University of Redlands

Building Complex and Site Categorization using Similarity to a Prototypical Site

A Major Individual Project submitted in partial satisfaction of the requirements
for the degree of Master of Science in Geographic Information Systems

by
Kim A. Wilson

Douglas M. Flewelling, Ph.D., Chair
James A. Ciarrocca, M.S.

December 2007

Building Complex and Site Categorization using Similarity to a Prototypical Site

The report of Kim A. Wilson is approved.

James A. Ciarrocca

James A. Ciarrocca, M.S.

Douglass M. Flewelling, Ph.D., Committee Chair

December 2007

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my mother for instilling in me an appreciation for education and the goal of obtaining a Master’s degree. Thanks to all of the MS GIS staff and my committee members. C-11 – thanks for keeping this year fun. John, Rueben, and Brandon – you guys rock! You were always so calm when my computer needed life support and were always able to bring it back to life and save my files. Nate, I really appreciate your assistance. ‘Dr. Doug’, thank you for your patience and guidance throughout the year. Theresa thanks for always making me laugh and for all your help with those seemingly little details that are always so crucial. Lastly, I would like to thank Ralph…for everything.

ABSTRACT

Building Complex and Site Categorization using Similarity to a Prototypical Site

by
Kim A. Wilson

This project presents an assessment tool for classifying building complexes using site-based relationships as calculated from ArcGIS 9.2 using model builder and Python scripting. Anthropogenic features extracted from imagery often form the foundation of spatial databases. These data are in turn used to inform situational awareness for relief, law enforcement, and military agencies among many others. Buildings and the complexes they form are critical features within the landscape. The categorization of complexes requires an understanding of the relationships of the buildings within the site. In this study, building complexes in California were assessed for similarity to a prototypical California high school defined with a training set of known high schools and compared to a set of uncategorized sites. Eighty-eight percent of the high schools were correctly classified as being highly similar to the control data set.

Table of Contents

Table of Figures

List of Tables

List of Acronyms and Abbreviations

AAS	Anaheim Angels Stadium
ATR	Assisted Target Recognition
CA	Percent of Total Component Area by Category
CSP	Candlestick Park
EDHS	El Dorado High School
F	Frequency of Categorical Occupation
GIS	Geographic Information System
GSHS	Golden Sierra High School
HGMS	Herbert Green Middle School
HHS	Highland High School
HNGTWN	Hangtown Speedway
HWPK	Hollywood Park
ICE	Indian Creek Elementary School
MMS	Markum Middle School
OBIA	Object Based Image Analysis (tentative name)
ORHS	Oak Ridge High School
PHS	Ponderosa High School
RHS	Redlands High School
SME	Sutter's Mill Elementary School
SSA	Site Similarity Assessment
STHS	South Tahoe High School
TA	Percent of Total Site Area by Category
UMHS	Union Mine High School
WSHHS	William S. Hart High School

1. Project Introduction

Humans tend to perceive objects within a view space, or simply *view*, collectively rather than discretely. For example, whether the focus is an object or set of objects, we intuitively assess the relationships of surrounding elements, which provide context for the object(s) under examination. Within the domain of image interpretation, a *site* is a subset of objects within a view space that has either a distinct purpose or identity. For example, the view may encompass a range of mountains, trees, roads, power lines, and buildings. Upon further examination, a subset of these buildings may be a residential neighborhood, a power generation plant, or a high school. Each of these collections of buildings has identifying, or signature, characteristics that make its designation recognizable. A formal site description defines essential elements or components of a prototypical site, utilizing these characteristics as a foundation. For example, a power generation plant *must have* cooling towers in close proximity to its generation buildings. The formal definition serves as a guide to assess similarity between different sites.

Assisted target recognition (ATR) is a rapidly growing field with applicability across many disciplines. Pattern recognition is a core element of ATR and a sub-discipline of machine learning. Pattern recognition seeks to classify data based upon predefined criteria, or from statistical data derived from the relationships between features. ATR integrates machine-learning algorithms, such as neural networks and decision trees, to facilitate the extraction of object-specific geographic features from high-resolution panchromatic and multi-spectral imagery. It significantly reduces the costs associated with extracting vector data from imagery by reducing the amount of human interaction required for extraction, creation, and maintenance of geographic data.

Extracted geographical features such as rivers, lakes, and shorelines, as well as anthropogenic features such as buildings, roads, power lines, industrial facilities, and power plants, are the foundation of spatial databases within geographic information systems (GIS) for both government and non-government organizations. This data enables military, law enforcement, and relief agencies to have clear understanding of critical geographic and anthropogenic features within their areas of responsibility, increasing situational awareness in crises. Features are typically categorized, for example, as a river, lake, or industrial facility. Full attribution of a given feature requires correlation to other sources, such as topographic maps, where this information is already included, or the use of *a priori* knowledge. This strategy may also work for a gross categorization of buildings, such as *industrial* or *residential.* However, it does not take into consideration the inter-relationships that can exist between the buildings.

Many facilities and installations, such as power generation plants, airports or airfields, and high schools, are composed of two or more buildings or components. Visually, these are obvious examples of multi-component facilities; each has at least one distinguishing feature that signifies possible classification, in addition to having at least two or more buildings or components. Airports and airfields have runways, power generation plants have distinctive cooling towers, and high schools have prominent athletic facilities such as football fields and running tracks. However, these signatures alone do not serve as positive identification; other factors, such as context and composition, also need to be considered.

Context is important for correct identification: power generation plants need to be close to a natural or anthropogenic water source; airports typically are in close proximity to major roads and away from residential neighborhoods; high schools are usually close to both residential areas and major roads. Composition is equally important as it can signify facility status or echelon. For example, an airport with multiple interconnected runways denotes a different capacity and use potential than an airport with a single, shorter runway. Schools with combined track and football fields usually signify a high school level rather than a middle or elementary school.

1.1. Problem Statement

The client has automated tools that extract and categorize vast amounts of feature data from imagery. However, relationships that exist between features within multi-component facilities, or sites, are not considered in this process. Image analysts are easily able to ascertain these site-based relationships through manual analysis. However, they do not extract features in tandem with this activity, nor would it be feasible to do so due to the quantity of imagery the client works with daily. Development of an automated process that considers site-based relationships would enhance the client's capabilities. An automated process may assist monitoring of known facilities in addition to narrowing the search for facilities in unknown locations.

By incorporating the ability to assess similarity to a prototypical facility or site, managers could feasibly use the process as an aid in the decision of where to allocate human and capital resources when searching in large unfamiliar areas. In order to conduct a search for a specific type of facility, the ideal composition of the characteristics of that facility need to be known. The first step in this process is to develop an understanding of the relationships that can exist between features within multi-component facilities and sites.

1.2. Research Question

The research question addressed by this project is: Do feature extracted data signatures sufficiently distinguish multi-component facilities?

1.3. Project Objectives

The objectives for this project were to develop an assessment tool for classifying sites using site-based relationships to include:

- A formal site description including related subcomponents.
- Creation of a geodatabase to organize and manage facility data, site description, and analysis.
- Generation of a profile from training sites.
- Mechanism to assess test data against the control data profile.
- Capability to communicate results.

1.4. Site Similarity Assessment Methodology

After consideration of many types of multi-component facilities, several California high schools were selected, due to the availability of data, a priori knowledge, and accessibility for field verification. A personal geodatabase was used within ArcCatalog to store and manage the feature extracted vector data that would be required for the proof-of-concept due to the anticipated small size of the data. Raster catalogs were implemented to manage a total of eleven images acquired for six of the facilities. The profile was derived based on statistical analysis of the aggregation of components within the high schools. Calculations were originally conducted in Microsoft Excel, and then later programmed within ArcGIS. Model Builder and Python scripting within were used ArcGIS to automate the workflow process, and a toolbar was created to facilitate the workflow and to guide the user through the process.

