Improving Landsat Scene Selection Systems

Kenneth McGwire

Abstract

With the goal of periodic global coverage, Landsat 7 will create a wide variety of opportunities for the use of high-resolution data at regional and global scales. However, significant improvements to the information systems for Landsat scene selection are necessary to support such wide-area applications. There may be significant phenological differences to be captured in such wide-area surveys, and trade-offs between image timing and quality must be considered on a scene-by-scene basis. Ancillary environmental information can be integrated into the scene selection to ensure the most phenologically consistent image acquisitions. Also, by defining a weighting function, users can substantially automate the scene selection process for extensive and complex compilations of Landsat scenes.

Introduction

Improvements in information systems for Landsat scene selection are necessary to support the its ever-broadening application to regional and global applications. The data acquisition schedule for Landsat 7 of 100 scenes/day has been stated as having the capability to provide periodic, terrestrial coverage of the globe (NASA, 1994). However, actually compiling such a coverage requires consideration of complex, spatially varying constraints in phenology, cloud cover, and data availability which are not adequately addressed by the current capabilities of the current Earth Observing System Data and Information System (EOS-DIS) (NASA, 1986) or the Global Land Information System (GLIS) (USGS, 1997), nor are such capabilities defined in recent documents for the EOS-DIS Core System (Clinard, 1995). While the Landsat 7 data system will provide inexpensive terrestrial coverage of the globe, mechanisms still must be developed to ensure that scientists can efficiently compile the most phenologically consistent sets of imagery possible for very large land areas. Such a system should allow interactive, semi-automated scene selection which integrates image metadata and ancillary data sources, such as annual profiles of the normalized difference vegetation index (NDVI) for each path/row as derived from the Advanced Very High Resolution Radiometer (AVHRR) sensor. The Humid Tropical Forests project (Skole and Tucker, 1993) and the North American Landscape Characterization project associated with NASA's Landsat Pathfinder program have put together some of the most extensive coverages of Landsat data ever compiled. However, the objectives of these projects are fairly modest in comparison to the ambitious goal of coordinating a global coverage.

The 16-day orbital period of Landsat provides significant challenges for integrating Landsat with global Earth observing strategies. This period provides much less flexibility in dealing with cloud cover than the near-daily coverage by sensors which are designed for coarse-scale, global observations such as AVHRR or MODIS (Moderate Resolution Imaging Spectrometer). The chance existence of cloud cover in consecutive acquisitions of Landsat data may significantly affect the ability to develop phenologically consistent sets of imagery, especially because cloud cover and phenological state are generally correlated. Further, phenological states will vary throughout a region and from year to year. If one is interested in biophysical characterization at the peak of "greenness," how does one compensate for variations in phenological timing throughout a large region when using static queries based on date? An improved scene selection system could explicitly identify the user's priorities for temporal and phenological consistency and apply these priorities in a rapid and objective manner using ancillary data sources.

Landsat 7 data management will be coordinated through EOS-DIS, with support for Landsat 7 provided by Version B of the EOS-DIS Core System (ECS). The Interface Control Document (Clinard, 1995) and Interface Requirements Document for the ECS (Caplan, 1994) detail the technical needs for searches of EOS-DIS inventories. However, these documents do not address the issue of improving the user's ability to explore the archives through graphical and/or analytical synthesis of ancillary environmental data. We need to go beyond simply determining which individual datasets or granules meet the constraints of a query. We need to identify the user's specific objectives in order to determine the best combination of images. By quantifying user priorities, a system could objectively prioritize scene acquisitions throughout the Landsat mission, identifying gaps in the global coverage and indicating the relative magnitude of this problem for each path/row.

