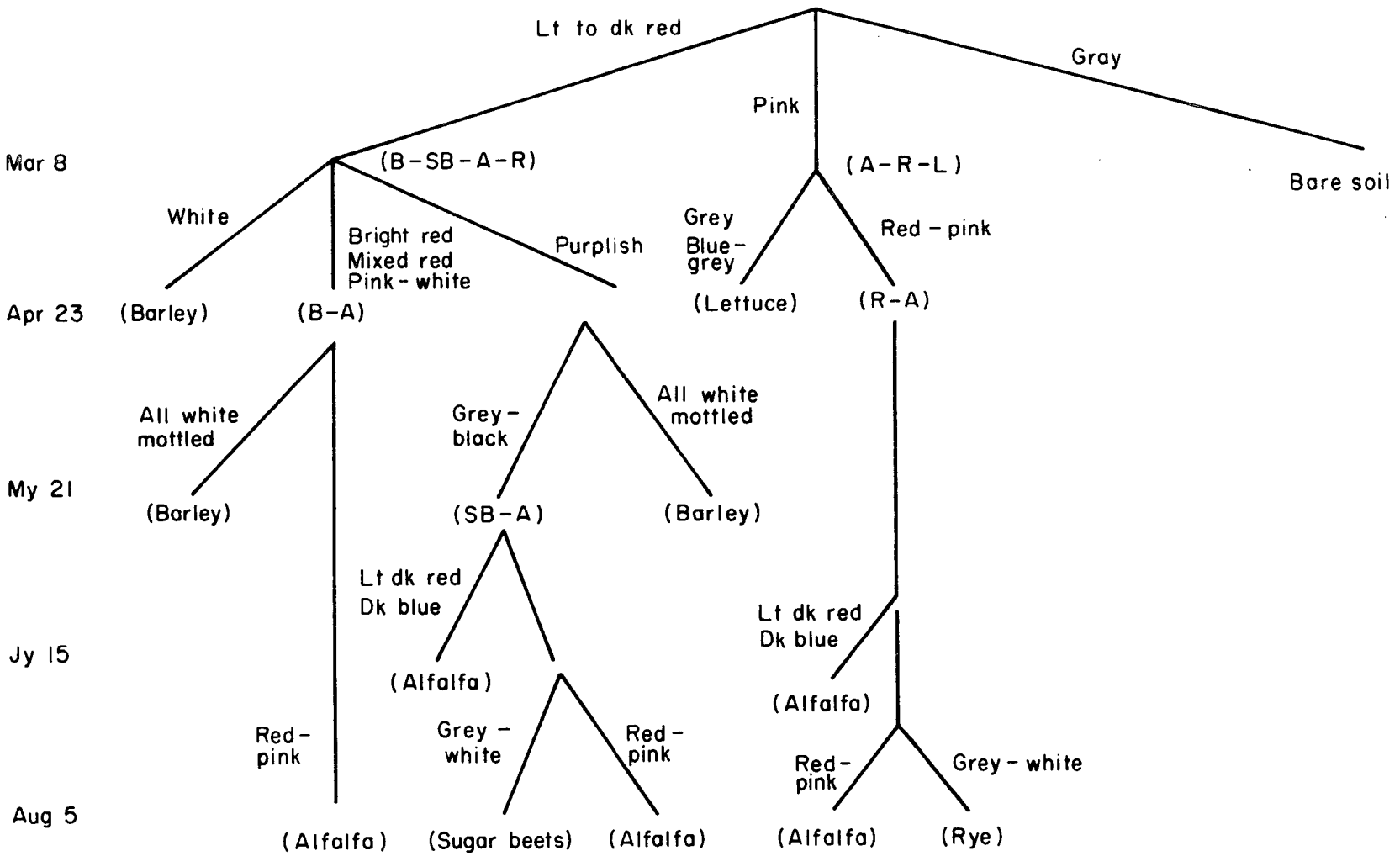

Figure 21. - Sequential Recognition Process.

the crop is alfalfa. Little multispectral data would be necessary in this process. This is borne out by the fact that in the manual sequential technique it was only necessary to notice that with IR Ektachrome the crop appeared light to dark red on March 8; bright red, mixed red, pink, or white on April 23, not white or mottled on May 21; and red or pink on August 5.

This technique utilized the fact that crops are easiest to distinguish from each other at certain critical times of the crop cycle. Thus, to put it into the mathematical terms of the pattern recognition algorithms, an interclass divergence chart (table 4) is prepared for each observation date. Interclass divergence will be explained in the following section. For present purposes, it is enough to understand that the higher the numerical value in the table, the more likely it is that the two crops may be distinguished from each other on the given date. There would not only be a best set of channels that would emerge but also a best set of dates to maximize the interclass divergence for each crop discrimination function. By combining the observations from the whole series of dates the overall performance of the algorithm would be optimized. For example, the divergence between corn and wheat classes is greatest (310) on July 15, in our hypothetical example and this date would give the best discrimination.

It is conceivable that a simplified method in which, for example, only a small amount of condensed information was stored for each field might even increase the accuracy of recognition; for example, rather than storing ten previous observations of the multispectral signature of a field it might be enough merely to store what the field contained at the last observation, the date when it last contained bare soil, or some other simple item of information which nevertheless contains much implicit knowledge of the previous crop history of the field because of the strongly determined crop rotation model of the basin.

TABLE 4 INTERCLASS DIVERGENCE BY OBSERVATION DATE

Date	Corn/Wheat	Corn/Oats	Corn/ Sugar Beets	Oats/Wheat	Oats/ Sugar Beets
March 8	75	80	350	120	25
April 23	200	120	310	25	100
May 21	45	70	65	310	80
July 15	310	35	110	45	200
August 5	20	60	300	200	120
(MAX)	(310)	(120)	(350)	(310)	(200)

Figure 22 is a typical crop calendar that illustrates the large amount of knowledge of crops that can be used to distinguish among crop types that such a calendar contains. Mature barley can be identified in April and May when it turns golden and is harvested. In April, barley is not golden and shows bare soil in May, so that, for example, rye and alfalfa can then be distinguished. Cotton will show bare soil in April and preparations for planting in March. Lettuce will show bare soil in April and harvesting in March. All of these differential variations can be employed to improve discrimination by using multidate techniques.

Channel (Feature) Selection

We have described the simplest form of operation of a multispectral signature processor for an analog, a digital, and a hybrid processor. In each case we have assumed that the number of dimensions N was a fixed quantity. But referring again to figure 16 it can be seen that the exact number of spectral bands used to approximate the continuous spectral signature is actually a design variable. The greater the number of dimensions the more complex will be the multispectral scanner equipment, the greater will be the bandwidth used to transmit the data from the vehicle to the processing equipment, and the greater will be the computer requirement for handling the multidimensional data. On the other hand, there is the question of what happens to recognition performance or misclassification as the number of bands is increased.

In Figure 23, the measure of the performance of the algorithm is shown increasing along the dotted line as new bands are selected at random are added to the feature set. Ultimately a point is reached after which only marginal improvement can be obtained by adding bands, since the continuous spectral signature (figure 15) is not a very rapidly changing one (bands highly correlated) and relatively few bands give an adequate representation of it. At the same time, computational costs are going up, some as the square, and others as the cube, of the number of

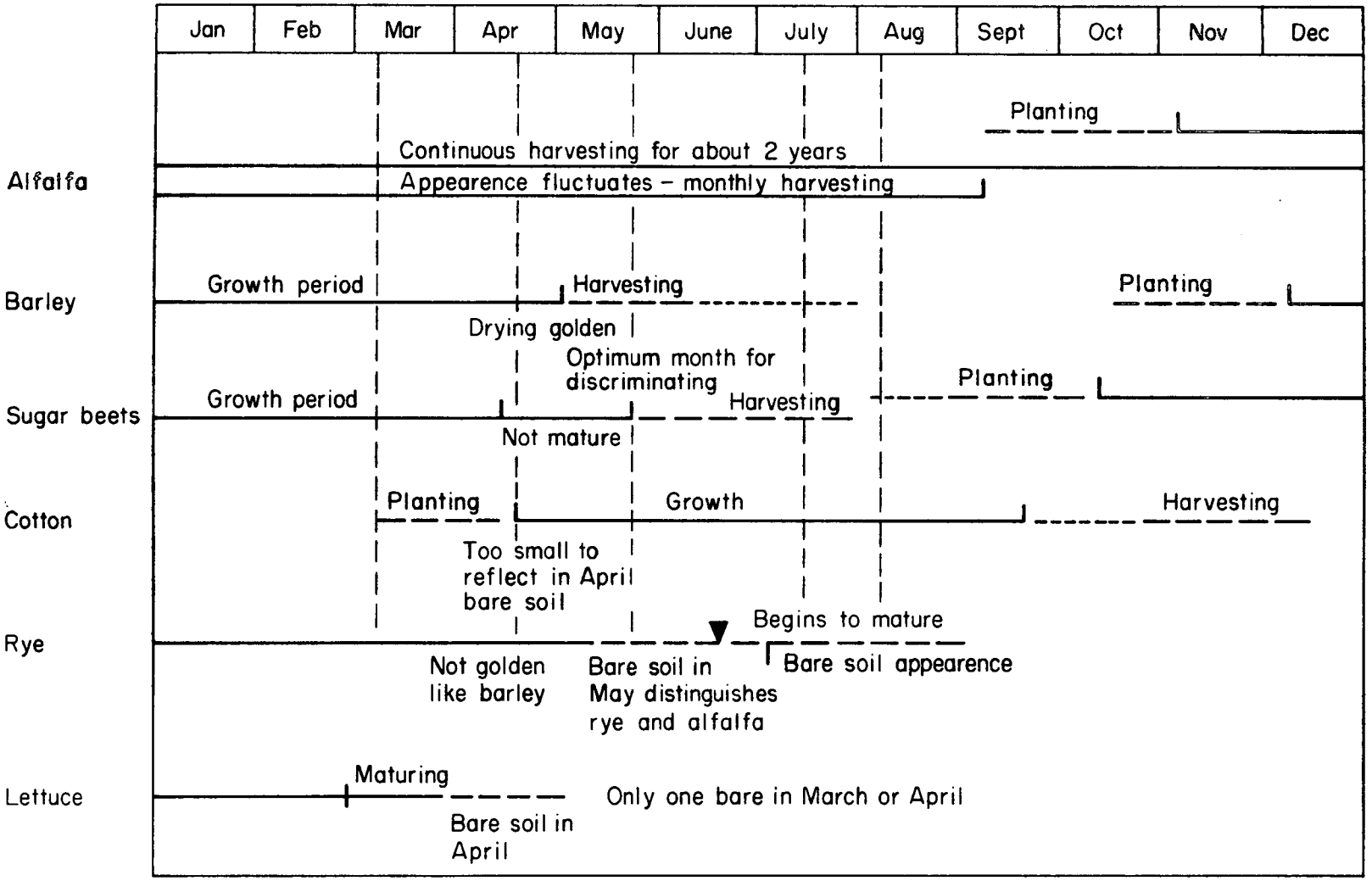

Figure 22. - Typical Crop Calendar.

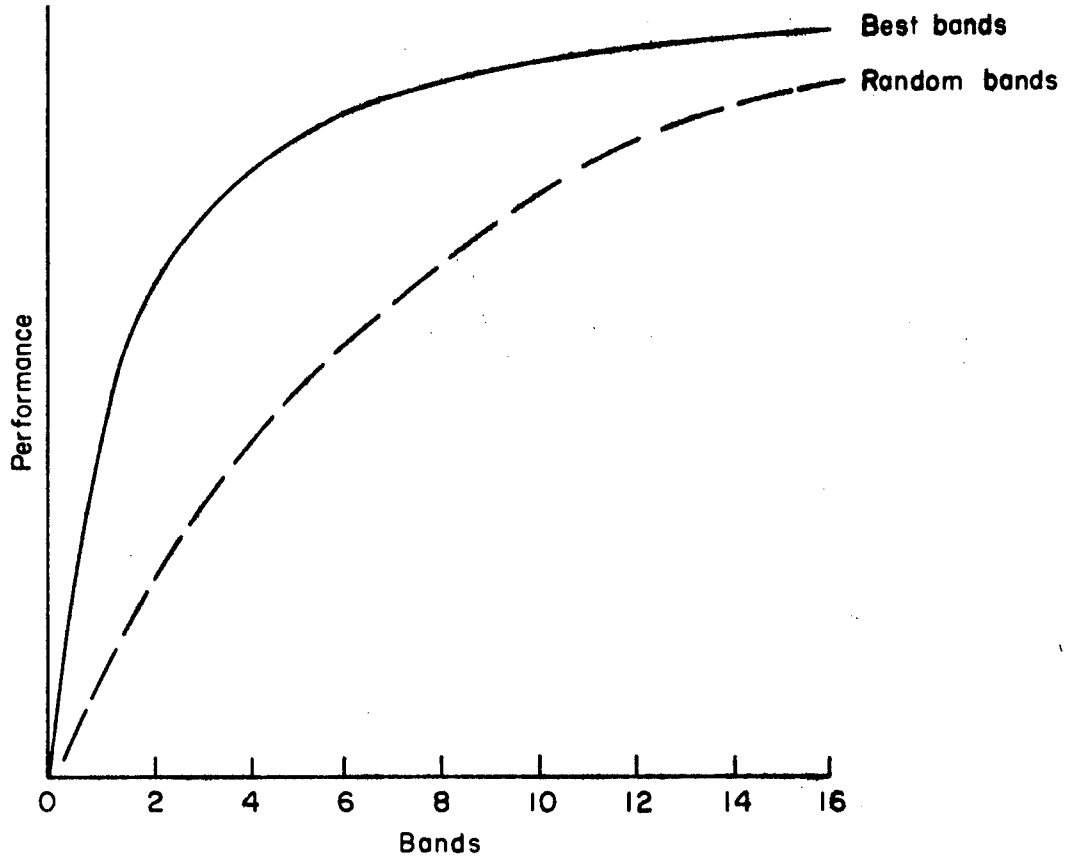

Figure 23. - Algorithm Performance as a Function of Bands Used.

dimensions so the relatively small marginal gain soon becomes offset by the increasing costs of additional bands. If an algorithm is included that enables one to decide which are the *best* bands to use then the situation is as in the solid curve; the performance increase occurs more rapidly and reaches the optimum number of bands much sooner. There are several methods for selecting such optimum bands ranging from the factor analysis approach used by Bendix to divergence methods used by Purdue. The simplification of being able to use fewer bands would be more than offset by the cost of the calculation to determine the best bands if the calculation had to be done frequently. Fortunately, this simplification can be used in practice. If the test is being made in an area in which the parameters of the problem are changing very slowly then the selection of the optimal set of features may be made once and an entire set of identifications may be run with the smaller optimal feature set. Now it is worthwhile to examine the feature selection.

The basic concept is that there is an optimum set of channels for separating classes of materials and that there is a relationship between numerical values of a measure called "divergence" and the degree of separation of material classes in terms of classification errors. In the Purdue system, this step may be done interactively, that is, with the experimenter supervising the operations "on line" and superimposing his judgments on the results of the machine tests.

The technique described by Fu (ref. 12) is rather complex mathematically, but a modified presentation of it will be given here. A performance measure for a classifier is postulated which is the sum of all probabilities of misclassification. The probability of a misclassification is the probability that a discriminant function indicates that an unknown belongs to class i when it in fact belongs to class j, or the reverse. As in previous examples we assume linear discriminant functions. A separability measure, d_{ij} is derived from these linear discriminant functions that permits the probability of misclassification to be reduced to a probability statement about a standarized Gaussian random variable.

$$P_{ij}(\varepsilon) = 1 - \text{Prob}\ (\varepsilon < d_{ij})$$

$$d_{ij} = \frac{B_{ij}\ (M_i - M_j)}{(B_{ij}^T\ K_i\ B_{ij})^{\frac{1}{2}} + (B_{ij}^T\ K_j\ B_{ij})^{\frac{1}{2}}}$$

where ε is the standardized (univariate) normal, variable d_{ij} is the separability measure, B_{ij} is derived from the combination of discriminant functions $B_i^T X - c_i$ and $B_j^T X - c_j$, M_i and M_j are the vectors of means of the i-th and j-th classes. Using this derived value of d_{ij}, the method then is to maximize the expected value of the separability for all pairs of classes,

$$\text{Max} \left[\sum_{i=1}^{M} \sum_{j=1}^{M} P(w_i)\ P(w_j)\ d_{ij} \right] , \ i \neq j$$

That subset of features is selected which maximizes the above criterion.