2. Background and Literature Review

Careful deliberation and investigation of what constitutes a site and how to measure relationships between the different components directed research work into three categories: similarity assessments; qualitative spatial relations, and object-based image analysis.

2.1. Similarity Assessments

Point objects, the simplest form of spatial data, may be used to represent natural objects occurring at an absolute location, or be a summary, or central point, of a larger distribution. Points are described by their patterns in terms of density and separation (O'Sullivan & Unwin, 2002). These patterns can be analyzed for relationships of occurrences and events (Longley, Goodchild, Maguire, & Rhind, 2005). Area objects, as with points, may be absolute representations of phenomena, such as buildings and lakes, or they may represent imposed areas, such as fire districts. Their shapes and areas are important analytical characteristics, though O'Sullivan and Unwin (2002) believe shape to be a difficult geographic concept to analyze.

Flewelling (1997) discusses spatial objects, classes, sets, and measures of similarity for sets. He explains that spatial objects can be discrete entities, such as a building, or aggregates (buildings, parking areas, athletic fields), as in the case of a school. Classes, on the other hand, provide a means to collectively describe and define objects, as well as their context. These class definitions "provide structures upon which to organize the phenomena they observe and form the basis of the classification of spatial objects" (Flewelling, 1997). A spatial class furthers this concept by incorporating location. Thus, a class definition for a high school may be 'has football field and running track combination', 'has many educational buildings', 'has multiple parking lots,' as well as a location component of being 'near residential areas' and 'may be bisected by road'.

Sets are groups of objects based on a common understanding and have a spatial ideal. For example, high schools are composed of educational and administrative buildings, athletic facilities, and parking components. A spatial ideal exists if all the elements within the set satisfy the class definition. Tversky (1977) proposed that similarity is essentially the inverse of distance and used cluster analysis to support his position. Thus, objects are considered to be closer if they are more alike. This important body of work led to database indexing structures that laid the foundation for search engines such as Google.

How similarity is assessed between sets of objects depends on the data's scale of measurement (Stevens, 1946). Nominal data measures only permit a Boolean assessment. Nominal values seek to classify the data, with its categorical classification being "…inclusive and mutually exclusive" (O'Sullivan & Unwin, 2002). If an object belongs to a particular category, it should not be capable of belonging to another category. Ranking or assessing distance between different categories is not possible.

Equality between categories of objects exists if there is a one-to-one relationship between all the objects in those categories. Similarity between categories of objects, on the other hand, simply requires that the categories share attributes (Flewelling, 1997). For example, two facilities may both have athletic categories containing different numbers of a variety of athletic objects, such as gymnasiums, running tracks, and football, soccer, and baseball fields. Therefore, their athletic facilities can be considered similar even though the objects may exist in both different type and numbers.

Similarity can be assessed between categories holding the same classification as a function of distance where "equality is distance 0 and inequality is as distant as possible" (Flewelling, 1997). Furthermore, "we can count category members to form frequency distributions. If entities are spatially located, we may also map them and perform operations on their (x,y) locational coordinates" (O'Sullivan & Unwin, 2002).

Ordinal data expands upon measures of equivalence to include greater than or less than comparisons, and thus permit ordering operations. Both nominal and ordinal data are commonly referred to as categorical data (O'Sullivan & Unwin, 2002). Interval data can include measures of differences or distances between categories in addition to ordering but does not have an inherent zero. Ratio data, on the other hand, is critical since it incorporates an inherent zero, thus absolute or relative magnitudes can be determined.

Bruns and Egenhofer (1996) discuss several different ways to measure similarity: assessment of deviation from equivalence and gradual change deformation. Objects are considered equivalent if they are the same, thus any deviation can be measured with a distance value. Gradual change imposes order on sets of spatial relationships, such as topology, distance, and direction. The gradual change deformation concept was developed in a formal model that describes the "…partial order over topological relationships and provides a measure to assess how far two relationships are apart from each other" (Egenhofer, Al-Taha, 1992). Thus, two spatial relations that require fewer deformation changes are more similar than those that require more changes.

Bruns and Egenhofer (1996) used the gradual change deformation concept to evaluate spatial scenes. Their research focused on assessing the similarity of controlled scenes that were composed of a few objects having a constant geometry type (region-region), shape, size, and with no rotation of the objects with a future work notation about the need to assess scenes that are more realistic. Egenhofer (1997) expanded on his previous work to assess which spatial constraints of a query could be relaxed while maximizing the return results in a similarity assessment. He determined that considering metrical refinements to the 9-intersection, such as length and area measurements, as well as cardinal directions, would facilitate a new paradigm in spatial queries wherein future queries would be based upon "…spatial relations rather than location in space…" (M. J. Egenhofer, 1997). Metric refinements, grouped into the categories of splitting, closeness, and approximate alongness, of the 9-intersection were later found to be "…critical to distinguish between … similar configurations" (Shariff, Egenhofer, & Mark, 1998).

Earlier work by Sharma and Flewelling (1994) introduced their qualitative spatial reasoning concepts, as well as a prototype implementation designed to work with incomplete and imprecise data to address this issue. Sharma and Flewelling's concept builds upon earlier work introduced by Egenhofer and Franzosa (1991) of representing explicit spatial binary relations between two spatial objects, such as cardinal directions, approximate distances, and the topological relations. The prototype combined two approaches: explicitly storing spatial relations, and utilizing relation algebra to describe the behavior of spatial relations that enabled qualitative inferences. Hong and Egenhofer (1995) followed this work and identified two possibilities for processing spatial queries by accessing explicitly stored qualitative spatial relations.

No reasoning or calculations are required for relations already stored in a database thus, relations are immediately available in response to a query. On the other hand, for relations not already stored, a reasoning mechanism infers the relation from those that already exist. They sought to translate quantitative location relations onto qualitative location relations between points and acknowledged that concepts of qualitative distances and directions often denote a range of valid values. Thus, the term 'near' can equally be used to describe a cooling pond about 50 meters from a power generation plant, or refer to the plant itself being within several miles of a town. Similarly, the actual azimuth of a power plant that is 'south-west' of the town may be different from the azimuth of a cooling pond that is considered 'south-west' of the plant. Therefore, they suggested utilizing a sector-based methodology to consider both distance and direction. Thus, "Objects within the same sector share the same qualitative locational relation with respect to the origin of the system (Hong, Egenhofer, 1995, p.671)." A problem they found with the approach was that the potential number of possible answers to a query was very high, thus the process to infer the relations could be computationally inefficient.

Other research into inferring direction relations started with a basic assumption that spatial databases would implicitly store direction relations of objects within a particular region, and algorithms would be used to infer the direction relations that were not stored between objects in different regions (Papadias, Egenhofer, & Sharma, 1996). They noted the need for a set of universally accepted formal definitions for direction relations (similar to those developed for topological relations), inferring relations, and relation composition in order to apply the algorithms correctly. Frank (1996) noted that while some precision is lost in the translation to qualitative directions from quantitative approaches, it "…simplifies reasoning and allows deductions when precise information is not available" (p. 270). Frank's research focused on utilizing an algebraic approach and identified 'inverse' and 'composition' as two operations.

2.2. Qualitative Spatial Relations

Egenhofer (1989) discussed the need to develop a formal definition of spatial relationships in order to "…clarify the users' diverse understanding of spatial relationships and to actually deduce relationships among spatial objects. Based upon such formalisms, spatial reasoning and inference will be possible (p. 457)." He suggested that a formalized mathematical theory of binary spatial relations with formal relation definitions was a requirement to spatial reasoning. Egenhofer introduced the concepts of topological relationships, corresponding formal definitions, and proofs based upon set theory.