Improved Scene Selection Methods

A semi-automated approach to Landsat scene selection, which includes graphical representations of ancillary datasets and scene attributes, is needed to maximize phenological and spectral consistency between images while minimizing cloud cover. Such a system would feature

- user interaction with image metadata and co-registered ancillary data sources to guide the specification of scene selection parameters;
- flexible scene selection based on ancillary datasets and user-established priorities; and
- interactive, graphical displays of scene characteristics and data richness in the Landsat archives relative to user-specified criteria.

Without these capabilities, scientists wishing to use Landsat in regional or global studies will be forced into an unnecessarily laborious exercise of browsing thousands of database records. EOS-DIS Version 0 and GLIS allow the selection of scenes by cloud cover and time of year, but there is no support for relating acquisition dates to the actual phenological

Photogrammetric Engineering & Remote Sensing,
Vol. 64, No. 7, July 1998, pp. 717–722.

0099-1112/98/6407–717$3.00/0

Desert Research Institute, Biological Sciences Center, 7010 Dandini Blvd., Reno, NV 89512 (kenm@dri.edu).

state. The listing of database records resulting from massive queries are not presented in an efficient graphical manner, nor are they presorted for the user based on expected suitability. The current system is essentially designed for queries in which relatively small numbers of scenes are considered independently from each other.

Despite the availability of swath processing (user-defined segments of an orbital track) with Landsat 7, the Level 0R product definition will report image metadata using path/rows of the previous world reference system (WRS-2). The WRS path/rows provide a systematic sampling framework for parameters such as image quality and cloud cover which are not meaningful at the scale of an entire swath. *A priori* processing could intersect the area encompassed by each WRS path/row with other spatial datasets, such as monthly values of the global vegetation index (GVI) derived from the AVHRR sensor. Parameters like

- month of peak GVI,
- maximum value of monthly GVI,
- minimum value of monthly GVI, and
- seasonality (e.g., annual range of GVI / average annual GVI)

could be presented graphically for each path/row and the time period being considered. Other environmental variables such as mean monthly precipitation or maps of ecological zones could be added to this scheme. Integrated access to coarse-scale remote sensing, climate, and map data would allow a user to consider the acceptability of scenes which do not directly match the desired characteristics (e.g., acceptability of scene substitution given cloud cover). This would be especially critical for arid and semi-arid environments which are subject to highly variable patterns of precipitation and productivity. Phenological patterns can vary greatly over large areas. Integration of ancillary environmental datasets with the scene selection process would allow collections of scenes to be compiled based on actual environmental status rather than approximate date ranges, and all the generalizations a simple date-based approach would force on unique subregions.

Figure 1 displays the month of maximum GVI for 1984 which was derived from the NASA AVHRR Pathfinder 1 degree datasets (James, 1994). Figure 2 displays the absolute difference in months of peak greenness between 1984 and 1985. In some cases, the differences highlighted in Figure 2 correspond to areas where intra-annual phenological variation is actually insignificant relative to the artifacts of the GVI compositing process (e.g., tropical rain forest or regions with minimal vegetative cover). Such regions might be differentiated from areas of significant phenological change by dividing the intra-annual range of GVI values by the annual average. Figure 3 displays this ratio using mean values for 1982-1993, with lighter tones indicating areas of higher value and hence greater phenological contrast (ocean masked as white). Those areas which do display meaningful differences in the month of peak GVI might be responding to inter-annual climate variations, such as the El Niño Southern Oscillation (ENSO). Such inter-annual variability should definitely be factored into the scene selection process. However, the ability to do this on a scene-by-scene basis is currently quite limited and labor intensive.

In addition to bringing ancillary environmental information into the scene selection process, users should be able to specify weighting functions which represent their priorities with respect to scene acquisition characteristics in order to automatically sort database listings. A basic multiple-year query of a single path/row in GLIS can result in several dozen scenes. Users typically must browse several consecutive web pages, searching for those entries which have the best combination of cloud cover, image quality, and date. While this process will not be particularly problematic for selecting small sets of scenes, developing a scoring method to automatically identify the best candidates could save large amounts of time and effort for wide-area studies. A weighting scheme leads the way to automating much of the scene selection process and provides the basis for creating much more useful graphical summaries of data richness and quality throughout the region of interest.