In practice the above operation is done interactively, so that manual assistance can be given to the computer in the selection of the best subset of features. Other methods mentioned by the previous authors include direct estimation of error probability, use of the feature space transformation (Karhunen-Loeve expansion) and use of a stochastic automata model. Factor analysis is also apparently used by some investigators.

It is not clear whether this method would have to be used on-line frequently in an operational system or whether it would be possible to select an optimal set of features for a given region well ahead of time, off-line, and merely look up the optimal set of features each time that region was being explored. For the purposes of this paper, we are going to make the assumption that this calculation will be done off-line and the appropriate features will be looked up when needed. This eliminates another slow step, and one which at least at present requires interaction

with a human as an important part of its operation. For the purposes of estimating the time required the time-consuming steps other than the human interaction steps are probably the computation of the Lagrange multiplier, λ_{ij}, and corresponding value of B_{ij} which maximizes d_{ij} in the above d_{ij} expression. This requires a matrix inversion of an N x N matrix:

$$B_{ij} = \lambda_{ij} K_i + (1 - \lambda_{ij})^{-1} (M_i - M_j)$$

Table 5 shows a sample calculation of optimal feature selection from a Purdue LARS Program annual report. For each subset of features and for each pairwise combination of classes the interclass divergence is computed and the subsets are ranked in decreasing order of total divergence. If a subset of k features from a total of N is desired* and there are m classes then the total number of interclass divergences to be computed is:

$$\binom{N}{K}\binom{M}{2} = \frac{N!}{(N-K)!K!} \cdot \frac{M!}{(M-2)!2!}$$

In the example shown here, N = 12, k = 4, and m = 5 so the number of interclass divergences is:

$$\frac{12!}{8!4!} \cdot \frac{5!}{3!2!} = 4950$$

That is, there are 495 subsets of twelve features taken four at a time and for each subset ten interclass divergences must be computed. If each calculation takes on the order of 10^{-4} seconds (on a 10^8 operations/second machine) then the entire calculation will take only 0.5 seconds, but that is not the difficulty as far as speed is concerned. It is not enough merely to select the feature subset having the highest total divergence; currently the selection of the best feature subset is left to human intervention, thus requiring setup time far in excess of the

---

* It is possible for a subset of features to classify better than the complete set. Fu (ref. 12) points out that this is probably due in part to the error in estimating the means and covariances.

TABLE 5 OPTIMAL FEATURE SELECTION

	Features	D_{ij}(Min)	D(Tot)	Interclass Divergence									
				SC	SO	SW	SM	CO	CW	CM	OW	OM	WM
1.	1,6,10,12	36	1538	36	84	194	179	120	111	347	95	133	0
2.	1,6,10,11	33	1534	33	81	196	184	120	111	111	90	130	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
30	1,10,11,12	28	1407	32	74	166	174	94	300	343	107	117	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
34	1,5,10,12	31	1403	31	72	170	181	91	279	111	89	140	0

computer time. The large total divergence may come primarily from a single very large interclass divergence, in which case the large total divergence is not truly representative of the ability of the feature subset to distinguish between all the classes it is required to treat. On the other hand if one of the interclass divergences is too small then the classifier will have difficulty discriminating between the corresponding classes, so that again the feature subset leaves something to be desired. In the example shown, the feature subset that was chosen turned out to be (1,5,10,12) with rank = 34, because it exhibited the best *balance* of the feature subsets displayed.

This step, with human interaction, might require fifteen to thirty minutes of setup time. It appears that the process could be totally automated, but it is not clear whether this would be satisfactory. In an operational system it may be that the best sets of features would all have been precalculated for the prevailing conditions in each area and that they would then just be looked up in a data bank. One approach would be to use a statistical Monte Carlo process to estimate the probability of miscalculation. A sample of say 10,000 known data points could be classified in .3 seconds per feature subset and the misclassifications could be totaled. The result of running one such case would look something like table 6. If a weight could be given to the importance of each type of misclassification the weighted total of misclassifications could be used as a criterion for selecting the subset of spectral bands that gave the lowest value of such a weighted sum. An entire batch of $\binom{12}{4} = 495$ subsets could be calculated in 150 seconds of computer time. The decision could be made by the computer or human interaction could be used as before. About ten minutes would now be required for this step if a human decision were required.

The classification algorithm used by Purdue is essentially the same likelihood ratio test used by Willow Run on their analog computer, except that here it is done digitally. Instead of taking the likelihood ratio they use the equivalent test:

TABLE 6 HYPOTHETICAL MONTE CARLO RESULTS

Classification Summary by Test Classes

No. of Samples Classified into

	Class	No of Samps	Pct Corct	Soyb	Corn	Oats	Whea	Red	Alfa	Rye	Soil	Thrs
1	Soyb	2368	85.9	2035	39	133	1	0	0	0	10	150
2	Corn	588	94.0	25	553	1	0	1	0	0	0	8
3	Oats	370	84.9	0	0	314	0	56	0	0	0	0
4	Whea	806	91.2	0	0	17	735	0	0	48	0	6
5	Red	1401	85.9	2	20	38	0	1203	133	0	0	5
6	Alfa	456	87.7	0	4	21	0	29	400	0	0	2
	TOTAL	5989		2062	616	524	736	1289	533	48	10	171

Overal performance = 87.5

Average performance by class = 88.3

$$\text{If } \log / K_i / + (X - M_i)^T K_i^{-1} (X - M_i) \leq$$

$$\log / K_j / + (X - M_j)^T K_j^{-1} (X - M_j),$$

then classify X as belonging to class ω_i, for all $j \neq i$. This is essentially equivalent to computing the density function for each class at the point in N-dimensional space represented by the unknown X and selecting the class that gives the greatest density, just as was done for the analog computer. Similarly, because of the difficulty of anticipating all the ground truth distributions that might eventually be required in this approach, an all-inclusive "rejection class" is formed by using a threshold for rejection as follows: reject from class ω_i if

$$\log /K_i/ + (X - M_i)^T K_i^{-1} (X - M_i) > T_i',$$

where T_i' is the threshold for the class ω_i. Because the distribution of this random variable is known to be χ^2 with N degrees of freedom it is possible by using χ^2 tables to adjust the threshold for each class to control the fraction of rejections that occur.

To estimate the time required, we note that we already know the time to calculate K_i^{-1}, and we can assume that it is only necessary to calculate $/K_i/$ once and store the value, so this leaves only the matrix multiplications to compute the quadratic form. This involves computing N^2 terms of the form:

$$(X_i - M_i)(X_j - M_j)\ \alpha_{ij}$$

each of which requires two multiplications and one addition. Assuming as before that one multiplication equals five additions, the computational cost is $11N^2$ addition equivalents. The cost is incurred each time the unknown is tested against a particular class, and for each unknown, so it is:

$$11N^2 \quad m \cdot n \cdot \alpha$$

where N is the number of features, m is the number of possible classes of material that may occur, n is the number of unknown samples to be tested, and α is the computer add time. If as before, we let N = 12, assume $\alpha = 10^{-8}$ and assume that m = 20, this is:

$11 \cdot 144 \cdot 20 \cdot 10^{-8} n$ or

$3.17 \times 10^{-4} n$ seconds.

Since n may be on the order of 10^7 to 10^8 samples per frame, the computation time is on the order of 3×10^3 to 3×10^4 seconds per frame, or from one to ten hours for each picture. This is clearly the calculation that is the most expensive, and anything that can be done here to reduce the calculation cost would help. It is for this reason that the analog computer has been used for this step. Again, since there is an N^2 term involved, it may justify the lengthy calculation of selecting, say, three best features, since this would reduce the computation time by a factor of $(12/3)^2 = 16$.

The relative importance of accuracy versus speed is not totally clear but it would appear that the presumably lesser accuracy of the analog computer would be at least adequate for solving the problem and the increase in speed would more than offset any possible loss in accuracy in going from digital to analog. If the classification performance of existing techniques is adequate (approximately 90 percent appears to be a typical result) then the bulk of the future work should probably be applied to increasing the speed rather than on new methods to improve the performance measure. The possible use of a potential function method and a nonparametric classification technique are thoroughly investigated by Wacker (ref. 13).

One interesting result given by Wacker and Landgrebe is that if results for many classifications are to be averaged together then the parametric classifier does just as well as the nonparametric classifier.

In general, however, it was found that the performance of nonparametric methods was better than parametric; in one experiment, for example, nonparametric was 95 percent compared to 90 percent accuracy by parametric methods.

"Training" the Classifier

A surprisingly small number of sample vectors is required to "train" the classifier. The method doesn't work with only one vector, but as figure 24 shows the performance is not much worse with only two training vectors than with twenty or forty vectors, whether with parametric or nonparametric methods, when using the best two or the best three features. This might appear to be a paradox since we have mentioned earlier that performance declined as the distance from the ground truth site increased, but it is really not a paradox at all. The present experiment used training samples scattered evenly through the field and the field was not large enough for the performance degradation due to wide separation from ground truth sites to begin to take effect. Thus, it is not so much how many training vectors are taken as it is how close they are to the unknown element being estimated. Wacker and Landgrebe point out that the number of samples required to estimate the covariance matrix is usually taken as 10 q, where q is the number of terms in the covariance matrix, ($q = N(N+1)/2$ in our calculations above).

One of the most important aspects of the recognition problem is that the training data are not truly representative of the data to be tested. When classifiers are used to identify the same data used to train them, they may score 95 percent, but when used to identify new data, they may only do so with 90 percent success. In general, the situation is even worse than this for new data, so that it is not uncommon to see test data scoring 20 percent to 25 percent below the training data. Wacker and Landgrebe (ref. 13) state:

> "Until training techniques are developed which ensure that the training data is truly representative of the test data

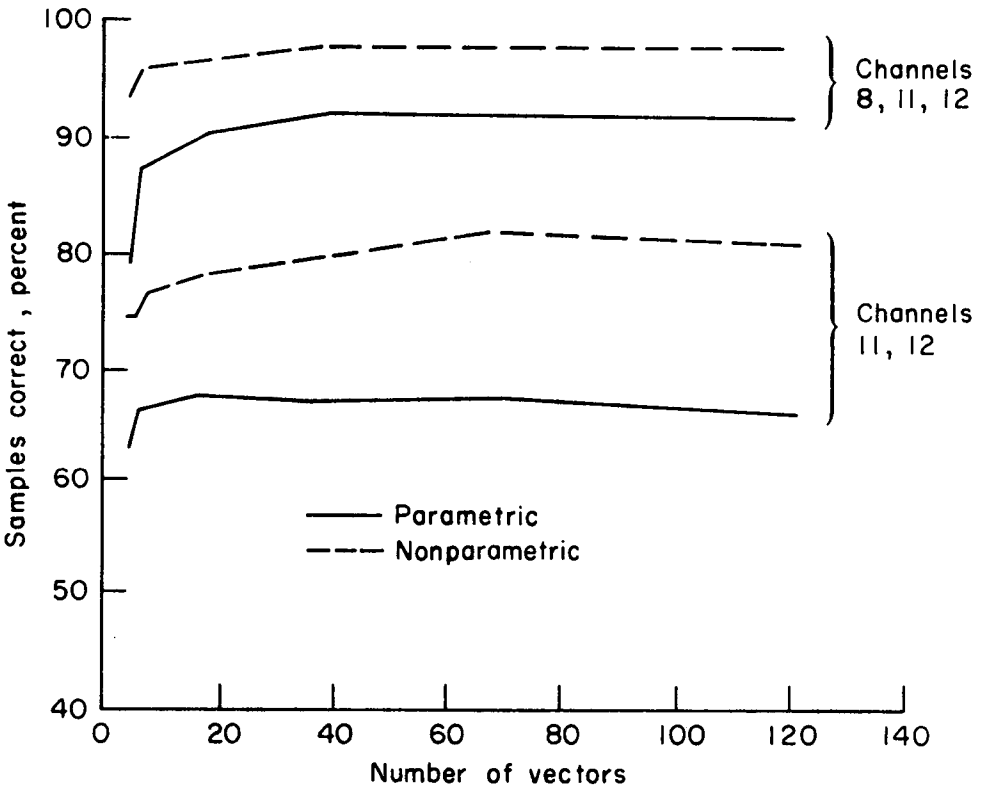

Figure 24. - Performance Versus Number of Training Vectors.

C.2

the choice of distance in a minimum distance classifier is not critical, and the extra complexity of a nonparametric classifier is not warranted."

Boundary Identification in Pattern Recognition

Locating boundaries is ordinarily thought of as a preprocessing task, but it may be important in pattern recognition as well. The most important use is the identification of training fields that are used to train the processor with ground truth. A simple approach is the one used for boundary recognition in preprocessing. A sample of say five points along the scan line is taken and averaged. The sample variance is computed. As long as all five points are the same material the variance will be relatively small, but as soon as the moving five-point average first includes a point from a different material the sample variance will increase. Thus, changes in the sample variance can be monitored as a simple way of detecting a boundary. Wacker (ref. 14) describes a method for using clustering to find spatial boundaries other than the one-dimensional case just described--that is, his method enables the use of several channels of data in setting boundaries. It is important to do this because a boundary may show up in one channel but not in another channel. The method described by Wacker is a complex heuristic algorithm that is said to work for approximately Gaussian data sets and at the time of writing did not necessarily produce closed boundaries.

Boundary enhancement algorithms are described by Su, *et al.*, (ref. 15) and a boundary following and recognition algorithm is described by Kuehn, Omberg, and Forry (ref. 16). Kuehn, *et al.*, gives the time equation for a pixel by pixel approach as:

$$T = (5/2)C^2F\,N^2\,(\alpha + \mu)$$

where C is the number of channels, F is the number of signature classes, and N is the number of pixels on a side of a square region. If we assume $\alpha + \mu = 6\alpha$, as before, and N on the order of 10 to 100, C = 12 channels,

and F = 10 signature classes, we have a time requirement on a 10^8 op/sec computer, of from 2×10^{-2} to 2 seconds depending on the size of square area compared to a pixel. The alternative algorithm they present for border following requires:

$$T = (4CN^2 = 80\ C^2N)\ (\alpha + \mu)$$

or 7.2 to 16.3×10^{-3} seconds which is one or more orders of magnitude faster than the previous method.

If a closed area can be identified without having to identify each element in it by using the maximum likelihood test with full dimensionality, then the data points in the entire field might be averaged and a highly accurate identification made of the field as a whole. Presumably this could be done with less computer time than the element by element classification, since the most expensive processing step is the likelihood test.

Spectral Signature Extension and Unsupervised Recognition Algorithms

The literature indicates that spectral signatures from ground truth are reasonably valid from "5 to 50 miles" from the point where ground truth is taken. Elsewhere we have indicated that it is not the amount of the ground truth data that is used that is important in successfully using a recognition algorithm but the distribution of the ground truth data. Why cannot the ground truth data be easily extended? What are the chances of ultimately storing enough spectral signature information in the processor memory to make it possible to do entirely without ground truth? Or, alternatively, to what extent can the expensive and time-consuming process of obtaining ground truth and of training recognition algorithms for a given large agricultural scene be reduced? Can enough basic signature data, accumulated under a variety of conditions, be stored to permit recognition of a scene under the same conditions?

These questions require a discussion of the nature of a signature. Assume that a resolution element was chosen that included only one plant and it was desired that the algorithm "recognize" the plant from basic spectral radiance measurements. What would be required? At a minimum the signature of a bright leaf, a shaded leaf, bright soil, and shaded soil would be required since all four objects would be present in the field of view. (See figure 25.) The radiance of each object would depend on the illumination angle with respect to the line of view. The relative proportion of the objects in the line of view would depend again on the illumination angle and the line of view. The scene would be illuminated by sunlight, sun transmitted through the plant, and plant-reflected light, by soil-reflected light, and by the light from the sky, in varying proportions; for example shaded soil would be illuminated by sky, sun transmitted through the plant, and plant-reflected sun. To reconstruct the overall picture from its elements then would require knowing the radiance response of a leaf and of soil to sunlight, sky, plant-reflected sun, and plant-transmitted sun, or eight signature profiles in total. The net effect of the scene would depend on the plant configuration, the illumination angle, the view angle, the percentage of ground cover, row direction, row width, crop canopy fullness, weeds, etc. For example, figure 26 indicates how the three-dimensional signatures for the four basic objects might combine as a function of viewing angle for a constant illumination angle as the relative proportion of the scene represented by various objects changes.