Egenhofer and Franzosa (1991) proposed extending the previously defined point-set topological spatial relations between objects of *equal, not equal, inside, outside,* and *intersects,* to having relations be "…defined in terms of the intersections of the boundaries and interiors of two sets" (p.161). Thus, four fundamental relationships (common boundary parts as the intersection of bounding faces, common interior parts, boundary as part of the interior, and interior as part of the boundary) are derived from the comparison of the interior and bounding faces of two objects, the result of which can be either an empty or non-empty Boolean value. This body of work, referred to as the "4-intersection model for spatial relations," provided the impetus for continued research and the subsequent extension to nine intersections.

In the early 1990's, it was recognized that the then-available commercial database query languages were inadequate to support typical spatial queries, such as identifying 'all the bridges that cross the American River within 3 miles of downtown.' This led to the creation of several experimental spatial query languages, however, "their diversity, semantics, completeness, and terminology, varied dramatically" (Egenhofer, Herring, 1994). Egenhofer and Herring continued the effort to develop a single formalized binary topological relationship standard. They refined and extended the 4-intersection spatial relation model beyond the comparisons between the interiors and boundaries of two objects to include their exteriors as well. The "9-intersection" model compares nine possible topological set intersections between the interiors, boundaries, and exteriors of two objects. This advance considered an objects' context within space and provided more detail than the 4-intersection spatial relationship model, its predecessor.

Egenhofer and Sharma (1993) conducted a formal analytical comparison of the 4- and 9-intersection models in response to questions raised about the use of the 9-intersection model in comparing line-line and line-region relations, as it carries more computational overhead than its predecessor. They showed that both models returned the same results when the objects compared were the same geometry type (i.e. points, lines, or regions). However, for comparisons between a line and region, a two-point line and a complex line, or a region and a convex region, the 9-intersection provides a finer resolution of the topological relationships that exist between these types of objects. Furthermore, it can make distinctions between relations that the 4-intersection would evaluate as being the same. This evaluation was important as it had a bearing on the implementation of spatial queries within current geographic information systems (GIS). Utilizing the 4-intersection rather than the 9-intersection when the objects evaluated are of the same type reduces computational requirements.

According to Hornsby and Egenhofer (1998), effective modeling of objects in the real world needs to encompass structure, meaning, and behavior. They discussed the notion of formally modeling multi-part or composite objects that result from abstraction methods of association and aggregation. These abstraction methods provide an emphasis on the essence of a composite object, thereby removing the focus from less relevant details. Association creates a *member-of* relationship, while aggregation creates *part-of* relationships. For example, Andrews Air Force Base is a *member of* Air Force bases, while a runway or a control tower is *part of* an airfield. These methods of abstraction lead to an intuitive understanding that a relationship exists between composite objects and their components. Additionally, they propose that the identity of a composite object is based upon a distinguishing characteristic that sets it apart from other objects. They also suggest that this characteristic will exhibit some sort of purposeful pattern or structure between the components. For example, a single runway with no control tower, located in an agricultural area, versus an interlocking system of runways, coupled with hangers, large terminals, maintenance areas, and a control tower, distinguishes a simple agricultural or private airstrip from a heavily used passenger airport. This idea supports the often-heard axiom that the whole is more than the sum of its parts. Finally, they conclude with the concept that object identity relationships, when applied to composite objects, are useful in determining the existence or non-existence of an object.

2.3. Object Based Image Analysis

Object-based image analysis (OBIA) is an emerging sub-discipline of geographic information science and remote sensing that has its origins in the field of biomedical imaging. Current multi-spectral imaging sensor technology has the ability to produce very high resolution products that are proving to be a challenge to the traditional pixel-based image classifiers, because the pixels are smaller than many of the objects being imaged. As with traditional image processing, OBIA is composed of three components: feature extraction, classification, and product output. However, OBIA does not limit itself solely to the spectral characteristics of objects within an image; rather it considers the context, or surroundings, of an object as well. Further, it brings together multidisciplinary knowledge, and introduces spatial topology, to provide information particularly suited for analysis within a GIS.

According to Dr. Maggi Kelly (2007), an expert in the OBIA field, there is currently only a single software package, Definiens, truly capable of this type of work. She stated that other software packages attempt to incorporate the same concepts, but are not as robust in addition to falling short of having equivalent functionality. One of these packages has a module that works within the ArcGIS environment; however, she stated that the issues cited above make it a distant second choice to Definiens.

2.4. Summary

There have been different bodies of research assessing the relationships between spatial objects, three of which were relevant to this project: similarity assessments; qualitative spatial relationships; and object based image analysis. The scene similarity work conducted by Blaser (2000) used controlled variations of a set of five objects of unalterable shapes to form a scene, and a single scene then formed the basis against which all other scenes were compared. He drew heavily on topological relationships. Similarity of objects based upon topological relationships measures the number of transformations that must occur between sets. Egenhofer, and others' topological relationships can be used to conduct very detailed relationship analysis, and consequently similarity assessments. The wide variation of size, shape, and number of components within high schools, however, may make this approach quite cumbersome.

Object-based image analysis is a three-process methodology (extraction, classification, product output) that seems quite promising. License costs for the Definiens software is beyond the means of the University of Redlands. Furthermore, the introduction to OBIA from Dr. Kelly was well beyond the time when a change of course for this project would have been feasible consequently, OBIA was not incorporated into this project. In addition, OBIA incorporates AI principles and may be better suited for ATR applications.

Because the data for this project use a nominal measurement scale, are categorical in nature, and compose sets of objects, aggregation and frequency methods of measures were used to conduct this research.

3. Data

Data gathered for this project correlated to utilizing high schools as the test facility. An initial search for publicly available vector data of high schools was disappointing since the data available lacked sufficient building-level data. Data that included both buildings and ground-based features needed to be extracted from imagery. Requests for imagery that could be archived within ArcGIS from data providers proved to have associated legal issues. Projected AirPhoto, USA imagery for 2003 and 2004 (hereafter referred to as the *core imagery*) was acquired in tagged image file format (TIFF) that covered the existing six high schools in El Dorado County, California (hereafter referred to as the *training set*).

Features extracted from imagery of these six public high schools served as the core for this project. In addition to having GIS ingestible imagery of these schools, *a priori* knowledge of their layout reduced the need for field verification. These six high schools are located in suburban or rural settings in northern California and are representative of west coast suburban and rural schools: each has parking, multiple buildings, and indoor and outdoor athletic facilities. Several of the schools are located with alternative-education high schools. The six school sites span an elevation range of 200 to 5500 feet above mean sea level; therefore, noticeable differences exist due to typical winter weather conditions and terrain. Additional differences, such as architecture and number of portable buildings observed, are largely due to age. El Dorado County's original high school, El Dorado, was built in 1928 with an enrollment of 163 students and currently exceeds 1500 students. The newest high school, Union Mine, built in 2000, has no portable buildings associated with its main campus. However, there are portable classrooms for its alternative education high school. This range of elevation, terrain, age, and enrollment were important considerations in selecting facilities that would provide a comprehensive picture of a prototypical Californian suburban or rural high school.

Three additional high schools in different parts of California were selected for this project, all three being suburban in nature. *A priori* knowledge of the layout of one of these schools, coupled with the other two located within a reasonable distance from the University of Redlands simplified field verification. In addition, three middle schools, and three elementary schools within El Dorado County were included for the feature extraction process. While high schools are the target facility, other facilities having a similar purpose (student education), but of a different echelon (elementary and middle schools) were also included. Having functionally similar facilities serves to test the efficacy of this methodology.

The primary goal of this project was to develop an assessment of site similarity. Therefore, several non-school facilities with similar elements to high schools were included in the test feature data set. A horse racing facility was included since it contains high school-like elements, such as several oval tracks, many grouped buildings, and parking areas. A small county owned property that includes a quarter-mile dirt-track speedway, several baseball/softball fields, parking, and multiple buildings was also included. Finally, two professional major league baseball parks rounded out the data set. The inclusion of these four non-school facilities, two having elements of a high school and two having no correlation, also serve to test the efficacy of the methodology.