Automating Regional Landsat Scene Selection

The recommendations outlined above were implemented using programs written in the PERL and C programming lan-

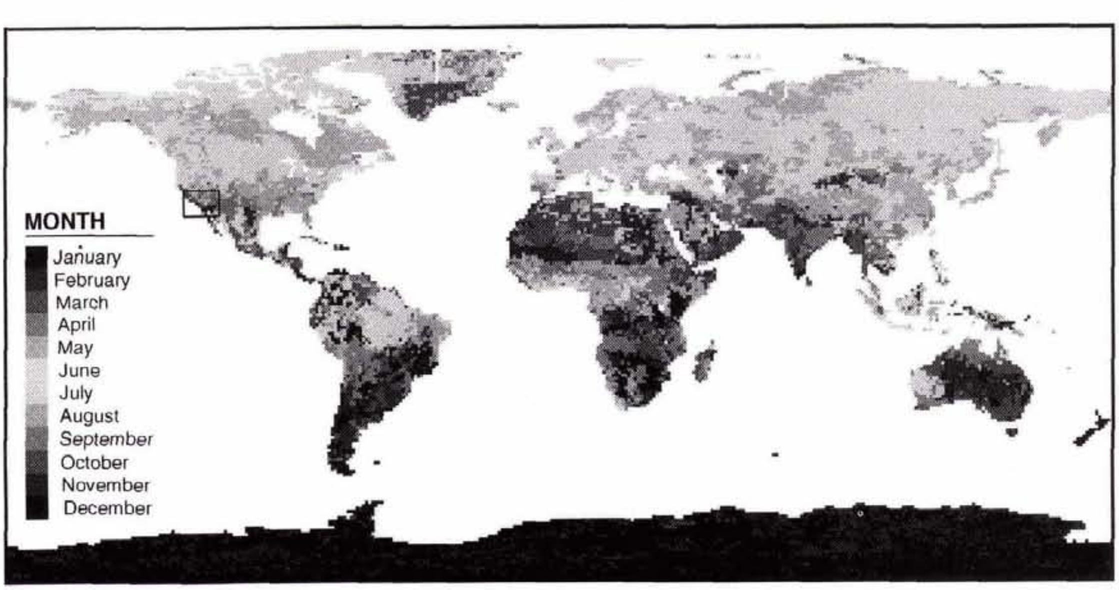

Figure 1. Month of maximum GVI for 1984.