The complexity of combining the basic signatures is evident and is the reason why a statistical approach is used. A large group of plants is measured, averaging together all the various elements, and this set of measurements is used for training under the assumption that if all the conditions remain roughly the same when the test set is measured then all the possible variations will cancel out. But the fact that it is impossible always to measure the test set under the exact same conditions is the heart of the problem. It is not always known precisely why "corn blighted to degree three," for example, has a different

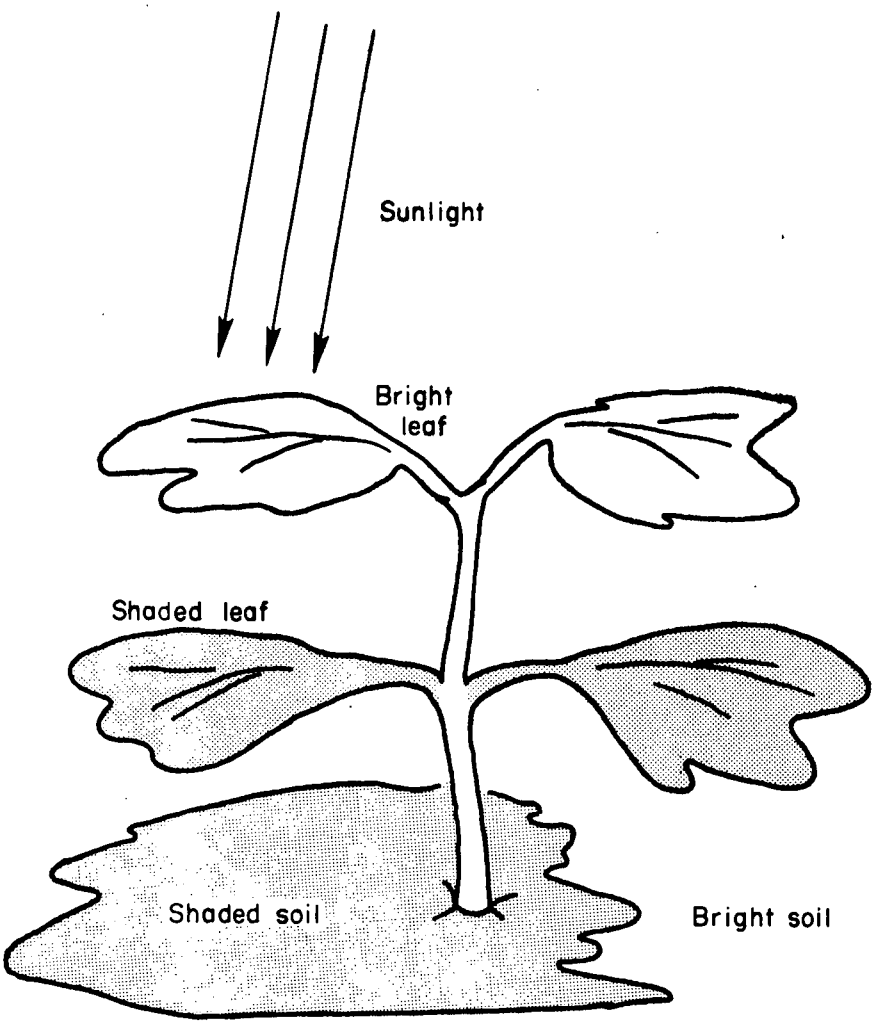

Figure 25. - Objects in Simple Signature.

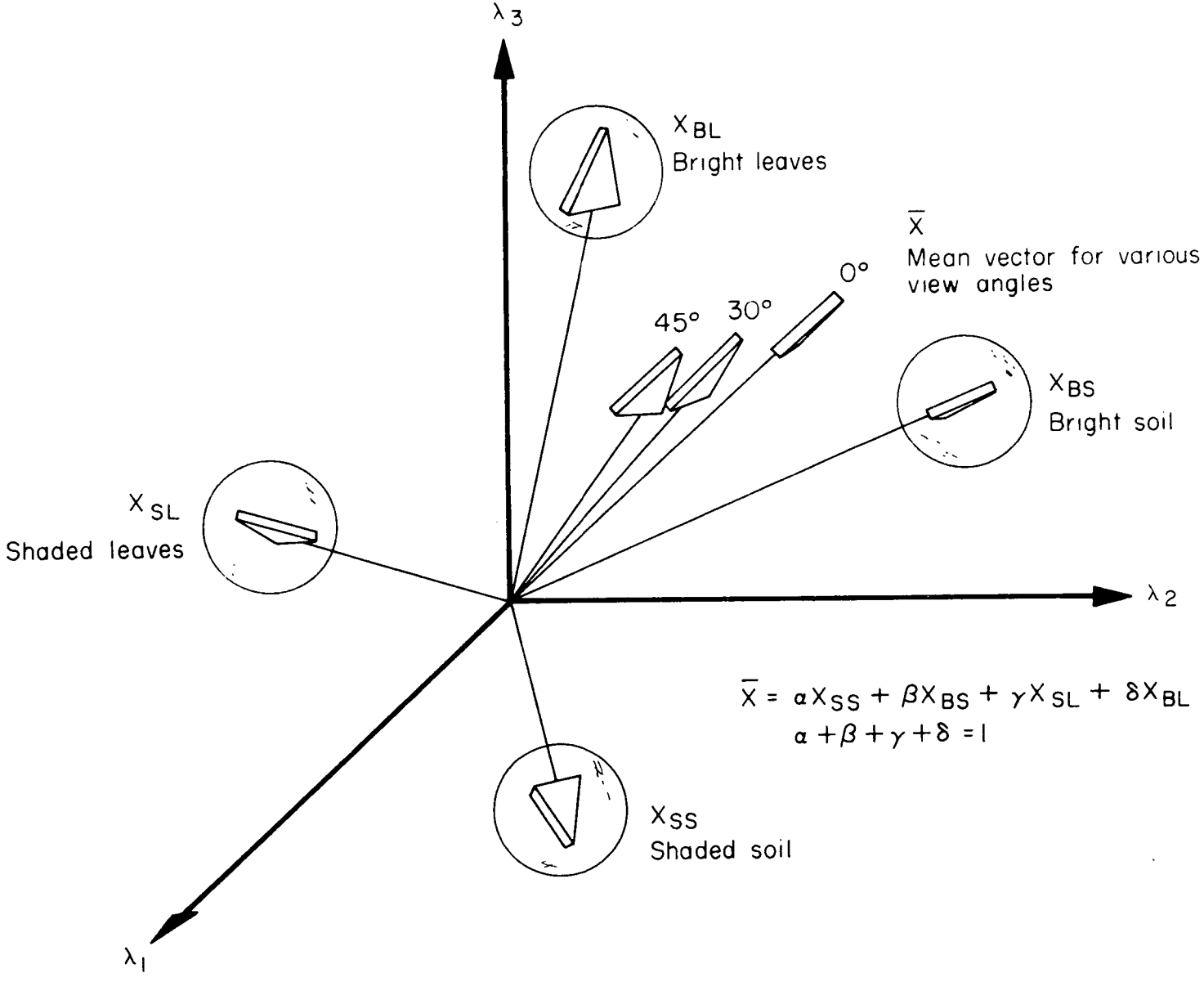

Figure 26. - Effect of Viewing Angle on Signature.

spectral signature than "healthy corn," only that the signatures are different--it may be partly because the relative percentages of bright and shadowed soil, bright and shadowed leaf are changed that the signatures are different. It could even be caused by drooping leaves.

It is possible that an entire scene or a large portion of it is shadowed. It has been found possible to obtain good recognition of a shadowed scene by using a ground truth signature from a non-shadowed area by appropriate filtering, both in the case where the area is known *a priori* to be shadowed and in the case where this is determined from the data. Kriegler (ref. 17) has developed a method for correcting for such variations in the data over certain spatial areas in preprocessing.

Summary of Processing Algorithms

Classification algorithms fall into three classes; parametric, nonparametric, and feature space transformation methods. The feature space transformation methods are used when there is no knowledge of the statistical characteristics available. The computation time is relatively low, but the performance is not satisfactory compared to other methods, so it will not be discussed further at this time. Parametric methods use known statistical characteristics to identify materials of the scene. They are most useful when the statistical characteristics of the scene are known, and their accuracy is not good if the statistical characteristics are not known. The computational requirements for the parametric methods are not as great as for the nonparametric. Nonparametric methods, strictly speaking, are those methods in which no knowledge of the statistical distribution is required; in the present discussion, we use the term to include such techniques as clustering. Nonparametric methods require extensive computer time compared to parametric methods, and since their performance on typical problems is not much better than parametric methods, the latter methods should probably be emphasized. In addition to whether the algorithm is parametric or nonparametric, it is important to distinguish between supervised and unsupervised algorithms. The only unsupervised

algorithm we have discussed is the clustering technique. The supervised algorithms compare a training set with the unknown material to be discriminated, whereas the unsupervised algorithms do not require such ground truth data for their operation. At one time the problem of developing unsupervised algorithms was thought to be absolutely unsolvable. As far as is known clustering represents the only type of unsupervised algorithm which makes no assumption concerning the nature of the underlying statistical distributions. In the parametric case, unsupervised learning consists of using a Bayesian approach to successively improve a distribution of the parameter as unlabeled samples continue to arrive in the system.

Table 7 compares the features of the algorithms discussed in this paper. These are clustering, likelihood ratio, table look-up, and sequential.

Of the methods considered, only clustering may be considered an unsupervised algorithm, although likelihood ratio and sequential could presumably be adapted to unsupervised approaches. The table look-up method could probably not be used in an unsupervised mode, since the storage of ground truth in tables is inherently a supervised mode of operation.

All of the methods considered could be implemented on a digital computer. The clustering technique could never be implemented on an analog or hybrid computer, however. Therefore, if some unsupervised algorithm is required on an analog computer, it would have to be a Bayesian technique. Due to the memory requirements for such techniques, it seems unlikely that an unsupervised algorithm will be implemented on an analog or hybrid computer. Similarly, table look-up would never be implemented on an analog computer, since it is an inherently digital technique.

Of the techniques considered, it appears that the potential accuracy of the sequential technique would be the highest, although a great deal

TABLE 7 COMPARISON OF CLASSIFICATION ALGORITHMS

	Unsup.	Digital	Analog/ Hybrid	Accur. Poten.	Speed	Remarks
Clustering	Yes	Yes	No	High	Slow	Good potential for unsupervised algorithm. Relatively slow--can't speed up by going to analog. Useful for classifying unknown data without ground truth.
Likelihood Ratio	No	Yes	Yes	High	Analog-very fast Digital-medium	Most commonly used. Very suitable for analog.
Table Look-up	No	Yes	No	High	Fast	30X faster than likelihood ratio with same accuracy. Not suitable for analog. Uses computer memory heavily. Unsuited to more than 4 features.
Sequential	No	Yes	Prob. No	Very High	Slow	Not yet automated. Potential for digital or analog. Potential for highest accuracy.

of work would have to be done before this potential could be demonstrated. In addition, the memory requirement of the sequential technique might prove to be insuperable. The accuracy of the likelihood ratio and the table look-up technique have been shown to be equivalent in a few test cases. The accuracy of the clustering technique is not strictly comparable because it is inherently an unsupervised approach so it will not classify into a set of fixed categories. In fact, one of the problems of such a technique is its use when it is known into what classes the objects to be discriminated must fall; the clustering technique tends to build its own categories which may be finer categories than are required by the problem, or coarser than required, or entirely separate categories than exist in the problem. Although this creates a problem when comparing accuracy of classification, it is felt that the clustering methods have the advantage when it is not known with certainty what classes are the proper ones for classification and, in particular, when there is no ground truth at all.

The fastest algorithm described is the maximum likelihood if implemented on an analog computer. Next would come table look-up implemented on a digital computer, and following that would come the likelihood ratio technique implemented on a digital computer. Probably slowest of all would be the clustering algorithm, which as we have seen must be implemented on a digital computer. Because the sequential technique has never been implemented we can only guess at how slow it might be; because of the memory requirement it, too, must be implemented on a digital computer. It seems likely that it would at least be slower than the likelihood ratio technique.

Of the techniques discussed, the likelihood ratio technique is the most frequently used, particularly with an analog computer. The clustering types of algorithm are also used extensively to classify data for which there is no ground truth; these are always digitally implemented. The table look-up algorithm is strictly experimental and is not widely used, while the sequential algorithm has not yet been implemented.

The speed of the clustering algorithm is limited by the amount of data that is clustered at a bite. There are also problems in ensuring that the data in consecutive bites are classified on a comparable basis as well as problems in selecting the parameters that the classification is to proceed under, such as the threshold parameter and the number of classification classes to be used.

The likelihood ratio requires training data in the near vicinity of the area where identification is to be made. It assumes that the data are normally distributed. Its speed is a function of the number of targets to be discriminated and the square of the number of features (channels) employed. It becomes extremely expensive in computation time as the number of channels increases. Implemented on an analog computer the expense of the algorithm is that the number of operational amplifiers increases as the square of the number of features and the set-up time for setting potentiometers goes up as the square of the number of features times the number of simultaneous targets.

The table look-up method suffers from memory problems. The memory requirement increases as 13^N where N is the number of channels. The method will work in the case of three channels, becomes awkward at four channels, and becomes impossible at more than four channels.

The sequential technique now works well in a manual mode but has not been implemented on a computer. If combined with multispectral data as suggested and implemented on a digital computer, it is probable that the biggest implementation problem would be the memory that would be required to store crop calendar information for comparison. Such a method would be accurate in classification if developed, but it is not certain that its implementation would be feasible.

All of the methods discussed here are possible candidates for a future earth resources processing facility. It appears likely that both analog and digital techniques will be used. The analog techniques could

be used for fast throughput of an inflexible, standard product, and the digital techniques could be used for flexible applications in which a variety of evolving techniques might be employed to process the data.

Processing Computer Requirements

The processing described in this section creates severe problems in terms of computer and buffering requirements. Figure 27 plots the required computer speed in operations per second as a function of the data rate into the computer in elements per second for various degrees of processing algorithm complexity. A number of conclusions can be drawn from this graph. It can be seen that there are nearly 2½ orders of magnitude difference in cost between a minimally complex algorithm and a fairly good one, represented by N = 3, m = 1 (N = number of channels, m = number of targets to be discriminated) and N = 12, m = 10, respectively. If we take 10^7 elements per second as the capacity of a processing facility, then the fastest existing computers could not process the load even using the trivial algorithm, whose value would in any case be doubtful, having N = 3 and m = 1. By 1985, the computer state-of-the-art would have advanced sufficiently that an algorithm with parameters N = 3 and m = 10 could be used, which would give barely adequate results.

Figure 28 shows the throughput requirement in elements per day as a function of the resolution in meters for a typical coverage mission. It is seen that with no data buffering, practically no useful data can be processed at 10 meters resolution, even with the 1985 digital computer capability, since the real time processing that would have to take place during the 237 second satellite pass is beyond the capability of any of the computers. If the data are buffered so that the load can be spread out over an entire day, the load for a ten meter resolution satellite can almost be handled by the fastest existing computers, can barely be handled by existing analog computers, and can be exceeded by 1985 digital computers. Before the data can be handled without buffering, it is necessary to relax the resolution requirement to 100 meters or more.