Imagery for all facilities outside of El Dorado County was obtained through a subscription, via the Redlands Institute, to the ESRI ArcGIS Image Server, housed at the University of Redlands. This imagery has a one-foot to two-meter resolution, with one-foot resolution being found only in metropolitan areas. A comparison of this imagery to that obtained for the facilities within El Dorado County revealed a two-foot resolution, the same as that of the core imagery data. This imagery was also consulted for the facilities within El Dorado County, as it was more current. The imagery obtained for the training set was archived within the project geodatabase and reflects the status of those facilities at a particular date. However, the imagery accessed via the ArcGIS Imager Server is not archived because the licensing agreement held by the Redlands Institute does not permit exporting the images into a geo-referenced format. Therefore, the imagery will change as the image server is updated.

3.1. Database Design Consideration

A personal geodatabase was selected for the management and administration of the data for this project. Domains and subtypes were included to enforce data attribution integrity. Domains constrain component categorization. In this case, a component is restricted to administrative, athletic, educational, parking, or unoccupied categories. Subtypes constrain the value choices of the component type. Facilities were manually extracted from imagery using heads-up digitizing in ArcGIS resulting in a feature data set. Raster catalogs provide links to the images of the training sites as a personal geodatabase does not have the capability of internally storing images. Although this poses a constraint, it was chosen over a server-based database system because there were only eleven images to manage.

3.2. Source Data

The datasets used in this project were derived from government sources and ESRI's ArcGIS Image Server. The feature-extracted data were projected in the same coordinate system as the core imagery (Appendix A). All data with the geodatabase were derived from the datasets listed in Table 3.1.

Table 3-1 – Table of Data

Dataset	Description	Source and Date	Data Model
hs_pars	Boundaries of High Schools in El Dorado County, CA	El Dorado County Surveyor's Office, April, 2007	Polygon (vector)
Parcels	Parcels within El Dorado County, CA	El Dorado County Surveyor's Office, May, 2005	Polygon (vector)
edcSchools03	Aerial Photography, 2 foot resolution	AirPhoto, USA, 2003 (Via El Dorado County	Raster Catalog

Dataset	Description	Source and Date	Data Model
		Surveyor’s Office)	
edcSchools04	Aerial Photography, 2 foot resolution	AirPhoto, USA, 2004 (Via El Dorado County Surveyor’s Office)	Raster Catalog
ESRI’s ArcGIS Imager Server	Aerial Photography	ESRI Image Service (accessed Via Redlands Institute) Varied dates	Raster Image Service

3.3. Geodatabase Architecture

The database architecture for this project consists of a personal geodatabase in Microsoft Access that manages both input data and results. The database houses facility data, the facility description implemented with domains and subtypes, derived facility signatures in the form of K-Score feature classes, and two raster catalogs. Feature classes and database tables resulting from the similarity analyses round out the database contents. Metadata, accessible from ArcCatalog, provides information for each feature class including projection and coordinate system documentation.

4. Methodology

The site similarity assessment (SSA) methodology began with the creation of a facility description. The facility description served as a guide to the database design; domains and subtypes were implemented for facility component categorization, as well as to enforce data attribution integrity. Feature extraction, via heads-up digitizing within ArcGIS, provided the vector data of each component within the facilities. Three different facility occupancy measures were analyzed: 1) frequency of components by category; 2) percent of total component area by category; and 3) percent of total site area by category. The result is a methodology to assist locating potential instances of a particular facility type.

4.1. Facility Description

Multi-part facilities are composed of two or more components. The facility description defines and describes all known components, and establishes component classification categories. A formal description, based on an analysis of the facility components, serves as a guide for the corresponding spatial database design, both for creating domains and subtypes to enforce data integrity, and for categorization purposes. An analysis of the elements of high schools revealed repeatedly recognizable characteristics. The following categories are integral to classifying components of high schools: administration, athletics, education, parking, and unoccupied space. Criteria on which these categories are based are discussed in more detail below.

Campus-style high schools often encompass buildings of varying sizes and configurations. Additionally, there are distinct differences between buildings that are permanent structures and those that are portable or temporary. Permanent structures generally have considerably larger footprints, have a single roof with no gaps across the span of the roofline, often stand alone, and have multiple well-defined pathways providing access to several different portions of the structures. Modular buildings, on the other hand, occupy a significantly smaller footprint, and are usually found in groups of three or more. The groups are typically located at the periphery of the permanent buildings or parking lots, though are sometimes sandwiched between various athletic facilities. Furthermore, their pathways are usually less pronounced and fewer in number and typically only lead to the front of the group. The height difference between permanent and portable buildings is evident from the different lengths of shadows they cast. Figure 4-1 depicts these differences, with arrows denoting the different length shadows. Based on *a priori* knowledge of the training sites, coupled with campus maps obtained either from the school web sites or from administrative staff, buildings were found to serve several distinct purposes.

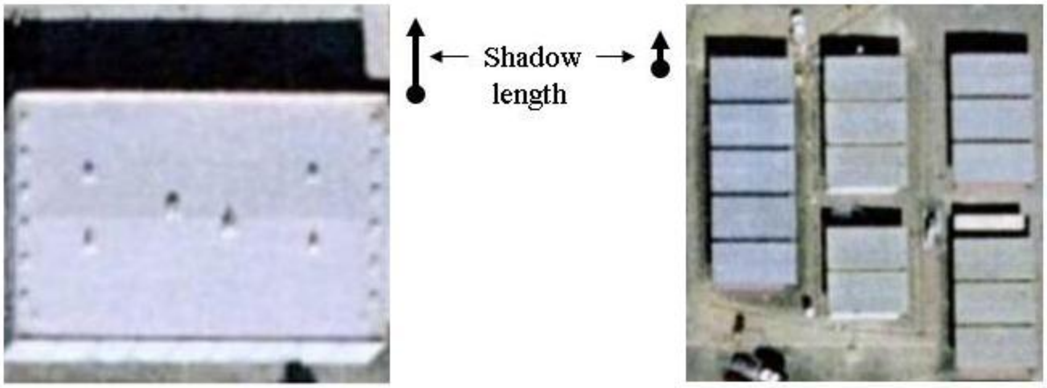


Figure 4-1 Differences Between Permanent and Portable Buildings

(Image: Google Earth accessed 24 Oct 2007)

Typically, portable buildings are used for various educational purposes, including traditional subjects and courses, and those taught by trade or regional occupation programs (ROP), or alternative education schools that are sometimes located with the high schools. Gymnasiums are located in large permanent structures, often with a larger central area flanked or adjoined by shorter buildings, which typically house separate male and female gyms and locker rooms.

Administrative functions, such as the administration office, teachers' offices, and career, guidance, and counseling centers, are typically housed in permanent structures, are often housed within the same building, and located in close proximity to a parking lot or centralized drive-through area. Cafeterias are located in permanent structures, and in some locations, house a multiple purpose room that doubles as a gathering location for school assemblies.

Components also include maintenance systems, such as centralized heating and air conditioning systems, or janitorial spaces. These are typically located in permanent structures, and often housed in an administration or educational building. They are often not distinguishable from other buildings, and their locations were identifiable only with the use of a campus map, or via a phone conversation with school staff. Due to the lack of a noticeable independent signature, these components are included in the administrative category for the purposes of this project.

Building styles vary considerably, with the most noticeable differences due to the age of the school. Older schools tend to have rectangular buildings and more temporary classrooms. The newer schools have buildings with multiple angles and configurations, and have fewer temporary buildings (Figure 4-2).