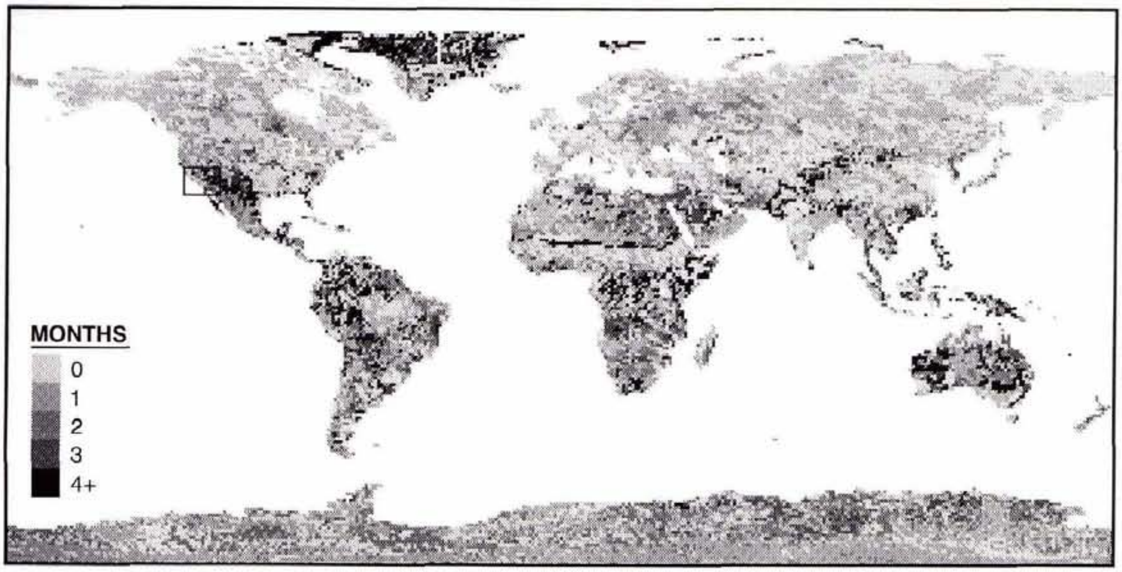

Figure 2. Difference in month of peak GVI (1984 vs. 1985).

guages. A PERL script which uses text files to specify input parameters was written to automatically perform GLIS queries, rather than relying on manual interaction in an Internet web browser. Programs written in C extracted information on phenology from AVHRR GVI data in order to create input files for the PERL script. The GVI data were taken from the AVHRR Pathfinder one-degree dataset covering 1982 through 1994. The example presented here automatically develops a list of candidate images covering all of Nevada, Arizona, Utah, and inland California for the mid-1980s in which

- cloud cover is minimized,
- image quality is maximized,
- scenes acquired during the month of peak of GVI are given priority, and
- acquisitions close to the target year of 1985 are given preference.

This example covers a relatively large region (69 path/rows) experiencing both intra- and inter-annual variation in climate and phenology (boxed area in Figures 1, 2, and 3).

Path/rows covering the study area were identified in a GIS database of the Landsat world reference system (WRS-2). The month having the maximum GVI value was identified for each path/row (using center-point latitude/longitude) and for each year from 1983 to 1987. Table 1 shows the high degree of intra-annual variability in the timing of peak GVI throughout the study area, highlighting the difficulties that would be encountered in making phenologically meaningful selections in a manual mode. This variability arises from climatic dif-

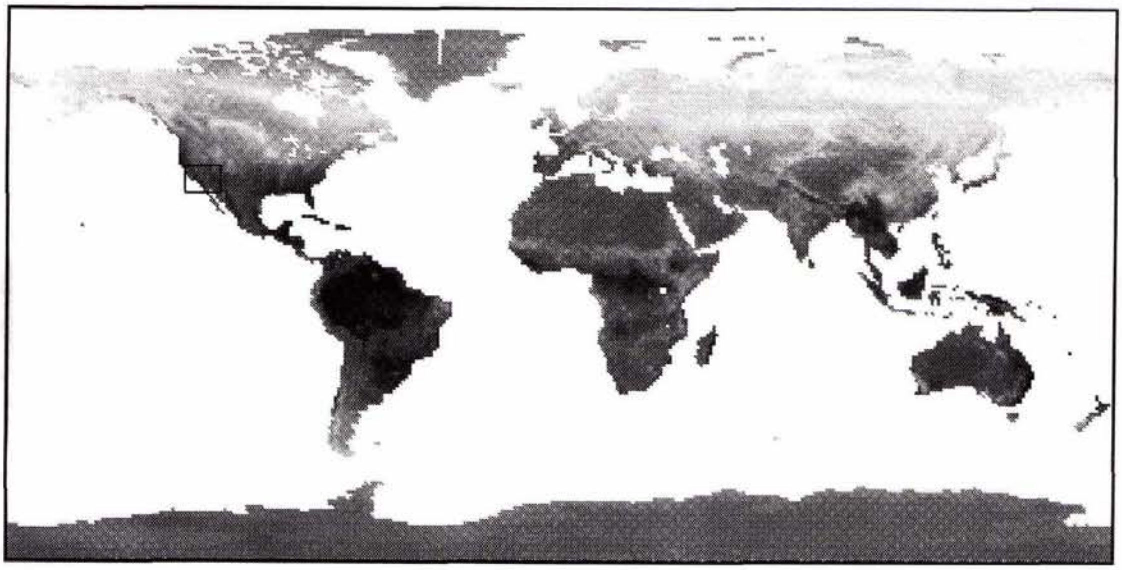

Figure 3. (intra-annual range of GVI) / (annual average GVI).