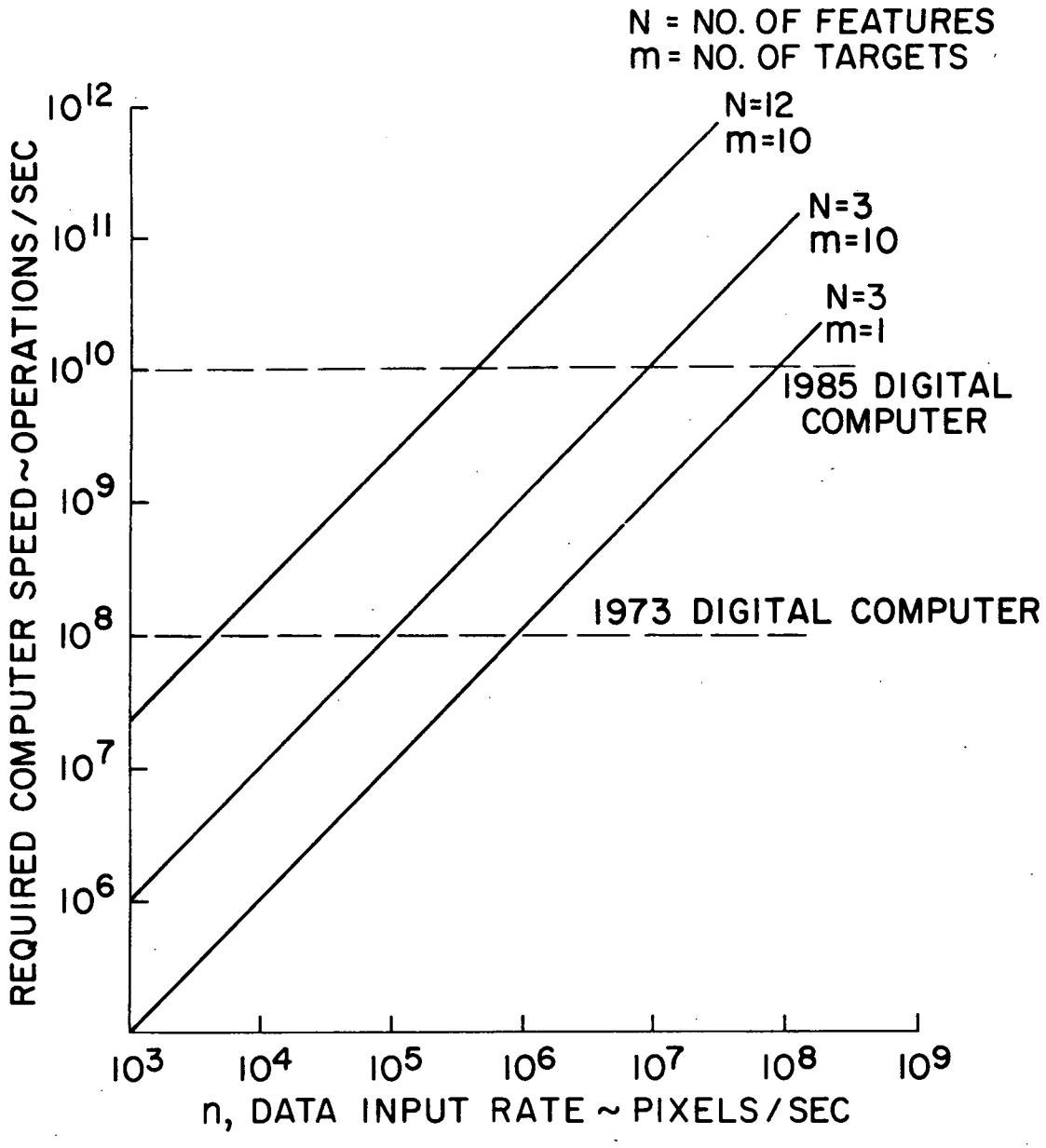

Figure 27. - Required Computer Speed Versus Input Rate for Various Classification Algorithms.

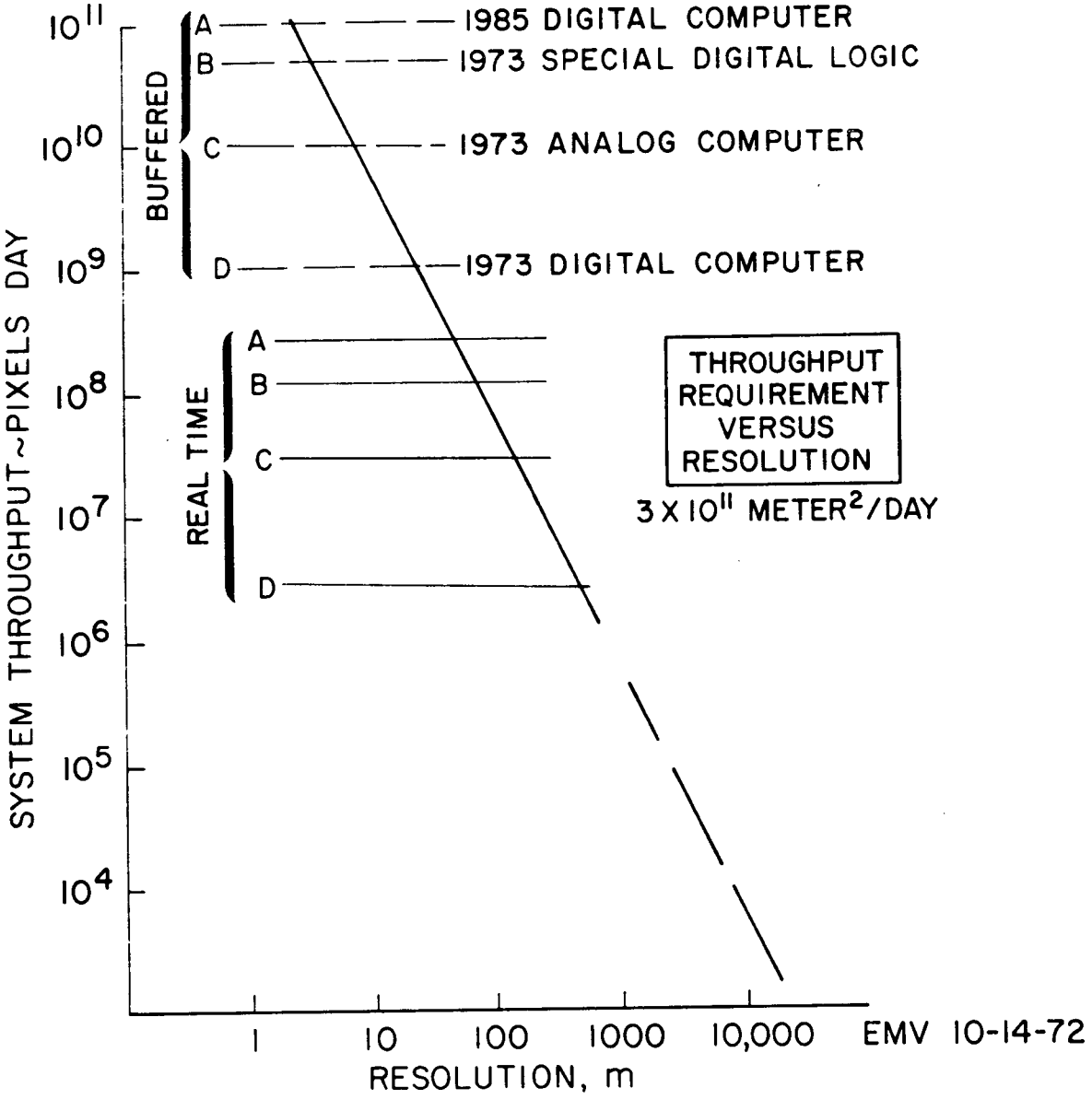

Figure 28. - Throughput Requirements Versus Resolution.

MACHINE CAPABILITIES

Implementation of the preprocessing and processing functions requires the use of modern data processing equipment. Indeed, the huge volume of data to be handled will require systems with throughput capabilities approaching or exceeding the present and projected state-of-the-art for computing equipment depending on the computing approach chosen. Since the 1930's, computers have increased in speed and capability from a few operations per second to present day systems that can operate at speeds up to more than 100×10^6 operations per second. This fantastic increase in computer speed and capability has, of course, come largely from the advances made in electronic technology in recent years. Along with these technology advancements have come changes in computer architecture from the normal sequential computer to parallel and stream processors in an attempt to increase speed.

While these great changes have been taking place in digital computing systems, analog computer technology has remained relatively static. The main reason for this situation, of course, is the fact that general purpose digital machines, because of their great flexibility and accuracy, have a significant advantage over analog machines in most applications. In situations, however, where a problem requiring a near real time solution is to be handled repetitively over a period of time, an analog machine is normally chosen. Where a large number of initial conditions must be readjusted each time, or a considerable number of bookkeeping type chores are required, a combination digital-analog machine, known as a hybrid computer, may be employed. This section of the paper will summarize the present and anticipated capabilities of digital, analog, and hybrid systems and associated peripheral devices.

Digital Computers

The organization of the conventional sequential computer is shown in figure 29. A counter in the control unit determines the address of

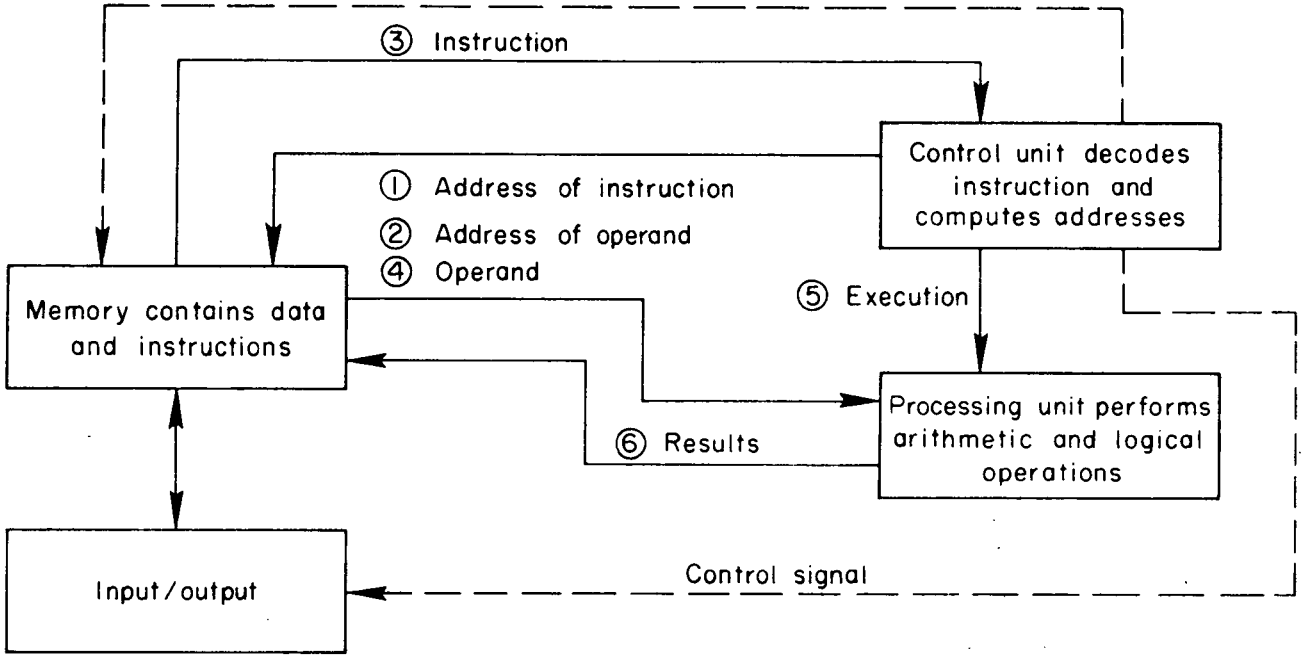

Figure 29. - Conventional Sequential Computer.

the next instruction in the sequence to be executed and transmits the address to the memory (1). The memory returns the instruction to the control unit (2). The instruction contains the address in the memory of the data (operand) on which an arithmetical or logical operation (also specified) is to be performed. This address is sent to the memory (3). The memory furnishes the selected operand to the processing unit (4). The control unit then transmits to the processor a sequence of electronic signals that contains the fine structure of the arithmetical or logical operation required by the program (5). The calculated result is then stored at a specified location in memory (6) for use in a subsequent operation or for conversion to printed form for the user of the machine. Some advanced computers carry out this entire sequence in a few millionths of a second; however, billions of repetitions may be needed to solve a complex problem. The sequential computer must await completion of one operation before the next one can be started.

The organization of one of the more recent parallel processor computer designs, that of Illiac IV, is shown in figure 30. This approach enables the control unit to arrange the operation of 64 processing elements, each with its own separate memory unit. Thus, large mathematical problems that are simply a repetition of a series of steps to be solved, can be handled simultaneously by a battery of independent processors. In the case of Illiac IV, each of the independent processing units operates faster than some single processors in advanced sequential computer systems.

A third computer design being used is the pipeline concept being employed in the central processor of the Texas Instruments ASC computer illustrated in figure 31. This example shows a "pipe" which performs an operation consisting of three separate and distinct steps. This operation can be performed on an operand by entering it in the pipe and collecting the results at the exit after some time has elapsed. The time required to perform the operation is equal to the sum of the individual steps. These steps, when separate and distinct as shown, allow the average operation time to be decreased by entering operands into the pipe

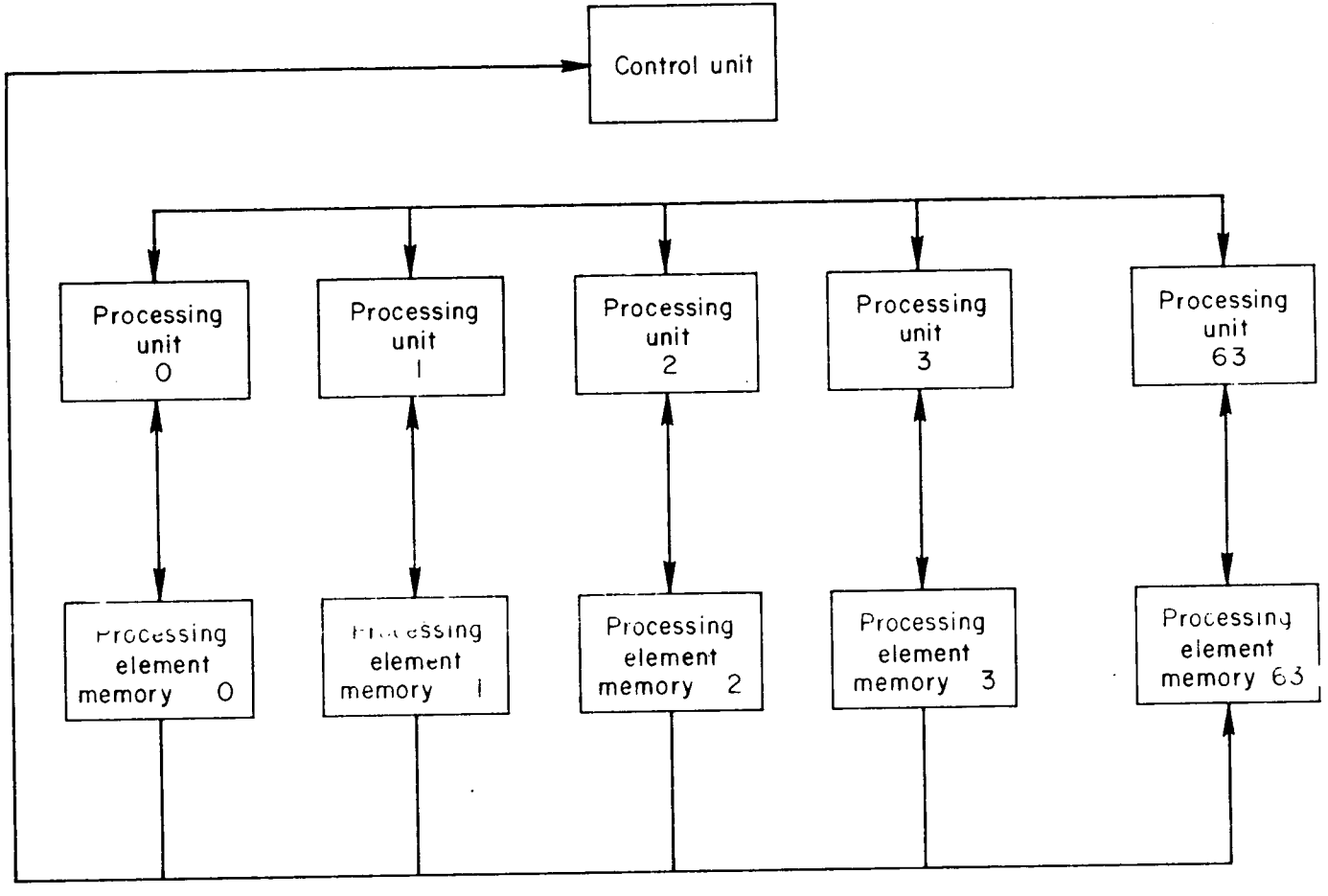

Figure 30. - Parallel Organization of Illiac IV.