Based on analysis of the building utilization, it is evident that they serve three functions: administrative, educational, and athletic. Although there are noticeable physical distinctions between the permanent and portable buildings, they are not segregated into a separate category for this project.

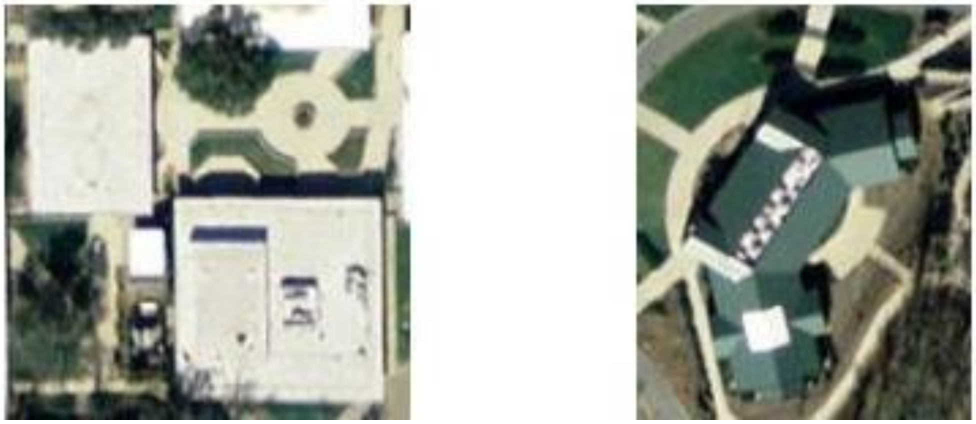


Figure 4-2 Visual Difference in Schools Based on Age

(Older School on Left, Newer School on Right)

(Image: ESRI's ArcGIS Image Server accessed 24 Oct 2007)

Outdoor high school athletic facilities are easily recognizable. For example, softball and baseball fields have distinctly shaped infields with associated outfields; running tracks are oval, and typically have a football field inside their boundary. The quality of the infields and outfields typically varies between better-maintained varsity fields and junior varsity fields. Football fields are normally found within a running track, have permanent goal posts at either end, and during football season will have visible chalk lines on the grass (Figure 4-3). Soccer fields are large expanses of maintained grassy areas devoid of trees, bushes, or structures, and may have goal posts in place at either end of the field, or stacked off to a side. Tennis courts are recognizable by the muted red and green ground surface, coupled with distinctive white markings outlining the different portions of the court. Basketball courts tend to be located on asphalt or naturally colored concrete surfaces with white markings delineating the courts. Often, shadows of the poles and backboards are required to provide a count of the number of courts (Figure 4-4).

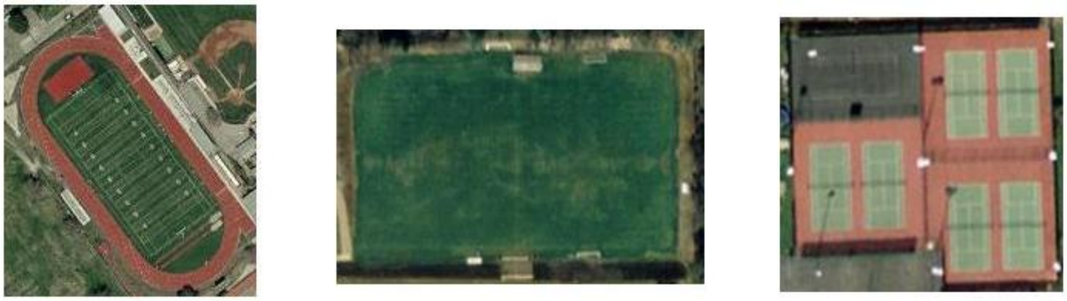


Figure 4-3 Various Outdoor Athletic Components

(Left to Right: Football/Track; Soccer Field with Goal Posts; Tennis Courts)

(Image: ESRI's ArcGIS Image Server accessed 24 Oct 2007)

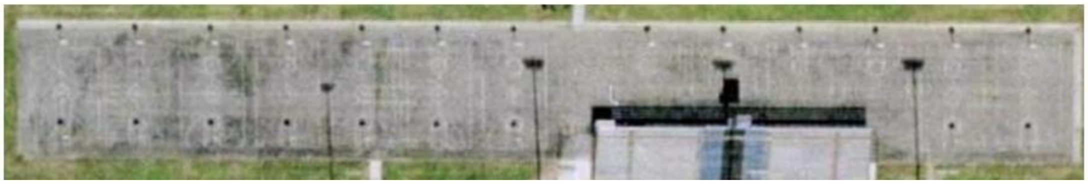


Figure 4-4 Using Shadows to Discern Basketball Courts

Black Dots are Shadows of Backboards

(Image: Google Earth accessed 24 Oct 2007)

Outdoor athletic facilities located at elementary and middle schools are markedly different. They do not have combined football fields and running tracks or tennis courts, and may only have one baseball and softball field combination. However, they typically have large multi-purpose athletic fields used for soccer, flag football, and other activities. A number of recreational courts, including basketball, dodge ball, and foursquare, are typically located on asphalt surfaces. Additionally, there may be several distinct playground areas at each site (Figure 4-5). These lower echelon schools were assessed during the facility description phase. This enabled their components to be included as subtypes within the categories identified for high schools.

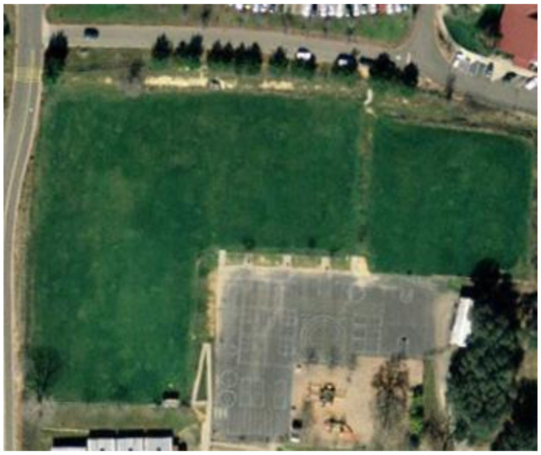


Figure 4-5 Elementary School Athletic Field/Multi-Purpose Playground

(Image: ESRI's ArcGIS Image Server accessed 02 Nov 2007)

Parking lots are considered an additional element signifying echelon because public transportation, such as subways, city buses, or taxis, are not available to students or staff attending schools in suburban or rural areas. Students must utilize either school buses or private transportation. Therefore, there must be a sufficient number of parking lots to accommodate students, staff, and school buses (Figure 4 -6).

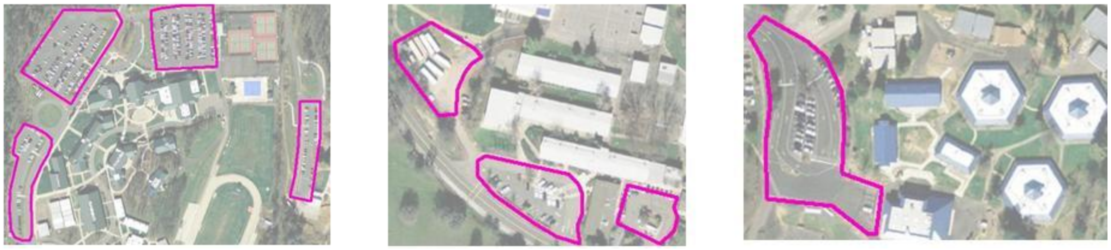


Figure 4-6 Parking Components at Different Echelons of Schools

Left to Right: High-, Middle-, and Elementary School

(Image: ESRI's ArcGIS Image Server accessed 02 Nov 2007)

Typically, there are separate lots for use by students and staff, and in some cases, school busses may not be located on-site due to space constraints. Unoccupied space encompasses all areas within the boundary of the facility that do not contain components.