TABLE 1. STATISTICS FOR MONTH OF PEAK GVI (JAN. = 1, DEC. = 12)

	1983	1984	1985	1986	1987
Earliest Peak	3	2	3	2	4
Latest Peak	11	11	10	11	9
Mean Peak	7.0	7.2	6.0	6.5	6.5
Std. Dev.	1.9	2.2	1.4	2.0	1.6

ferences between frontal, monsoonal, orographic, coastal, and continental regimes which are found throughout the four States. Interestingly, 1987 shows a very different growing season corresponding to an El Niño event which produced anomalously high rainfall in the region. Such inter-annual events would be very difficult to compensate for without integrating ancillary environmental data into the scene selection process.

Automated queries were submitted to GLIS for a period covering three months around the date of peak GVI for each path/row and year, resulting in 1,134 candidate Landsat images. On retrieval, a weighting function was then applied to the scene metadata to automatically select a candidate image for each path/row. This weighting function created a normalized score (0.0 to 1.0) using the following formula:

$$\frac{\sum_i w_i{*}s_i}{\sum_i w_i{*}\max(s_i)},$$

where i is the attribute index, w_i is the weight of i^{th} attribute, and s_i is the standardized suitability index of i^{th} attribute. The weights and suitability indices for selected image attributes are presented in Table 2. The target date for the suitability index relating to the time of year was the middle (day 15) of the month of peak GVI. The weighting scheme in Table 2 was developed by testing the scoring system on a few selected path/rows.

Figure 4 displays the scores for the best scene in each path/row throughout the test area. This graphical presentation could be used to determine which images must be critically evaluated or to indicate where alternate data sources might be considered (e.g., SPOT or IRS data). The weighting function used here placed the greatest emphasis on selecting scenes with the lowest possible cloud cover, as indicated by the field for estimated percent cloud cover in each scene which is returned from the GLIS database query. However, it can be seen in Figure 5 that the distribution of cloud-cover estimates for selected scenes does not correspond directly to the distribution of scores. Instead, Figure 4 shows how multiple objectives interact, a situation which would be very difficult to handle in an objective fashion with existing methods. Using the proposed method, the user could also change the weighting scheme on-the-fly to more strongly favor other attributes, such as the year of acquisition or phenological consistency. Scores, scene selections, and graphics portraying image attributes like cloud cover would be immediately updated for the user to examine. As an example, by changing the weight for year of acquisition listed in Table 2 to 10.0, we increase the likelihood of selecting scenes from 1985 or immediately adjacent years. As expected, Figure 6 shows that cloud-cover estimates for the highest scoring scenes from this weighting scheme are generally higher than Figure 5. These changes are not entirely predictable though, because of the influence of time of year and image quality on image scores at each path/row. One might imagine a graphical user interface where displays like Figures 4 and 5 update in real time as a user moves sliders representing the various weights, thereby allowing a rapid and detailed examination of the benefits of different selection schemes.