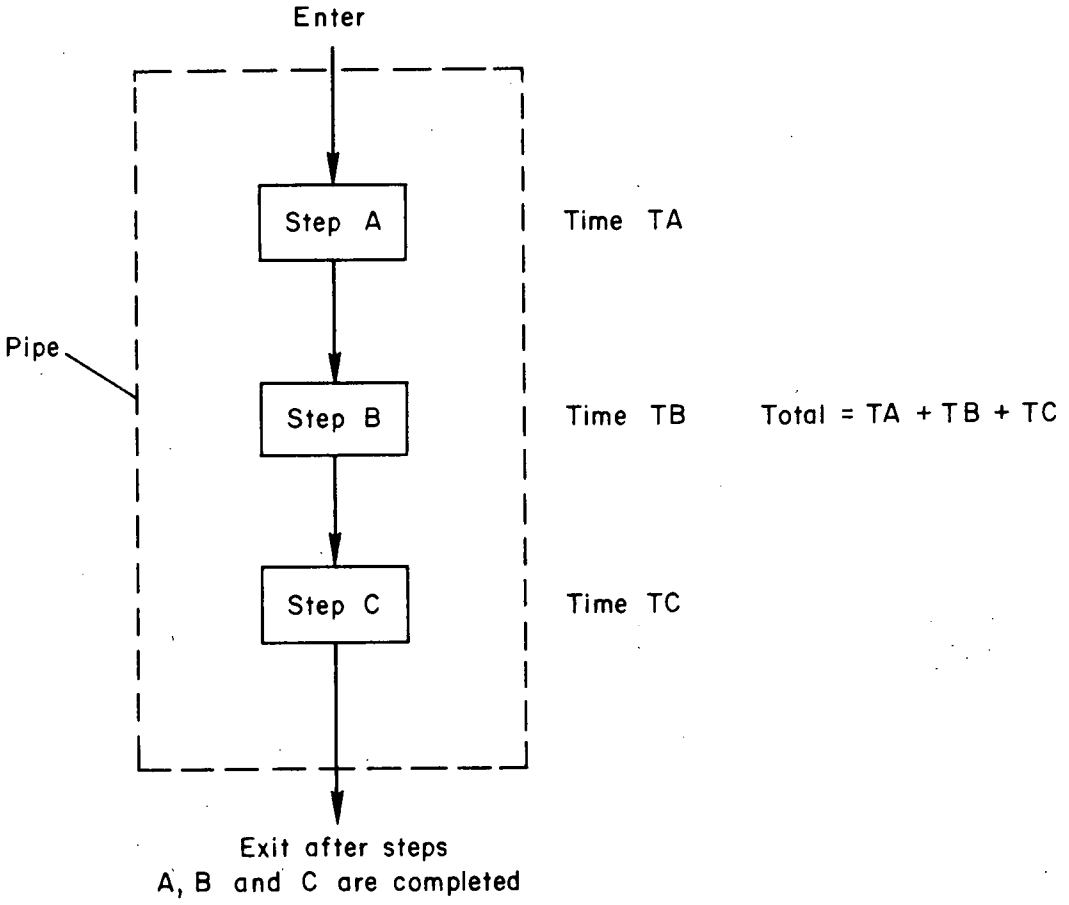

Figure 31. - Pipeline Concept.

so that different operands are at Steps A, B, and C simultaneously. If a long series of operands are routed through the pipe so that the "fill-up" and "empty" times are kept to a negligible amount, the average time required for an operation will be the average of TA+TB+TC/3.

This pipeline concept is used as a design because of its inherent ability to achieve high-speed operations on large volumes of well ordered data. If the data are arranged so that a large number of identical operations are required in sequence, the pipeline can be filled, achieving an average operation speed equal to the time required for only one section of the pipe. Problems that are well ordered and lend themselves to pipeline processing are vector and array processing.

Present Capability. - The present computer capability varies widely from computer to computer. As one tool to evaluate the performance of a computer, the three floating point operations were examined and are shown in table 8. The first seven representative computers are sequential computers while the remaining three are parallel or pipeline computers. From table 8 it may appear that the advanced design computers (Illiac IV, ASC, and STAR) operate at speeds that are little or no faster than some of the sequential computers; however, due to the hardware design such as pipeline computing in the ASC and STAR, and parallel processors as in the Illiac IV, the advanced computers are able to produce throughput several times faster than the conventional sequential computers. As can be seen, the Illiac IV's execution times for the functions shown are about equal to those of a CDC 6600 computer. One must remember, however, that these times are for each of the separate Illiac IV processors. There are 64 such processors in the Illiac working in parallel which could result in an increase in throughput of a factor approaching 64.

How much does this computer improvement cost? The computers shown in the table vary in price from $2-15 million per system. The three advanced designed computers have price tags several times higher due to the development cost that has been necessary to produce such systems. Current estimates on such systems run as high as $35 million. This cost

TABLE 8 COMPUTER COMPARISONS

Computer	Time to Get and Complete One Operation in μ sec.			Word Size (Bits)	Storage Nominal (Words)
	+	X	÷		
CDC 6600 (S)	.4	1.0	2.9	60	131 K
CDC 7600 (S)	.11	.137	.55	60	65 K
UNIVAC 1108 (S)	1.875	2.62	8.25	36	?
UNIVAC 1110 (S)	.9	1.65	5.3	36	?
IBM 360/67 (S)	5.4	6.8	10.0	32	64 K
IBM 370/165 (S)	.4	1.9	2.65	32	128 K
IBM 370/195 (S)	.11	.16	.59	32	?
CDC STAR (P)	.16	.32	1.4	60	524 K
Burrough ILLIAC IV (P)	.437	.625	3.56	64	128 K
Texas Inst. ASC (P)	.3	.24	.9	64	?

could, of course, be lower on a per computer cost if more units were produced and the development cost were amortized as it has been in the conventional computer systems. Each computer system will, of course, vary in cost as more or different peripheral equipment is added, be it an advanced or conventional system.

Future Capability. - A different look at computers may be to examine how the computer speeds have increased with time. Shown in figure 32 is a plot of computer speed versus year. Computer speeds have been increasing since the mid 1940's at a compound annual growth rate of 81 to 112 percent per year with the most rapid growth in the most recent period. If the growth curve is extended into the future, it appears that by the early 1980's an additional factor of 100 in speed might be expected to occur. Such a growth rate would bring the computer speeds to current foreseeable cooling limits; growth during the next decade will have to slow down as speed of light and information theory limits begin to be reached.

Modifying the Processing Problem. - There are a number of ways to scale the problem down to size so that the calculations can be done in real time. Faster computers may be used in the timescale we are referring to. Advantage may be taken of the 237 second per day duty cycle of the satellite, by buffering data during the time the satellite is over the area of interest and then later allowing the computer to process the data at a rate many times slower. A calculation could be made to determine a smaller subset of features than the set provided by the satellite, such reduction in dimensionality scaling the computing cost down as the square of the number of dimensions. A smaller number of possible targets may be used in the calculation for each area investigated by employing table look-up to identify a smaller subset of potential targets that may be found in the given area--this would bring a linear reduction in computation time.

Parallel processing and hard wiring might be employed. Parallel processing is ideally suited to this problem. As previously mentioned,

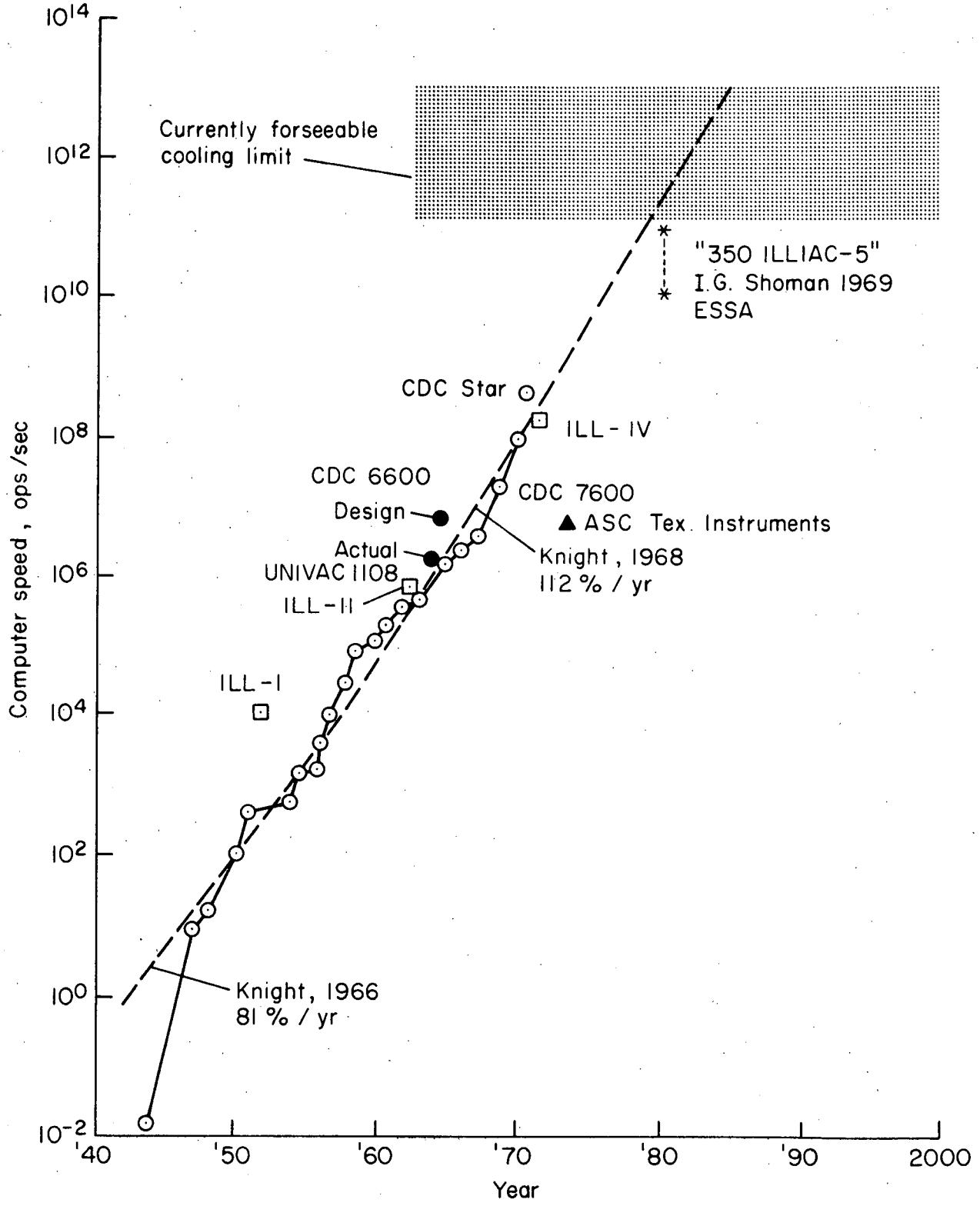

Figure 32. - Computer Speed Versus Year.

the Illiac IV employs 64 parallel processing units (P.E.'s) working simultaneously on portions of a problem. In the pattern recognition problem, 64 pixels could be processed simultaneously, the steps of each of the 64 matrix multiplication steps proceeding completely synchronously.

Hardwiring

Most general purpose computers are designed to handle a large range of problems and therefore may not provide the most optimum means of solving some data processing task. The data processing associated with the earth resources data will no doubt be the same basic computing procedure repeated many times day after day. This type of computing has been viewed with the thought of hardwiring in mind. Hardwiring is the design of the computer hardware to perform computing sequences that are done by computer programming on the general purpose computer. The resulting computer configuration then becomes a special purpose computer whose sole function is to process data by only one technique. When this special design is used, the results have been decreased computing time, for comparable jobs, of as much as two to three orders of magnitude.

This hardwiring, or special digital hardware design as it is sometimes called, has been implemented on some problems by the General Electric Company.[18] Two example problems run by General Electric are as follows.

Example A - ERTS. - Assume we want to classify each of the ERTS pixels into one of eight classes (arbitrary). Where the class decision boundaries are each represented by a four-dimensional Gaussian distribution, we have approximately 300 products (1,000 operations) per pixel which is 3×10^9 products or approximately 10^{10} operations per image.

ERTS-MSS 4 channel system		1×10^7 pixels
4 channels x 7 bits/channel		28 bits/pixel
Number of operations/pixel		1000
Total operations required		1×10^{10}
Total bits to Process		2.8×10^8
Processing times		
GP computer 360/75		250 minutes
GP computer - table look-up software		25 minutes
Special digital		16 seconds
Operations/second	special	6.25×10^8
Bits/sec processing rate	digital	18×10^6

Example B - 12 Channel Scanner. - In this example, each pixel is also classified into one of eight classes. Assuming each class is represented by a twelve-dimension Gaussian distribution, we have approximately 1,500 products (5,000 operations) per pixel, which is 15×10^9 products or about 5×10^{10} operations.

12 Channel scanner system	1×10^7 pixels
12 channels x 8 bits/channel approx.	100 bits/pixel
Number of operations/pixel	5,000
Total operations required	5×10^{10}
Total bits to process	1×10^9
Processing time:	
GP computer 360/75	1,250 minutes
GP computer with look-up software	125 minutes
Special digital	80 seconds
Operations/second	6.25×10^8
Bits/sec, processing rate	12×10^6

Close examination of the two General Electric example problems shows that as the number of operations per pixel increases from 1,000 to 5,000, the processing time is also increased by a factor of 5, that is, 16 to 80 seconds for the special digital hardware. The total number of operations, even though increased from case to case, yield the same constant operation per second value. If we now employ this information and substitute values that represent the operational system hypothesized in this study, we have an estimate of the required computation time. Using the same Gaussian distribution as in the two previous examples, we can construct the following:

Total pixels	2.0×10^9
Bits/pixel (12 channels x 8 bits) approx.	100 bits
Operations/pixel	5,000
Total operations required	1.0×10^{13}
Total bits to process	2.0×10^{11}
Processing time seconds	1.6×10^4
Operations/second	6.25×10^8
Bits/sec processing rate	$12. \times 10^6$

As stated above, the operations/second were fixed by the hardware at 6.25×10^8; therefore, the time to process 11×10^{12} operations would be 1.6×10^4 seconds, or slightly over four hours. This, however, would be a total data load from one 24-hour period collected using some sort of data buffering device enabling us to reach a data rate that could be handled by the computer system.

As can be seen from the first two examples, the processing time from general purpose computer to the special digital or hardwiring was a saving of computing time by almost three orders of magnitude, i.e., 15,000 seconds GP computer to 16 seconds special digital.

Now, how does the performance of the general purpose computer compare to its advertised operation per second rating? From the literature, we can see that 360/75 was introduced in late 1965 and was rated to perform at better than 3.5×10^6 ops/sec, but by simple examination of

the examples above, we see the 360/75 only performed about 6×10^5 ops/sec or 1/5 the speed one might expect. This comparison is made here to again emphasize how the performance of any computer can be varied by the nature of the problem being solved. It is therefore not enough to simply examine the operation rates of a computer. The special digital computer in the above examples solved the computational task at a rate of 6.25×10^8 ops/sec. How does this compare to computers from a newer generation? If we take the average advertised ops/sec ratings for several of the large present-day computers, remembering that they can be in error as shown above, we can construct table 9 to show how the computer capability of today will handle the projected 1×10^{13} computer operations load.

Special digital computers, although they are fast, do of course have some disadvantages. The general purpose computer, as stated earlier, can be programmed to solve many different problems. The method of solution and the computer program can be altered and changed with very few difficulties. These changes can certainly be made without the need to alter the electronic design of the machine. This is a real advantage for the general purpose computer as compared to the special digital which must be designed electronically to perform a predetermined task. There are some means to change the special digital computer but it will no doubt be a much more difficult task than changing the general purpose computer. There is, of course, a tradeoff that should be considered. Does the increased computation speed achieved by the special digital system offset the inflexible hardware design that limits program changes as compared to the general purpose computer? The data processing associated with the earth resources data does appear to be the variety of problem suited to hardwiring; it is the same computer process performed many times on a continuous flow of data.