4.2. Feature Extraction

Features can be extracted from imagery in several different ways. The most popular method is automated via image processing software. Due to economies of scale, this allows a tremendous amount of data to be gathered at less cost. This type of process involves the creation of 'training areas', wherein specific image pixels are identified as belonging to a particular type of feature. Typically, thousands of pixels representing different features of the same type are incorporated into the training set in order to provide a comprehensive knowledge base against which to compare other features during the extraction process. Once incorporated into the training set, pixels are classified according to feature type, such as, paved or dirt road, deciduous tree, conifer, building, etc. This process is repeated for each feature type targeted for extraction, and the training sets become part of the formal definition of the features they represent. Optimizing the results may require several refinements of the training sets.

After the training sets are optimized, they are saved into a *definition* file to be used during future automated extraction sessions, and therefore can be reused as many times as desired. While this greatly reduces the time and costs involved with extracting features, it is not a perfect process. Therefore, post processing is required. Different phenomena, such as the angle at which the imaging sensor acquired the image and shadowing, can affect the visibility and distinction of portions of the targeted features. Post processing entails ensuring the features extracted are correctly classified and complete. No gaps in the lines representing linear features such as roads and rivers should exist. Similarly, features that represent areas such as buildings or lakes should result in polygons that have no gaps. Features that are broken with gaps require manual intervention for correction. Post processing can be a time intensive process.

The second method of feature extraction is manual, wherein images are loaded into a GIS. The outlines of features are traced and the extracted features are classified. While this process is time consuming and redundant, creating training sets, and the subsequent post processing, are not required. This methodology is typically used only when a few features need to be extracted from a small area.

4.2.1. Feature Extraction Overview

All facility data were manually extracted from project imagery using heads-up digitizing within ArcGIS. The decision to employ heads-up digitization to extract the components was two-fold. First, the expected learning curve to use available image processing software was deemed excessive. Second, the subsequent post processing required of the resultant data was determined to be significantly greater than that of the manual process, given the small number of facilities analyzed.

The feature extraction process started with a determination of the boundaries of the facility properties, thereby ensuring that only components directly related to the high school would be extracted. In the case of the training set, property boundary polygons were obtained from the data provider. In a few cases, such as suburban El Dorado High School (Figure 4-7), the property boundaries (indicated by a blue outline on the image) are fairly evident. There is a clear indication of the extent of the facility grounds denoted by fencing, public roadways, and the presence of residential properties.

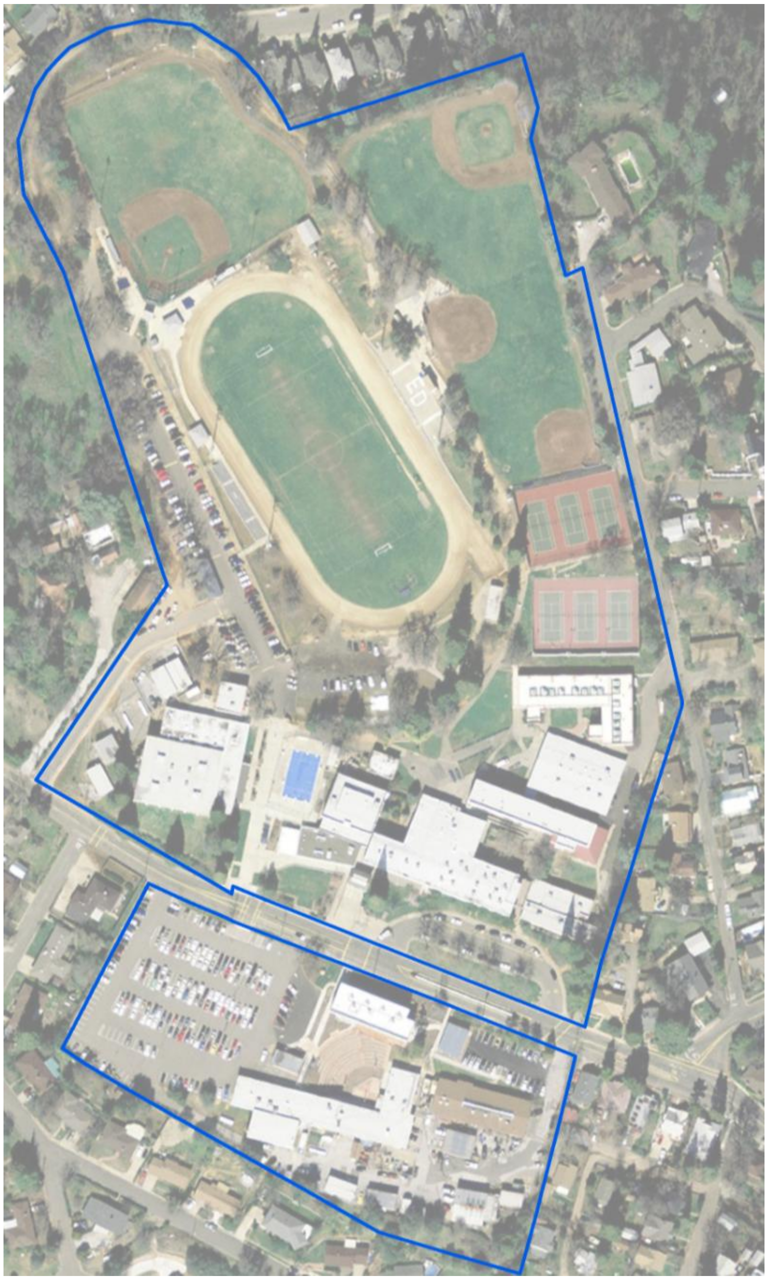


Figure 4-7 Easily Discernible Site Boundary in El Dorado County, CA

El Dorado High School

(Image: ESRI’s ArcGIS Image Server accessed 24 Oct 2007)

The other El Dorado County high schools included land well beyond observable school facilities with no clear indication (fencing, roads, or neighborhoods) of their boundaries, with some being very oddly shaped. Imagery of South Tahoe High School (Figure 4-8), demonstrates an extreme example of a school in a rural location. Without the acquired parcel boundary, the extent of the school's property could not be determined.

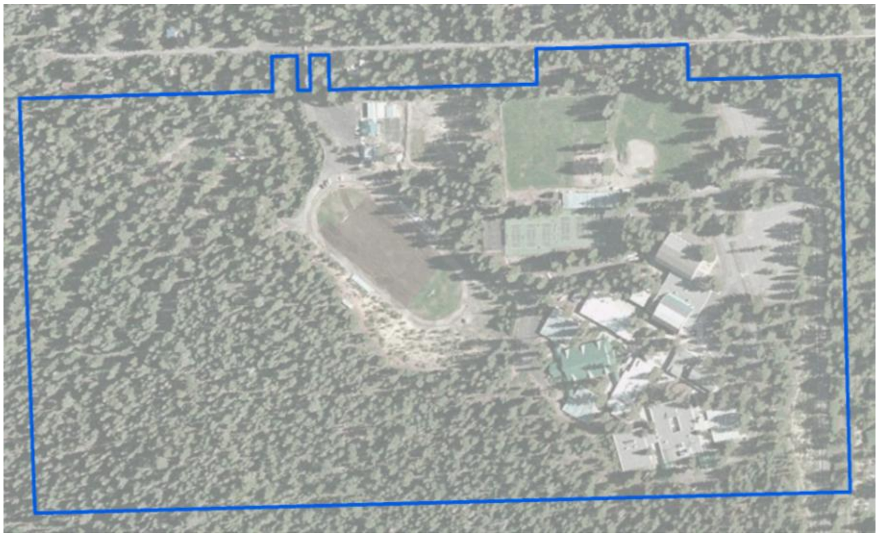


Figure 4-8 Indiscernible Site Boundary in El Dorado County, CA

South Tahoe High School

(Image: ESRI's ArcGIS Image Server accessed 24 Oct 2007)

Boundaries of the seven other facilities located within El Dorado County were derived from the same parcel layer. The boundaries of the remaining facilities (outside El Dorado County), were created through image analysis and heads-up digitizing. Ascertaining the boundaries of these facilities proved to be quite straightforward, as noted in the example imagery of William S. Hart High School located in Valencia, California (Figure 4-9) due to clear demarcation by roads and surrounding neighborhoods.