TABLE 2. SCORING SYSTEM FOR SCENE SELECTION (ABS -> ABSOLUTE VALUE)

Attribute	Range	Weight	Suitability Index
Cloud Cover	0–9	10	(9 - cloud_cover)/9
Image Quality	0–9	7	quality_rating/9
Time of Year	3 months	5	1 - abs(acquisition_date - target_date) / (0.5* date_range)
Year Acquired	5 years	3	1 - abs(acquistion_year - target_year) / (0.5* year_range)

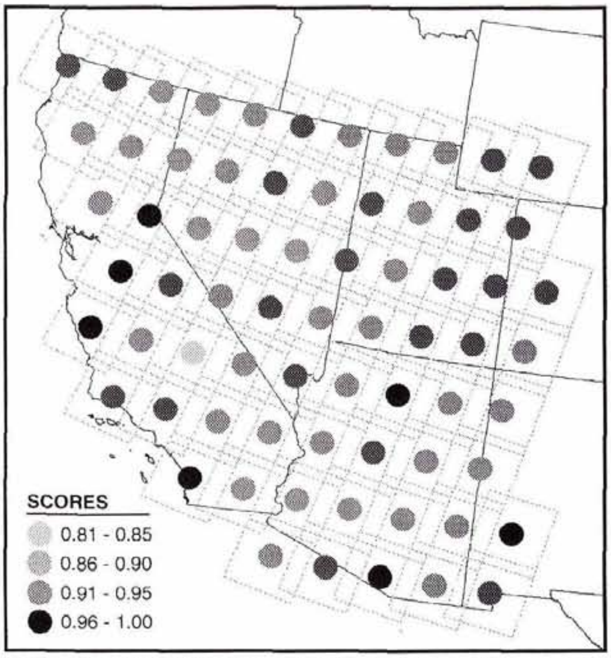

Figure 4. Suitability scores for the automated scene selection.

Because the metadata for cloud cover and image quality in the Landsat archives are not totally reliable or consistent measures, it will still be important to actually examine browse images for each path/row. The methods described here could easily be used to automatically download the GLIS browse images for each selected scene as well. If a scene was found to be undesirable because of dropouts or undetected clouds, an alternate browse image could be immediately selected and downloaded using the scoring list for that path/row. It would also be possible to create a "quick-and-dirty" compositing of browse images across the entire study area for rapid inspection. This could be done on-the-fly using image corner coordinates, or by implementing a browse standard with embedded georeferencing such as GeoTIFF (Ritter and Ruth, 1997).

It is clear that, with current query methods, the level of effort would be quite large to do a good job of selecting 69 path/rows from 1,134 possible scenes. The person performing such a selection would need to have a considerable understanding of the region in order to compensate for different phenological regimes, not to mention inter-annual variations. The difference in time and effort between manual and automated methods would then be multiplied if other years or phenological states were required. This test required approximately $1^1/_2$ hours from start to end, with the automated GLIS queries being performed during peak daytime hours of Internet traffic. For the majority of this time (> 1

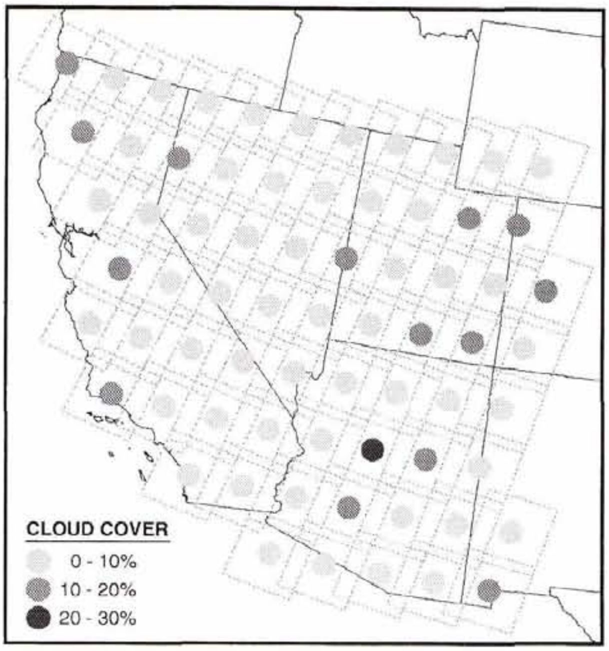

Figure 5. Estimated percent cloud cover for selected scenes.

hour), the program operated in a completely unsupervised mode. It is certain that the processing time would be dramatically reduced if this functionality were implemented and optimized at the host data center rather than involving numerous remote queries. After retrieving the requested set of archive records, the weighting function and its associated selections and displays could then be manipulated in real time.