The speed of the hardwire processor can be increased, thus giving a further advantage in speed over the general purpose computer but at an increase in the cost of the hardware. The cost for increasing speed was

TABLE 9 COMPUTER TIME TO PROCESS DAILY LOAD

G.P. Computer	Advertised op/sec x 10^6	Time to Compute 10^{13} op	
CDC 6600	3 - 5	3.3×10^6 sec 920 hours	2.0×10^6 sec 550 hours
CDC 7600	9 - 20	1.1×10^6 sec 305 hours	5.0×10^5 sec 139 hours
Univac 1108	2 - 4	5.0×10^6 sec 1390 hours	2.5×10^6 sec 695 hours
360/75	3 - 4	3.3×10^6 sec 920 hours	2.5×10^6 sec 695 hours
STAR	50 - 100	2.0×10^5 sec 55.5 hours	1.0×10^5 sec 27.8 hours
ILLIAC	100 - 200	1.0×10^5 sec 27.8 hours	5.0×10^4 sec 13.9 hours
ASC (Texas Inst)	50+	2.0×10^5 sec 55.5 hours	

From the above table, one can see that the only system approaching the ability to handle the load is the Illiac IV and it only at the maximum rated op/sec throughput and working at a better than 50 percent reliable uptime. The other computers range from 27.8 hours to a high of 920 hours, 1 to 38 days to process one 24-hour data collection.

estimated by the General Electric Company as cost ~ $(\text{speed})^{\frac{1}{2}}$. As an example, the approximate cost for processing 5×10^{10} operation in 40 seconds using the aforementioned formula would be, cost $\cong$ \$150,000 x $\sqrt{2}$ or about \$210,000.

The cost could and most probably would vary from problem to problem the same as the cost of programming a problem in software is very problem dependent.

Analog Computers

The analog computer is still another way one might solve the data processing associated with earth resources data. Improvement with time has not been as impressive as that of the digital but the analog could offer some very strong points. Current analog systems are able to obtain an accuracy of .01 percent or 1 part in 10^4, much poorer than digital systems but adequate for many problems. Since the analog computer works in a true parallel processing mode, it is a high speed machine. However, comparisons with digital machines are difficult. One acceptable gauge used to rate digital computers is the number of "operations per second" a particular computer can perform. A similar gauge for the analog computer would be useful.

The state-of-the-art of the analog computers has been estimated to be a system with a bandwidth of up to 1.0 MHz (10^6 cycles/sec) and having up to 600 operational amplifiers. What does this mean in terms of computing power? In an attempt to arrive at an answer to the question, we attempt to establish a measure whereby the digital and analog computers can be compared. Each operational amplifier (op amp) performs a particular function by the amplification of input voltages. The function may be to add, substract, or perform an integration on some voltages which, in turn, represents numbers. This function performed by the op amp is equivalent to some given number of digital computer operations; the precise number is a variable and will be very problem-dependent. A most

conservative comparison of digital to analog would be to say each operational amplifier is equivalent to one digital computer operation. We will, however, still choose to use a 1:1 ratio* as the lower base for comparison purposes. A simple addition on the digital computer takes from three to five machine cycles so we will choose a 5:1 ratio as representative of the add function. This 5:1 ratio is reasonable as the analog can add several numbers at the same time while the add time for a digital is for two numbers only. Now, if we take the number of op amps on a system and multiply them by the bandwidth BW (the times per second each op amp performs its function) and again multiply by the assigned ratios from above, we can get what we call analog equivalent operations per second (AEOS), i.e., AEOS = op amps x BW x Ratio*.

As we vary the bandwidth, the AEOS is shown on figure 33 for two ratios and two systems having 100 and 300 op amps. The straight lines show the advertised op/second for some large scale computers and for the three advance design computers. No time has been assumed for input or output to any of the systems, analog or digital, which could slow both systems by a considerable amount. The comparison shown in figure 33 will vary depending on the problem being solved and the ratio assigned to each op amp in the system. One must also remember that the analog system will not produce the accuracy that the digital computer can achieve, but the tradeoff of speed and accuracy could be made depending upon the problem being solved.

The cost of a system must also be a consideration when tradeoffs are made. We have already given an idea of the cost for large digital systems; now the cost of an analog system must be addressed. The cost of an analog system exclusive of the input-output choices depends on system characteristics such as accuracy, bandwidth and the number of op amps in the system. The current estimate based on the use of integrated

---

* Digital/analog operations ratio

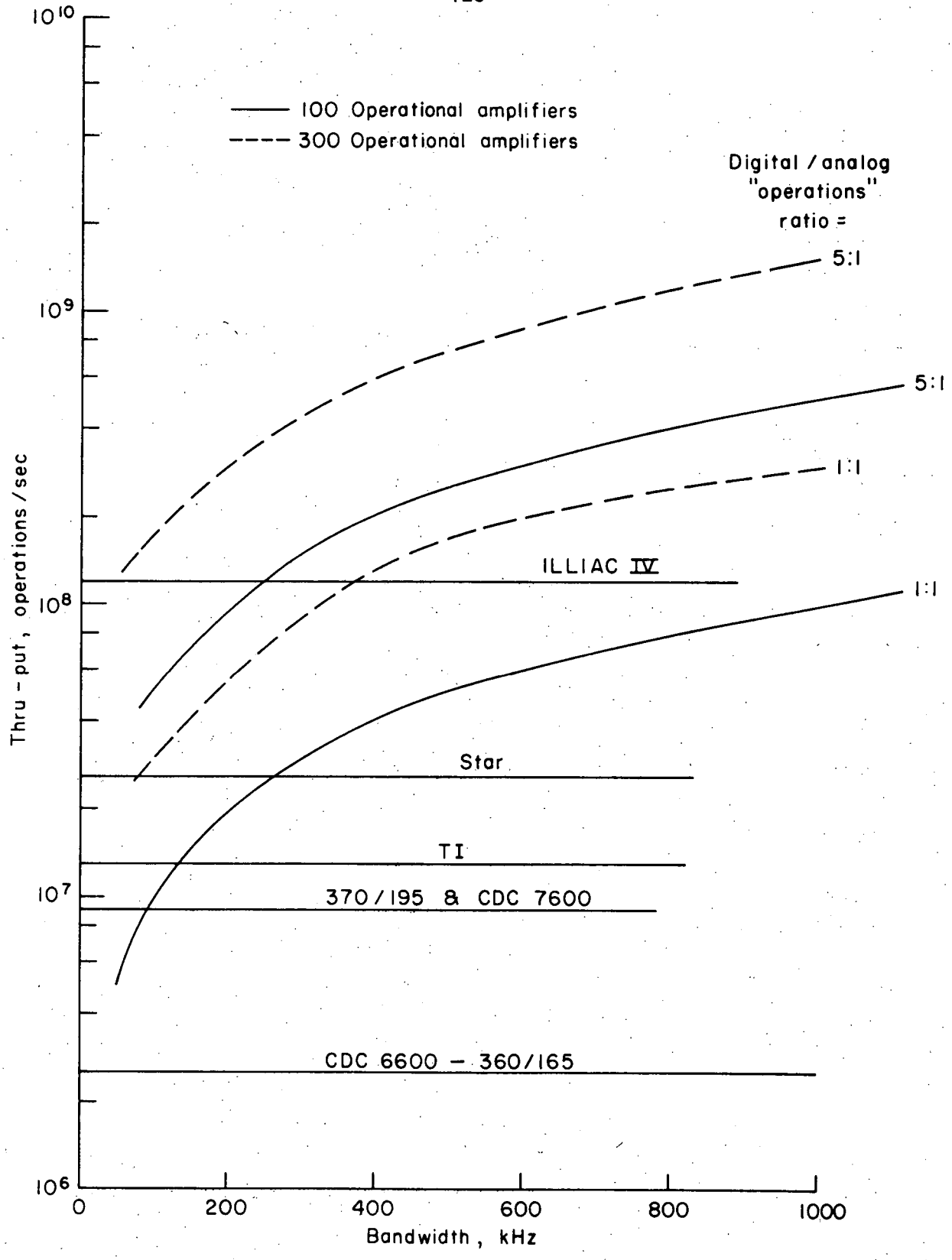

Figure 33. - Analog-Digital Performance.

circuit operational amplifiers and associated technology advances is $1,000 per operational amplifier contained on the system. This estimate includes all auxiliary hardware and interconnections associated with a complete analog computer system.

Output Requirements

The task that must now be examined is what do we do with the processed data and how large will the data load be that must be handled. The large volume of input data, 2×10^{11} bits per day, represents a covered area, as stated earlier in the report, of 1,850 km by 185 km with a resolution of 10 meters. The 2×10^{11} bits of data are also collected in the very short time period of about 4.5 minutes, yielding a data rate of approximately 10^9 bits/second. This, of course, is at the state-of-the-art limit for data rates of recording devices and might require, even in 1980, some sort of buffering device to make use of multiple recorders to handle the data as it comes from the satellite. It is for this reason, then, that we choose in the discussion of output requirements to consider that the data have been recorded from the satellite and will be fed to the computer in a continuous buffered data stream of about 2×10^6 bits/second. The real volume of output data will, of course, vary depending on how the data is processed and to what degree of refinement it is handled. The simplest form of processing would, of course, be that of enhancement of the original data and the volume of output would be in the same order of magnitude as that of the input stream, i.e., 2.2×10^6 bits/sec or 3.6×10^5 char/sec. (1 char. = 6 bits)

Shown on figure 34 are some examples of the type of output that might be required from earth resources data. Along with the output type, the estimate of the data volume is shown and is seen to vary two orders of magnitude from the most simple form to the requirement of handling the enhanced data for archival storage. In an operational system, not all of the output requirements would be needed at a single location, but

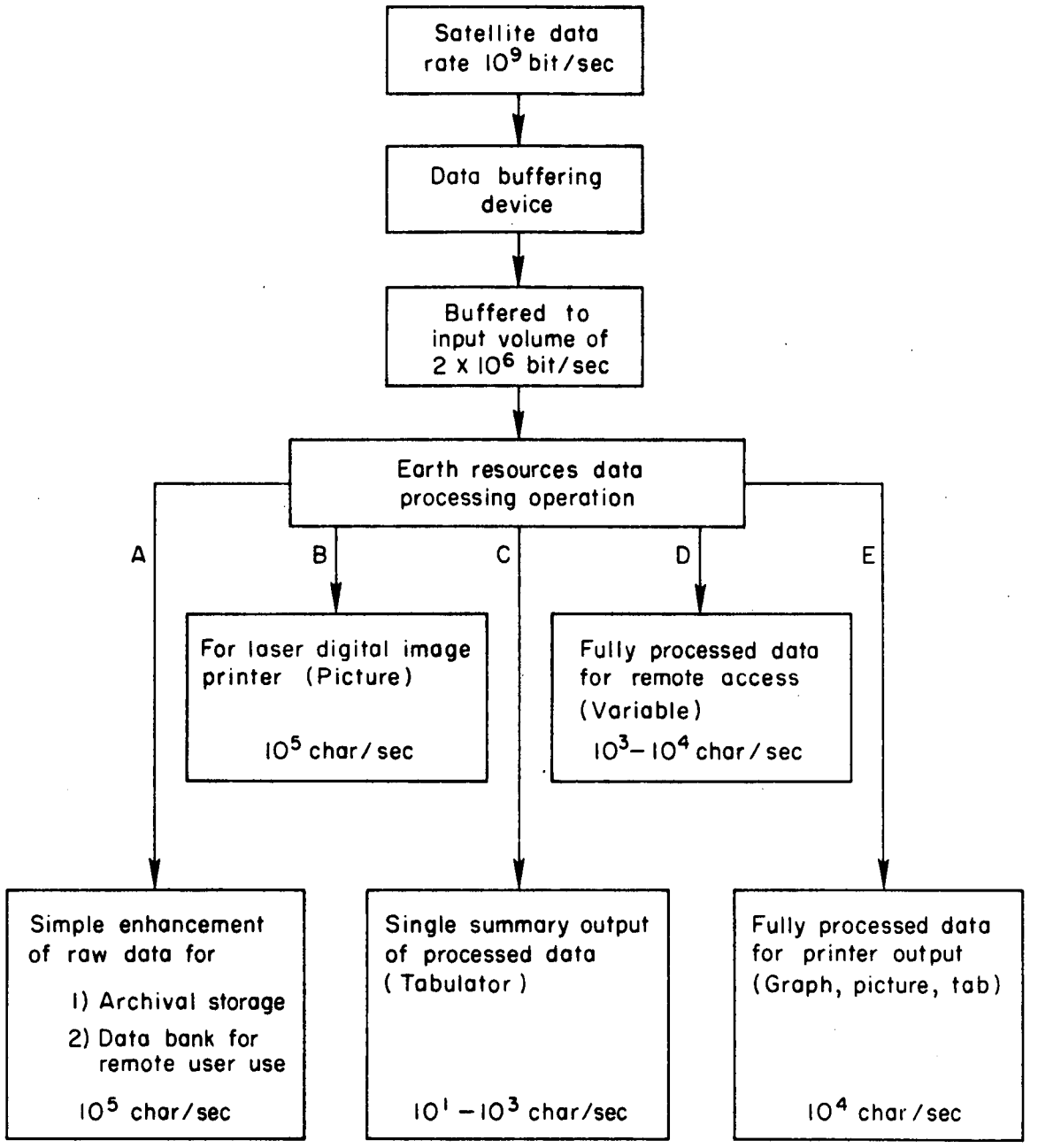

Figure 34. - Possible Data System Output Volumes.

the figure serves to illustrate the possible options that would be available to the user.

Addressing the first type requirement, that of simple summary data, the standard impact printer should adequately handle this data volume. The impact printer state-of-the-art hardware will produce up to 3,000 lines a minute output which represents approximately 7×10^3 characters per second.

When fully processed and ready for output, the data sufficient to produce graphs, tabulated tables, or printer generated pictures would produce a data volume such as represented by the Block E on figure 34. The impact printers would fall short of the required capability, and a printer of the electrostatic variety would be required to handle a data volume of 10^4 char/sec.

Mead DIJIT Printing System. - A new fast versatile printer is being developed and refined by the Mead Corporation. This printer will print on any ordinary paper using tiny ink jets to form the printed images at a very high resolution and at speeds of up to 1.5×10^5 char/second. The system will also be able to produce output in color or tones of grey if desired. This type printer could be used to produce multi-color pictures of the covered area as an output from the earth resources data. The system is called the "Mead DIJIT Printing System" and seems to offer real promise for earth resources data handling.

Laser Digital Image Printer. - Another new printer technology with the possibility of the printout being in several different colors or various tones is the Laser Digital Image Printer. The amount of data for this type output is an order of magnitude above the others discussed thus far. The volume is about the same as the enhanced raw data and requires a device able to handle 10^5 char/sec+ and able to print the data to a very fine resolution. The technology to produce such a picture image does, in fact, exist and is in a working model. The printer is referred to as a "Laser Digital Image Printer" and has the ability to print on film at the rate of 10^5 to 10^7 char/sec, well within the required data

rate shown on figure 34. The output that is for storage to be called upon at some later date presents an additional consideration other than simply the speed with which the data can be recorded. The size or number of storage devices needed to hold the large amounts of data should be considered. This problem is addressed to some extent later in the report. The volume of data expected, 10^5 char/sec, could be handled by the laser mass memory system, or by magnetic tapes using the video recording technique.

Table 10 shows several types of output devices and the char/sec capability range for each. The output types from figure 34 have been assigned to the best suited device, and as can be seen, seem to be adequate to handle the projected data volume.

Table 10 Output Media Performance

Type	Output Char/Sec	Could Handle Output from Fig. 34
Impact Printer	7×10^3	C
Non-Impact Printers	6×10^4 to 1.5×10^5	D,E
Laser Digital Image Printer	2×10^5 to 3×10^7	A,B
Magnetic Tape	6×10^4 to 3×10^5	C,D,E
Video Magnetic Tape	to 8×10^6	A
Laser Store Unicon	3.4×10^6	A,B,C,D,E

Memories and Storage Systems

The large data volume and the resulting data transfer rate will require new and larger storage systems. Some of the storage will be required as on-line storage for use with the computer system and still other space may be needed for archival storage. There are several methods being looked at by the computer community that could provide the needed storage capabilities. Some of the technologies under consideration are electron beam recording, magnetic recording, holographic recording, and magnetic-optic recording. All of these technologies have one thing in common; that is, the recording is done on a surface whether it be rigid

or flexible. As a representative example of present and future technology, we will examine the magnetic recording and electron beam recording devices in some detail.