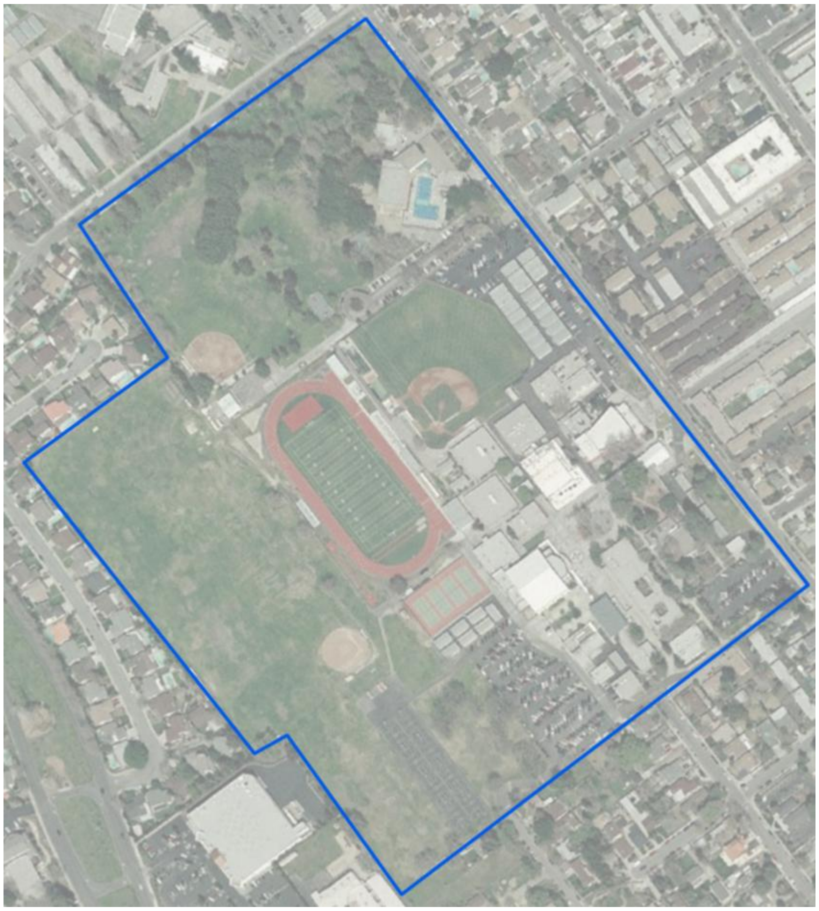


Figure 4-9 Easily Discernible Site Boundary Outside El Dorado County, CA

William S. Hart High School

(Image: ESRI's ArcGIS Image Server accessed 24 Oct 2007)

4.2.2. Feature Extraction Process

Polygons of the boundaries of each component were obtained by heads-up digitizing within ArcMap, and entailed starting an edit session and tracing polygon outlines over the features visible within the imagery. *Components* were deemed to be both structural, such as buildings, and ground-based, such as outdoor athletic and parking areas. Multiple portable buildings were digitized with a single boundary when located immediately adjacent to one another in distinct groups, as this type of grouping can be correlated to a single permanent building where multiple classrooms are contained within a single permanent building rather than within individual buildings for each classroom. Baseball fields are often located with softball fields — sharing an outfield but having separate in-fields — separate polygons were digitized in order to accurately convey that they are distinct entities. Tennis courts were digitized separately to show the difference in quantity observed between schools, and because each court is easily discernible from the imagery used for extraction. Similarly, basketball courts were captured as individual entities with a single court consisting of two poles and backboards at either end of the court. Shadows aided in the identification of individual courts due to the degraded condition of the painted court markings. Figure 4-10 depicts a high school with extracted components overlaid on the imagery.

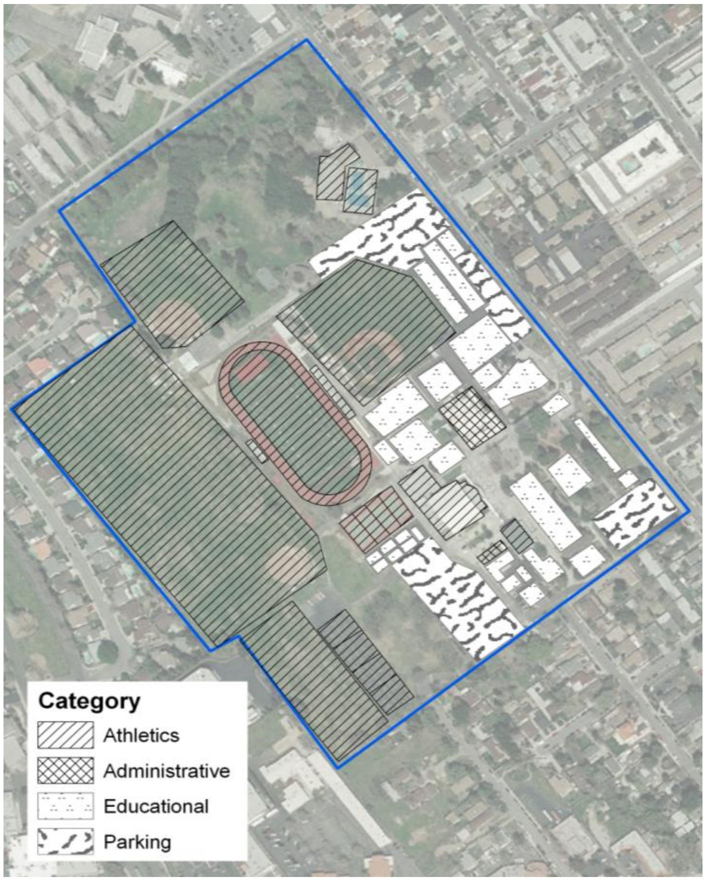


Figure 4-10 William S. Hart High School with Components Categorized

(Image: ESRI's ArcGIS Image Server accessed 24 Oct 2007)

In order to consistently portray potential space utilization throughout the different echelon of facilities, large maintained grassy areas that exhibited no clear indication of sport type were extracted and classified as athletic fields if the areas were not dominated by trees or surrounded by buildings, which would indicate non-athletic use.

4.2.3. Assignment of Component Type and Category

Assignment of facility component types is based primarily upon imagery analysis and *a priori* knowledge. In several cases, campus maps were obtained from school websites or administrators, though a few were of limited assistance as they were either incomplete or outdated. Component types are restricted to the values contained in the *Associated Component Sub-Types* (Table 4-1). Unoccupied space is the inverse of the occupied space within the boundary of a given facility.

Non-school facility components are categorized as if they were schools. This is an important point, as the goal of this project was to create the ability to assess facilities of an unknown type, or being suspected of being a particular type, against prototypical facility statistics to determine their probability of being the type of facility sought. An example in this project is the inclusion of Hollywood Park, a horse racing facility. Hollywood Park does not have educational buildings; however, it does have two large oval race tracks similar to the running tracks seen at high schools — one of which is also similar in size — several parking areas, and groups of similar sized and shaped buildings, which are actually barns. This facility's components are classified as if the features might be those found in a high school. Therefore, the barns are justifiably categorized as 'educational' and the racetracks as 'athletics'. Table 4-1 depicts the valid values for the 'Category' domain with associated subtypes.