Additional Design Issues

All the suggestions made here would be easy to implement using current technical capabilities. However, additional considerations are required for full implementation. The example presented here made use of precompiled global GVI datasets, so there were no problems with geometric coregistration. Automated geo-referencing would become an issue if such data were not being produced on a reliable basis or if user queries were to cover recent dates for which the data were not yet processed. For the purposes of selecting Landsat scenes which cover 185 kilometres on a side, it might be sufficient to use ephemeris data to perform a rapid but imprecise correction. The emphasis on systematic global coverage for upcoming sensor systems like MODIS and methods being developed for sophisticated access to data in EOS-DIS (e.g., Short et al., 1995; Cromp and Dorfman, 1992) should ameliorate this difficulty as information systems for accessing Earth science datasets evolve.

The two different world reference systems of historical Landsat data would also pose a challenge if Multispectral Scanner (MSS) data were to be supported by the system. This would also be an issue if options for substituting other imagery for Landsat were considered (e.g., SPOT, IRS). If so, the design should efficiently integrate these systems, such as performing an automated search for gaps in coverage when query results are drawn from different reference systems. Because of overlap of scenes between and within the WRS systems, the topological data structure used in many GIS software packages might not be appropriate for such tests of completeness. It may also be desirable to include a preference for consecutive scenes in an orbital track in the scoring system in order to make use of the swath processing which will be available for Landsat 7.

A further consideration involves the types and scales of ancillary environmental data to make available in the system. While the one-degree resolution GVI data used here seemed appropriate for characterizing average scene characteristics, there are limitations to its sole use. For example, the coarse scale GVI data cannot differentiate between greenup of natural vegetation communities versus that of irrigated agriculture. Higher resolution data, possibly combined with other map datasets (climate, ecozones, etc.), would provide more information on within-scene variability of phenological patterns due to elevation, coastal effects, or other causes. Another problem with the coarseness of the 1-degree grid is border effects. For example, some coastal path/rows in California which might have been considered in this test were not used because the image center point was in an area mapped as ocean in the one-degree dataset. Such problems could be alleviated by using higher resolution datasets such as the 8-km AVHRR Pathfinder data (Smith *et al.*, 1997), or by searching neighboring cells in the case of missing data. Cromp and Dorfman (1992) describe the implementation of a hypercylinder data structure that could facilitate the rapid retrieval and intersection of such disparate datasets.

Conclusion

A much more sophisticated system, which provides informative displays of environmental variability within the Landsat world reference system and provides a high degree of automation, is within the reach of current technology. The efficiencies offered by such a system would go a long ways towards making feasible the concept of global Landsat scene selections for a given time period and/or phenological state. A user could work region by region around the world, perhaps only looking at certain ecosystems, while the system identifies the required path/rows and their unique phenological patterns, and then tailors scene selections to meet the objectives for overall temporal and phenological consistency.