Magnetic Tape. - The old standby, and long a tool used by the computer industry to store data, interface with the computer, and drive off-line equipment, magnetic tape has undergone much improvement. Since the first magnetic tape was introduced, improvements have come in both quality of the product and increases in the packing densities of the data. The data density that is obtainable with magnetic tape today ranges from 200 to 2,400 BPI (bits per inch). The amount of data that can be stored on a standard reel of tape 2,400 feet in length, one-half inch wide depends on such things as size of record gaps, end of file marks and how they are used or placed on the tape. For the sake of comparison, we will assume the maximum storage possible is simply the number of inches of tape in a reel times the density of the tape, or $C = DxS$ where C is the tape capacity, D is the tape density, and S is the inches of tape in a reel. The value of C for a standard 2,400 feet by one-half inch reel would be 5.7×10^6 bits to 6.9×10^7 bits for tape densities of 200-2,400 BPI, respectively. When the capacity of a 2,400 BPI tape is compared to the data volume of 2×10^{11} bits that must be stored, the required number of tapes needed would be about 3,000. This would hold the data for one 24-hour period and a similar amount would be needed each day for several days before the tape could be reused for data recording.

Laser Recording. - Laser recording is now a reality with the memory system built by Precision Instruments for the Illiac IV computer system. This technology also offers a very fast, high density recording system. Laser recording may offer a stiff challenge to other recording methods in the future and must be examined here. The recording is done on strips of polyester material containing a thin magnetic coating. Each strip is 4.75 by 31.25 inches and can be encoded by a single laser beam to contain 1.6×10^9 bits of data. This represents a packing density of about 10^7 bits per square inch. The total Unicon memory, as it is called for the Illiac IV, consists of eighteen data strip packs. Each data strip

pack holds twenty-five data strips for a total storage capacity of $.7 \times 10^{12}$ bits of user information. Any one or all of the data packs can be replaced as more memory is needed to hold new data. The data load being considered, 2×10^{11} bits per day, would therefore use around 125 data strips per day to store the data. The data transfer rate for this system is 3.4×10^{6} bits/sec on each channel.

There is an additional consideration when considering laser beam recording of data. The data once recorded then becomes a read only memory and the recording material cannot be erased and reused for future recording.

Optical Memory. - When thinking about 2×10^{11} bits of data that must be stored and used each day, the need for a large capacity system with read/write and erase ability is apparent. Recently, the Laser Computer Corporation[19,20] announced their development of a 10 trillion bit optical memory (designated the LC-100 optical mass memory) with some very outstanding performance specifications. The read/write cycle requires only 20 nanoseconds while the full read/write/erase cycle requires 40 nanoseconds. This large, very fast memory system has non-destructive read out capability. The reported input/output data rate is 500 megabits/second while achieving an error rate of 1 bit in 10^{9}. The heart of the memory is said to be lithium niobate deposited on a thin film along with six other materials on a substrate of glass. A memory plane four feet on a side, consisting of ten panels of lithium niobate is required to achieve 10^{13} bits of storage capacity. Even this large size still has a storage density of approximately 10^{7} bits/mm^2, much higher than many optical memory devices. The cost of bit storage is said to be 2×10^{-5} cents per bit.

This current technology compares favorably with the projected future requirements. The data rate of 5×10^{8} bits/second is only a factor of 2 less than the 10^{9} bits/second required to record the data directly from the satellite. The 10^{13} bits capacity would hold a full 50 days of data before the memory would need to be erased and reused again. This system is reported to be in testing at several locations

as well as within the company. This system offers real promise for handling the vast amount of data to be generated within an operational program.

As a matter of additional interest it is reported that LCC is planning to develop an "atomic lattice" memory with a storage capacity of 10^{40} bits. It is expected to be developed by 1980.

Video Recording. - The need for very high densities and large capacities for memories and storage devices have brought forth other recording techniques. One such method is the video digital recording first introduced in 1961. This recording technique has been used by Ampex in their TBM (Terabit memory) system. The system uses standard two-inch wide video tape on 3,600 foot reels and is able to achieve a packing density of 1.5×10^6 bits per square inch. This density enables storage of about 10^{11} bits on each reel of magnetic tape. This recording technique would accommodate the earth resources data load of 2×10^{11} bits per day on only two reels of tape. As with standard magnetic tape discussed above, the data volume would be needed for several days before the tape could be reused for recording. The present data transfer rate for this system is 4.5 megabits per channel. The reuse of the tape would be possible for several years as the video tape is reported to have a useful lifetime of 2,000 recordings.

One such memory system has been built and delivered to a user at this date, so this is a present day technology.

SYSTEM REQUIREMENTS

In the previous sections of this paper, we have presented some general considerations of data load, generic system concepts, preprocessing and processing software, and computers. These are the elements of an ultimate system design. But the various characteristics and limitations and advantages alone are not enough to define a system. Of the many system alternatives that have been discussed, limitations of the system elements will rule out several even without performing extensive

tradeoffs. There remain other system alternatives that will require more extensive tradeoffs before the selection of the ultimate system can be made. Finally, it becomes possible and necessary to apply user requirements criteria to take the system selection properly. In this paper, we limit ourselves to discussing the factors that will affect system design in a major way and omit detailed tradeoffs among system parameters.

System Assumptions

To obtain an initial estimate of the size and scope of the data processing task, we may ignore the actual geographical placement of system elements and argue that certain functions will be required of all alternatives. Thus, the minimum system consists of the elements that will just process the requisite data at the speed required. In other words, from an equipment standpoint, a centralized facility may well be the least costly approach. In any event, the amount of equipment required for a given mechanization cannot be less than that required in a centralized facility. Obviously, there are considerations other than the amount of equipment required and the cost of that equipment. These other considerations which relate to the overall system effectiveness and, ultimately, to overall (as opposed to equipment) system cost including manpower are the basic reasons for considering various alternative schemes. Nonetheless, a very good estimate of the minimum equipment required can be obtained from a consideration of the centralized approach. This philosophy is based on the assumption that manual processing is not feasible for an operational system producing multispectral data.

Thus, the assumption of the centralized approach will not restrict the generality of the conclusions and it is a reasonable assumption that will simplify the analysis. Similarly, because the extremely high data rate of 2×10^{11} bits per day that we have assumed limits our flexibility in system design at many points, we may expect that a brief consideration

of the system design process may enable us to make some additional simplifying assumptions that will permit a first tentative look at a scenario of an earth resources processing system.

Before proceeding with this process, let us look at the user and his requirements in more detail than was given in the data characteristics section above.

User Requirements

The user and his requirements are an essential element in planning a system. Unfortunately, at the time that a system is planned, it is usually impossible to identify the user. And their requirements, even *after* users have been identified, are subject to a great deal of guesswork. Therefore, some assumptions must be made about such user requirements. These assumptions should be as representative as possible so that when the user community begins to form, and user requirements data become more firm, it will not require major reorientations or redesign of the system.

This report: (1) postulates an overall user requirement typical of the 1985 time frame; (2) outlines the basic requirements of a system to satisfy such a requirement on a broad, parametric basis that temporarily ignores the details of the user requirements; (3) introduces refinements in the user requirement as they become critical to the analysis.

There are several basic types of user requirements that must be introduced. Each requirement has a different set of system implications, all requirements are compatible, though possibly only at considerable expense, and it is impossible to predict the relative mixture of the several requirements present in the user community. Therefore, we analyze the system under all sets of user requirements, discuss corresponding system requirements, and analyze the compromises that might be effected to merge the requirements under one system.

The first type of user receives data on a specific geographical area at varying levels of detail that cannot easily be predicted in advance. The second type of user is interested in drawing broad conclusions, perhaps even of national interest, from a sampling of data. The third type of user is the farmer, farm goods manufacturer, etc., who ultimately must use the data. The fourth type of user is a "data broker" or commercial user who processes data into a convenient form for others at a price.

"Interactive" User. - The first type of user might interact with the system as follows. He receives a standard product from the system which might include a listing of areas in which corn blight has been detected during the last pass of the satellite. The user wishes to review his own area of interest for corn blight, perhaps even a specific farm that had corn blight last year. He uses his remote console and calls for a display of his own area, perhaps using symbols looked up in a glossary type handbook. If the computer has already processed the data for the given area, these recognition data are displayed on a scope. If the computer has not already processed the desired information, it does so at this time, and as before the display is made. Using a light pen, he marks out an area for enlargement. The enlarged area is scanned. Perhaps several special routines in the system library are called out and used to process the data, such as a masking routine that will suppress all recognition display characters except those for corn blight, so that greater visibility may be obtained.

Perhaps it is decided that a false color presentation of the data is required to make the required decision. Then the computer outputs through a false-color output device, the print from the device is placed on a color facsimile transmitter, and in a few minutes it is received over ordinary narrow band transmission lines on the facsimile receiver at the user facility. If time is not of the essence, it and similar hard copy products could simply be mailed. A high speed printer is available for the printing of grey scale maps as well, or even color at the more sophisticated user stations. It is decided that the standard library routines in the control facility are not adequate and it is desirable to

perform additional tests on the data. If the system is so designed, it will be possible to have either recognition data (processed) or raw data (unprocessed) available for transmission to the user facility on demand. If time is relatively unimportant, then the data can be listed on a magnetic tape and mailed to the user. If data timeliness is important, it may be desirable to have high speed lines leased to those facilities where computational equipment is available. The data are transmitted to the distant computer where it is processed using special algorithms.

On the basis of the special processing, together with the other data products just described, the user decides that on the next pass of the satellite he needs a more complete set of data analyzed; perhaps he would like the output from all of the satellite's multispectral scanner channels rather than those that were automatically selected by the system as containing the most information (based on some test such as interclass divergence). He makes his order for such additional data using his remote console again. The system then makes the appropriate modification to its report-generating program and the additional data and analysis will automatically be made following the next processing. If enough special requests for data are made, the system will automatically notify the appropriate decision-makers. These managers will use such reports of trends in data requests, outbreaks of infestation, unusual phenomena spotted by the system routines, etc., in their decisions as to where to dispatch aircraft. The aircraft are used routinely to monitor the performance of the satellite and in multistage sampling, etc., and so are available for contingency missions such as described.

"Multistage Sampling" User. - The second type of user uses a basically different product. He is interested in such broad questions as the total yield of wheat for a given region or the nation; the fraction of the total crop subject to blight or infestation; the total volume of timber in a specific region. His interest is still a management interest, like that of the first type of user, but he is interested in planning rather than in minute specific decisions related to small areas. His mode of interaction

with the system can be decribed as follows. The satellite multispectral data are processed and from it are drawn, partly automatically, and partly by a staff of analysts, a set of statements about the current state of resource variables of interest to planners, trends in such data, and perhaps predictions related to certain regions and to the total land mass. At each pass of the satellite, such predictions are updated and transmitted to the users. On the basis of these data, it is decided that a specific item of information concerning a given resource is required. Only an estimate of this value is needed; say, an estimate of the amount of wheat under cultivation.

It is decided that for the purposes of the investigation, it is important to determine this number with x percent accuracy. A sampling plan, probably a multistage sampling plan, is drawn up, using a computerized sampling routine which will specify the exact satellite data to be employed, the precise location of sampling sites to which to send remote sensing aircraft and the type of data to take there, and the precise location of ground sampling sites that are to be used in conjunction with the satellite and aircraft data. The aircraft itself might perform its sampling at several altitude/resolution combinations in a refinement of this technique. In another refinement of the technique, the aircraft might perform preliminary analysis onboard or through relay of data to the central facility to determine conditionally the additional sampling sites that might be selected, based on preliminary results. The aircraft personnel select the final ground truth sites in accordance with guidelines as to numbers and general locations printed out by the computer, together with subjective feelings of representativeness of the ground sites. The ground observations are made and transmitted to the central facility, using narrow band lines such as telephone lines, and the satellite, aircraft, and ground truth data are processed. The input data are used, through a stratified sampling formula, to estimate the total amount of wheat under cultivation to within the required accuracy earlier established.

The "Ultimate User". - The third type of user, the ultimate consumer of data, would either receive data directly from the system, as a fixed data product, or perhaps through a commercial processor type of user, to be described later. Such a user, for example, might be a farmer who wishes to receive regular reports on the status of his own plantings, estimates of crop health, yield, stresses, dryness, etc. Another type of ultimate user might be a manufacturer of fertilizer, who wishes regular status reports on the progress of planting by region so that he can estimate the amounts of fertilizer to stock in various warehouse locations, for planning shipping schedules, etc. This information might be received via radio, television, or cable television as special reports that the ultimate user could learn to interpret for his own needs; or they might be made available, for example, through a County Agent, who would maintain a remote terminal capable of receiving a standard data product, perhaps with the capability to request special supplementary data to aid in answering non-standard questions. It is probable that this kind of user would be connected to another user of the interactive type, either a user within the system or a commercial user. In this case, by satisfying the interactive user the requirements of the ultimate user would be automatically satisfied as far as the design of the data handling facility is concerned. For example, one might think of an interactive station at the state level connected with separate relatively noninteractive terminals, rather inexpensive in design, at locations such as County Agent offices. From the County Agent's office the information could be given directly to the ultimate user in the case of special requests, or disseminated to such users via the media in the case of a standard regional product.

The Commercial User. - From the standpoint of the system, the commercial user would be exactly like the first two types of user, either strategic or tactical, and would make no additional demands on the system design. An interconnection link would have to be provided, and system specifications would have to be supplied so that the commercial user could either purchase terminal facilities like those of the first two types of user or develop terminals to the given specifications.

Differences Between User Types

As far as system design is concerned, the above-mentioned four user types amount to only two basic classes of user--the "tactical" and the "strategic". There are significant differences between these two broad classes of user. The first, or "tactical", user required the system to have on hand very detailed data in large quantities that he could scan to perform his decision making function. He was not interested in data in the aggregate, but in specific areas and items. And yet a very large amount of information had to be provided for him to sort through. It is as if such a user would prefer to use even the satellite interactively, to search for and acquire precisely the data that he wants when he wants it and with a set of prescribed sensors. If synchronous earth observations satellites ever become available, it may be possible to completely satisfy such users, but in the meantime their needs can only be closely approached by collecting large masses of data and then providing the user with the tools for sorting this data out interactively, almost as a crude simulation of using the satellite itself interactively.

The second class of user is interested in data in the aggregate. For example, there are algorithms to be used with low resolution data that do not attempt to recognize specific materials within the pixel, but rather make an estimate of the relative proportion of specified materials that are present. Such an algorithm would not be useful to the first type of user, who wishes to know precise information concerning precisely located geographical regions; but the second type of user could make use of information from such an algorithm, since he is interested only in highly aggregative data on large geographical areas.

The second type of user is easier for the system to satisfy. It is only necessary to have a very large amount of broad satellite data, perhaps not with as good a resolution; a much smaller but much more detailed set of aircraft information; and a selected but rather small set of ground truth data.

The first user requires that large quantities of data be stored. For this user, it would be desirable to store raw data, but it is conceivable that the job could be done by storing only processed data, with consequent memory savings. The amount of time that the data would need to be stored is an important design decision, but one whose resolution would depend on the development of a user community with the capability of defining such requirements. It seems reasonable to assume that the data should be stored for at least 100 days. Before it is purged from the erasable memory, it should probably be stored photographically where it could be retrieved and rescanned (obviously with some loss of quality) should it ever again be required.

The second type of user does not have a strong requirement for interactive search capability. However, if this capability is provided for the first type of user, it could also be used by the second type of user, probably to good advantage.

System Design Implications of User Type. - In general, it appears that the requirements of the first type of user dominate the system design. The design should be interactive; raw data should be stored for a considerable period of time; all channels of the multispectral scanner must be saved; a high resolution system must be employed. Although keeping open other options, we can concentrate on the options dictated by the first kind of user.

System Design Criteria

Now that we have discussed the effect of user type on system design, although briefly, we shall proceed with other considerations that dictate design choices in the system. These factors have to do primarily with data transmission, storage, and processing, both on the ground and in the spacecraft. For the following discussion refer to figure 35.

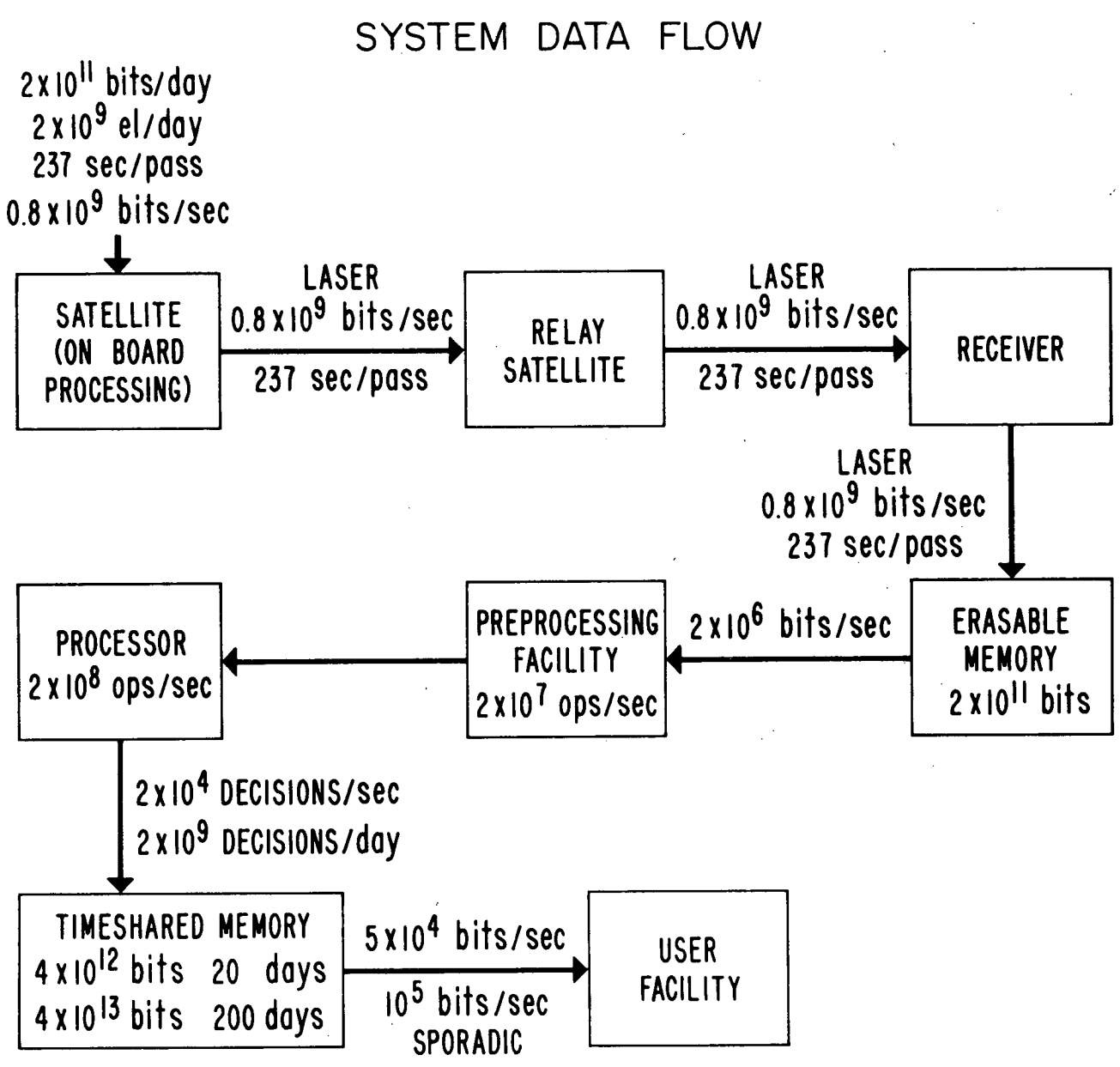

Figure 35. - System Data Flow.

Spacecraft Storage and Transmission Criteria. - There probably will be no onboard storage of data or onboard processing. The data rate of 2×10^{11} bits per day, assuming all channels are transmitted, is probably too high to permit buffering onboard. This is equivalent to a data rate of 0.8×10^{9} bits/sec during the satellite pass over the area of interest. The data would be relayed from the earth resources satellite in sun synchronous orbit to the data relay satellite in synchronous stationary orbit probably via a laser system. The technology developments of interest here would be a laser system for space capable of modulating a gigabit laser beam, which is just barely possible on experimental ground links in 1972, and the space-to-space link technology which has not yet been demonstrated. Future studies should examine the question of onboard storage in more detail than was possible here. In particular, it may be possible to: (1) determine which subset of channels to transmit; (2) compress the data on the order of 10 times before transmission. Both onboard buffering and onboard computation should be analyzed.

The next step is the transmission of the data from the synchronous relay satellite to the central command and control facility. The most important problem with such a link is outage due to clouds and storms. If outages in the data are not to be permitted to occur, then there would have to be a storage buffer onboard either the earth resources satellite or the data relay satellite capable of handling 2×10^{11} bits or, as another alternative, multiple ground stations capable of recording and re-relaying the data when the central station is not cloud-covered. This is because it is possible to experience an outage of 250 seconds which would eliminate an entire day's output of the earth resources satellite if it occurred at the wrong time.

The laser transmission mode also probably means that the data would be transmitted to a central facility at least once before going to the user. If data were to go directly to the user, there would have to be some provision for splitting off the data pertinent to each user while still in space. In addition, each user would have to have a laser reception facility capable of handling the entire bandwidth of one gigabit during the few seconds each day when pertinent transmission was occurring.

In view of these considerations, it seems that the data would first go to a central facility. Then, after preliminary processing, demultiplexing, etc., the data could be retransmitted via the satellite to the user in a narrow band transmission scheme if it were desired to do so.

Erasable Memory Buffer. - The data will have to be transmitted from the command and control facility to the mission control center via laser beam. This probably means that the two facilities should be located closely together or co-located, as in the ERTS case, so that the transmission distance can be kept small. The data enter the erasable memory in the mission control center at 0.8×10^9 bits per second. Investigation will have to be made to identify erasable memories with read-in rates this high; multiple parallel memories may be required. A laser memory may be able to handle this input by 1985. This memory should have a capacity of several days' data in case of computer difficulties and to permit buffering the fluctuating data load satisfactorily; this would mean perhaps 5×10^{11} bits. Memories that can hold one hundred times that load are available.

Processing Facility. - Data is read out of the erasable memory at the rate of 2×10^6 bits per second, which is well within the state-of-the-art for transmission but is extremely high as a memory readout rate. The processor must accept data at 2×10^6 bits per second. Depending on the type of data formatting, this translates into a decision processing speed of 2×10^4 decisions per second. An analog processor using a simple algorithm such as maximum likelihood ratio testing can process at such a rate or slightly faster. Because of down time considerations for training the classifier in the analog mode, it will be desirable to have available the fastest analog processor that the 1985 state-of-the-art permits and in addition hybridization must be used to reduce the training time to an absolute minimum. Because of the lack of flexibility dictated by the analog approach, we must still keep open the option of digital processors. At a decision rate of 2×10^4 decisions per second, a good digital algorithm might require 10^3 computer operations per decision, or a total rate of 2×10^7 operations per second. Parallel processors can already achieve such rates as we have pointed out elsewhere.

The problem is inherently a parallel processing problem, since any number of pixels may be operated on independently. This is not true, however, for the clustering algorithms and certain other algorithms such as boundary recognition, which require simultaneous interrelated handling of many pixels. Because the analog processor would be cheap and inflexible, while the digital processor would be expensive but flexible, it is possible that the system would be based on both types of system. The digital processor would be employed to give the system growth potential and the analog (or rather hybrid) processor would be used to provide a reliable, steady output of a limited product. It seems unlikely that the inflexible hybrid processor would ever be used as the heart of a sophisticated processing system. The existence of a digital computer in the system will create a natural environment for several dozen processing routines in addition to the pacing one of pattern recognition, although many of these may reside in a separate computer.

One of the system criteria for which we have an option has to do with tradeoffs between storage and computation. We can either classify each pixel as it arrives and store the classified data or store all the raw data and classify them as they are called for by the user (or on some preagreed basis). The first scheme would require a larger computer. The second scheme would require an exceptionally large memory--to store all the raw data for 100 days would require a memory of 2×10^{13} bits. As an assumption merely to allow us to continue with the analysis, perhaps it is reasonable to assume that the resource management type user would again dominate the system design and that a capability for storing raw data for 100 days will be provided. It should be noted that memory capabilities of 5×10^{13} bits are available now and are increasing so that such a memory will be possible in the 1980's; costs per bit are decreasing sufficiently rapidly that such a memory might be available for a price of approximately $4 million.

Accessible Memory for Processed Data. - The output of the classifier amounts to 2×10^9 pixels per day. For each pixel, something on the order of five bits of data would be required for classification labeling, so that 10^{10} bits of classified data per day would need to be stored in

an accessible memory. Assuming the terabit type of memory, this would permit storing 100 days of classification data. One option might be to store such classification data on a field-by-field basis rather than a pixel-by-pixel basis. This would not be adequate for certain of the user requirements that we described under the resource management type of user, but might be a workable compromise. This would cut the data storage requirement to 10^7 bits per day or a 10^9 bit memory for 100 days.* If it turned out that only a fraction of the users were interested in data at such a level of detail, perhaps only their share of the data could be stored on a pixel-by-pixel basis with the others on a field-by-field basis. This latter memory needs to be machine accessible but need not be erasable. It is felt that such a memory would be within the state-of-the-art in 1985. A timesharing environment such as that provided with the IBM 360/67 should provide an ideal way of using this memory. In addition, the bulk of the special processing routines of the system might reside on the 360/67 or its equivalent, thus reserving the large scale parallel processing machine for the difficult recognition tasks.

Of course, no one knows how many users the system may ultimately have. In order to completely specify the system, this number should be known. To perform the analysis, it is necessary to make some arbitrary assumptions.

The number of users may start out as low as thirty in the initial stages and finally increase to as many as 100, distributed through the agricultural area plus several other locations such as Washington, D.C., where decisions regarding the separate agricultural areas will be made. The system should provide for 100 users at the outset. The precise number will, of course, depend on the way the user community evolves, but 100 may serve for planning purposes. This means that the timesharing computer should be able to handle 100 users; a computer equivalent to the IBM 360/67 could presumably do so, but some consideration

---

* This reduction is based on a field size of 0.1 km^2 so that there are 1,000 pixels per field at 10 meters resolution. Such a field is smaller than the average field size.

should be given to providing a larger computer that can comfortably handle such a load without undue processing response delays.

User Terminal Equipment. - A scenario of user terminal usage will be presented to outline its requirements. The user equipment should include a keyboard for such things as ordering up data prior to need, requesting data to be transmitted for local processing, etc. There should be a high speed printer for digital maps on a quick look basis. There should be a color facsimile receiver for transmission of false color maps and space photographic imagery. There should be a high speed data transmission device for data to be transmitted at high speed from the central facility directly to a memory or tape at the user facility. Some users will receive bulk data via mailed tapes. Certain standard data products will go to the user over these transmission facilities on a regular basis, and others will be called up as required using the remote terminal.

When data are transmitted to local facilities, it will be via a relatively narrow band line, a group of twelve channels at most. The central facility, however, will have to provide for the simultaneous transfer of perhaps ten to twenty such groups over separate wires but from the same central memory. This could perhaps best be done by providing a high speed readout from the memory onto tape. The tapes would then be placed in the transmission devices, of which there would be one for each communication output line (10-20) and the transmissions of all the tapes could take place simultaneously. We will assume that only twenty of the 100 users will need to have raw data transmitted to them for local processing--these would be primarily the scientific workers who would be working with experimental processing techniques. The system as a whole would transmit much of its data over facsimile, much as the National Weather Service does with its WEFAX network.

Some of the data product will be sent on an *a priori* schedule rather than being requested interactively. There might be a recognition map together with predictions, statistical resumes of findings of the processor, etc. The pictorial data could be broken into local fragments by

the processor and transmission could be by facsimile, while the textual material might be transmitted by the high speed printer. There might be machine-generated special reports where timeliness was not so important and these might be simply mailed. Where critical, they might be sent either by facsimile or high speed printer.

It should be possible to use the same remote terminal to interact with the central computer in a timesharing mode. The user would have the option of performing manipulations of his data remotely, using the central computer from his remote console, without the data ever being transmitted to the local facility. Of course, for special processing functions that had not been provided for in the central facility library there would still be a need for transmission of the data to the local facility for processing locally.

The design of the data handling facility for the information generation rate we have specified would be a complex process requiring many tradeoffs to resolve the design decisions properly. In the foregoing, we have attempted to encapsulate this process briefly in a scenario so that some of the important design considerations could be brought into focus without actually doing these extensive tradeoffs. An attempt should be made to start a preliminary design process and perform some of the major tradeoffs so that the system characteristics can be developed in greater detail than is possible here. We have shown that the data rate requirements of an advanced earth resources satellite would result in pushing the state-of-the-art rather hard in several areas in the corresponding data handling facility. It does appear, however, that a data handling facility capable of handling 2×10^{11} bits per day may at least be feasible.

REFERENCES

1. *Agricultural Statistics 1971.* U.S. Department of Agriculture.

2. Shahrokhi, F.; Rhudy, J.P.: Remote Sensing Techniques in Evaluating Earth Resources--A Study of Remote Sensing for Southeastern U.S., *Remote Sensing of Earth Resources,* Vol. 1, Shahrokhi, F., Ed., Conference on Earth Resources Observation and Information Analysis Systems, Tullahoma, TN; March 13-14, 1972.

3. Baldwin, C. J., *et al.: Functional Design for Operational Earth Resources Ground Data Processing.* Final Report NAS9-12336, Earth Resources Technology Office, Applied Technology Lab, Houston Operations, TRW Systems, Houston, TX; September 15, 1972.

4. Stratton, A. J., *et al.: Earth Resources Objectives and Measurement Requirements.* NASA Working Paper MS-70-1; April 20, 1970.

5. Planning Research Corporation: *A Study of the Economic Benefit and Implications of Space Station Operations.* Vol. 1, Summary, PRC R-1218.

6. Deerwester, Jerry M., *et al.: Data Acquisition Systems for Operational Earth Observations Missions.* NASA TM X-62,107; February 1972.

7. Fu, K. S., *et al.: Information Processing of Remotely Sensed Agricultural Data.* Proceedings of the IEEE, Vol. 57, No. 4; April 1969.

8. Marshall, R. E.; and Kriegler, F. J.: *An Operational Multispectral Surveys System.* Infrared and Optics Laboratory, Willow Run Laboratories, IST, University of Michigan, Ann Arbor; 1971.

9. Eppler, W. G.; Helmke, C. A.; and Evans, R. H.: *Table Look-up Approach to Pattern Recognition.* Lockheed Electronics Co.; 1972.

10. Colwell, Robert N.: Significance of the Results Obtained in Relation to User Requirements. *Monitoring Earth Resources for Aircraft and Spacecraft.* NASA SP-275; 1971.

11. Carnegie, D. M.; Pettinger, L. R.; Hay, C. M.: Analysis of Earth Resources in the Phoenix, Arizona Area. *Monitoring Earth Resources from Aircraft and Spacecraft.* R. N. Colwell, ed., NASA SP-275; 1971.

12. Fu, K. S.: Statistical Pattern Recognition. *Adaptive Learning and Pattern Recognition Systems.* J. M. Mendel and K. S. Fu; Academic 1970.

13. Wacker, A. G.; Landgrebe, D. A.: The Minimum Distance Approach to Classification, Purdue Univ. Information Note 100771, The Laboratory for Remote Sensing; 1971.

14. Wacker, A. G.: A Cluster Approach to Finding Spatial Boundaries in Multispectral Imagery. LARS Information Note 122969, Purdue Univ.; 1971.

15. Su, M. Y.; Pooley, J. C.; Hand, C. G.: *Statistical Algorithms and Computer Programs for Analysis of Multispectral Observations.* NASA CR-63182, December 1970.

16. Kuehn, R. L.; Omberg, E. R.; Forry, G.: *Processing of Images Transmitted from Earth Resources Observation Satellites.* Colloque International, L'Espace et la Communication, Paris, 1971.

17. Kriegler, F. J.: *Implicit Determination of Multispectral Scanner Data Variation over Extended Areas.* Inst. of Science and Tech., Willow Run Laboratories, Univ. of Michigan, 1971.

18. Economy, R.: Private communication, General Electric Co., Valley Forge Space Center, Philadelphia, PA; September 19, 1972.

19. Anon: *It's the Real Thing: 10 Trillion Bit Optical Memory.* Electro-Optical System Design; October 1972.

20. Martin, Bill: Private communication, Laser Computer Corp., Anaheim, CA; November 1972.