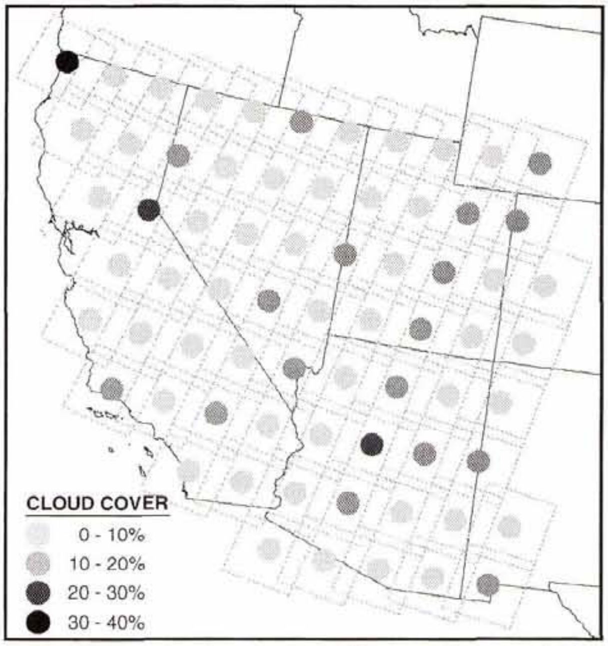

Figure 6. Estimated percent cloud cover when emphasizing year of acquisition.

This could be contrasted to limited manual queries using approximate dates for northern and southern hemisphere summers. The automated approach could also prove quite useful for planning acquisitions over a large number of geographically diverse field sites, such as the National Science Foundation's Long Term Ecological Research (LTER) network.

Both the ancillary environmental data used to support the scene selection and the scene metadata itself will be imperfect. The AVHRR NDVI data suffers from undesirable effects associated with cloud cover, off-nadir viewing, and sun angle. Cloud-cover estimates for Landsat scenes are unreliable. Though assigned a numeric score, the Landsat image quality indicator is actually a quasi-ordinal measure. The sensitivity of automated methods to these problems will require some examination. In any case, the final scene selection should be supported by a visual inspection of browse scenes. It is almost inevitable that the described method would be considerably more objective and dramatically more cost-effective than current methods when selecting scenes to cover a very large area.

Acknowledgments

This work was supported by the NASA Landsat Pathfinder Global Land Cover Test Sites project (NASA Grant NAGW-4214).

References

Caplan, M.A., 1994. *Interface Requirements Document between EOSDIS Core System (ECS) and the Version 0 System*, Hughes Applied Information Systems 194-219-SE1-004, Landover, Maryland.

Clinard, R.E., 1995. *Interface Control Document between EOSDIS Core System (ECS) and the Landsat System*, Hughes Applied Information Systems 209-CD-013-001, Landover, Maryland.

Cromp, R., and E. Dorfman, 1992. A spatial data handling system for retrieval of images by unrestricted regions of user interest, *Telematics and Informatics*, 9:221–241.

Goward, S., 1989. Satellite bioclimatology, *Journal of Climate*, 2:710–720.

James, M., 1994. The pathfinder AVHRR land data set, *International Journal of Remote Sensing*, 20(15):3315–3316.

National Aeronautics and Space Administration (NASA), 1994. *Landsat 7 Data Policy Plan*, http://geo.arc.nasa.gov/esdstaff/Landsat/l7policy.html, a WWW Home Page.

———, 1986. *Report of the EOS Data Panel*, NASA Technical Memorandum 87777, Washington, D.C.

Ritter, N., and M. Ruth, 1997. The GeoTiff data interchange standard for raster geographic images, *International Journal of Remote Sensing*, 18(7):1637–1647.

Short, N., Jr., R. Cromp, W. Campbell, J. Tilton, J. LeMoigne, G. Fekete, N. Netanyahu, and K. Wickmann, 1995. Mission to planet Earth: AI views the world, *IEEE Expert*, Dec. 1995, pp. 24–34.

Skole, D., and Tucker, C., 1993. Tropical deforestation and habitat fragmentation in the Amazon: satellite data from 1978 to 1988, *Science*, 260:1905–1910.

Smith, P., S. Kalluri, S. Prince, and R. Defries, 1997. The NOAA/NASA pathfinder AVHRR 8-km land data set, *Photogrammetric Engineering & Remote Sensing*, 63(1):12+.

United States Geological Survey (USGS), 1997. *Global Land Information System*, http://edcwww.cr.usgs.gov/webglis, a WWW Home Page.

(Received 17 June 1997; accepted 20 November 1997; revised 15 December 1997)