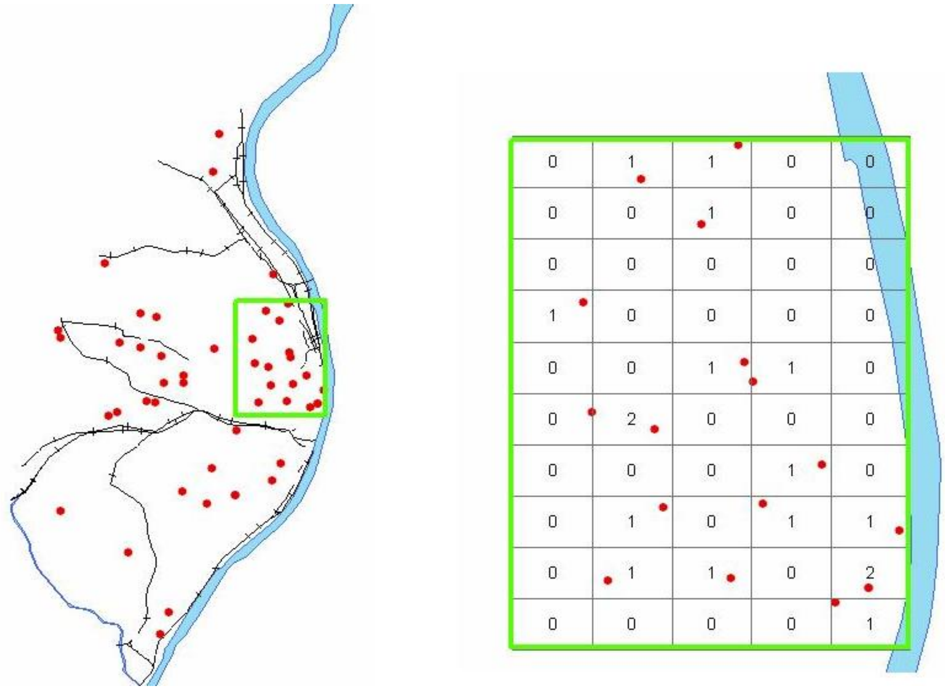
Table 4-1 Category Domain with Associated Subtypes

Component Category	**Associated Component Sub-types**
ADMIN	Administration Maintenance Commercial Bldg Government Building
ATHLETIC	Athletic field Baseball/Softball Field Basketball Courts Football Field Gymnasium Multi-purpose Playground Race Track Running Track Soccer Field Speedway Stadium/Bleachers Swimming Pool Tennis Court
EDUC	Classroom Barns
PRKG	Parking: Staff

Component Category	Associated Component Sub-types
	Parking: Student Parking: Bus
UNOCCUP	All areas within boundary not occupied by any component

4.3. Measuring Similarity of Nominal (Categorical) Data

Geographers commonly use quadrats to count occurrences of events. A quadrat is a measured area, usually square but may be of any shape that is used for sampling spatial phenomena. Quadrat counting can be conducted in one of two ways. The *census* approach encompasses the entire study area with quadrats that do not overlap and may not exceed the boundary of the study area. Conversely, quadrats may be replaced randomly, thus may overlap each other and the study area boundary and leave areas uncovered (O'Sullivan & Unwin, 2002). Both methods provide an approach for counting occurrences of events within the study area. Figure 4-11 shows an example of using the census approach. The figure on the left shows locations of attempted street robberies within a city. The figure on the right shows a study area that has been tessellated into evenly sized quadrats. Each quadrat has a count of the number of robbery attempts found within its boundary.



0	1	1	0	0
0	0	1	0	0
0	0	0	0	0
1	0	0	0	0
0	0	1	1	0
0	2	0	0	0
0	0	0	1	0
0	1	0	1	1
0	1	1	0	2
0	0	0	0	1

Figure 4-11Quadrats: Census Approach

Quadrats are discrete, independent entities and are not referred to by a numbering system or in terms of their proximity to each other. They are simply containers that provide the means to count the events that are contained within them. In effect, they take spatial data that has location as part of its attributes and reduce it to an aspatial count of occurrences within each quadrat. The categories used in this project are thus analogous to quadrats. They are used for the sole purpose of counting the occurrences of objects that fall within their boundaries. Data is reduced at this point to a nominal, categorical scale as it either falls within a category or does not.

Nominal (categorical) data cannot be ordered, although "…it is possible to distinguish observation memberships in the categories…" (Wong & Lee, 2005). Therefore, it is possible to assess the similarity of nominal data values of a known ideal against observed values. Similarity can be measured by evaluating the frequency of each categorical value in both the expected and observed sets.

Flewelling (1997) used this approach in his doctoral thesis and developed a derivation of the statistical formula traditionally used in the quadrat method of counting occurrences. "The K-score sums the differences between the frequencies of nominal values (f_o) in a set with the expected frequencies (f_e) (Equation 4.1). The result is normalized by the maximum difference that could occur by substituting the number occurrence (N) for (f_o) in the category with the lowest expected frequency (f_e) and zero is substituted for all other observed frequencies" (Equation 4.2) (Flewelling, 1997).

$$\text{K-Score} = \frac{\sum \left| f_0 - f_e \right|}{K\text{max}} \qquad (4\text{-}1)$$

$$\text{K max} = \left| N - f_{e_{\min}} \right| + \sum \left| 0 - f_{e_n} \right| = 2\left(N - f_{e_{\min}}\right) \qquad (4\text{-}2)$$

The normalized form enables comparisons of different size data sets. The K-Score relies on a set of expected theoretical frequencies and can only be used where the expected set is known. However, it is possible to derive expected values by evaluating a training set (a group of facilities whose median coalescence will represent the ideal).

Multi-part facilities of a similar type (i.e. high schools, power generation plants, or airfields) may consistently be composed of components that can be classified into a known set of categories. For example, high schools are composed of components categorized as administrative, athletic, education, and parking, yet the number of components within each category, or the percentage of total occupation by category, can vary. As this variance precludes the existence of a *universal ideal*, an estimation of this ideal can be derived by gathering statistical data from a known group of different high schools, in order to develop a representative ideal. The resultant data can then be used as the ideal, or expected, values against which other facilities may be compared. Six high schools were used as the training data to establish the universal ideal for this project.

Statistical distance, the difference between observations that share a group of attributes (O'Sullivan & Unwin, 2002), is the inverse of similarity. Conventionally, similarity is calibrated with a value of zero representing complete dissimilarity, and a value of one meaning no measurable distance. Conversely, when considering distance between a target set and another dataset, a value of zero reflects high similarity, while distance values approaching one are very dissimilar (Flewelling, 1997). This project relied on the latter, wherein the derived universal ideal was the control dataset, or benchmark, to which test datasets were compared.

4.3.1. Training Data Process

In order to conduct a database search of vector features for a specific type of facility, the ideal composition of the characteristics of that facility need to be known. If the facility type is of a very strict nature, in that there is absolutely no variance between facilities in different locations, then a single facility can be used as the ideal. However, if the facility type can exhibit a wide variance, then averaging a representative sample group of facilities derives the ideal. This was the methodology employed for obtaining the prototypical high school.

The term *training data* is borrowed from artificial intelligence (AI) applications and involves the use of neural networks and complex computational models. However, since training data serves as a control set against which to compare other data, the term is used in this project when referring to the creation and comparison of the control dataset.

4.3.1.1. Creating Training Data

High schools were chosen as the target facility for this project, in part due to the general perception that they are quite similar in composition. They all have educational and administrative buildings, indoor and outdoor sports components, and several parking areas. However, this perception was altered after their components were extracted from imagery. Beyond the base composition, their appearances proved to be extremely varied (Figure 4 -12).

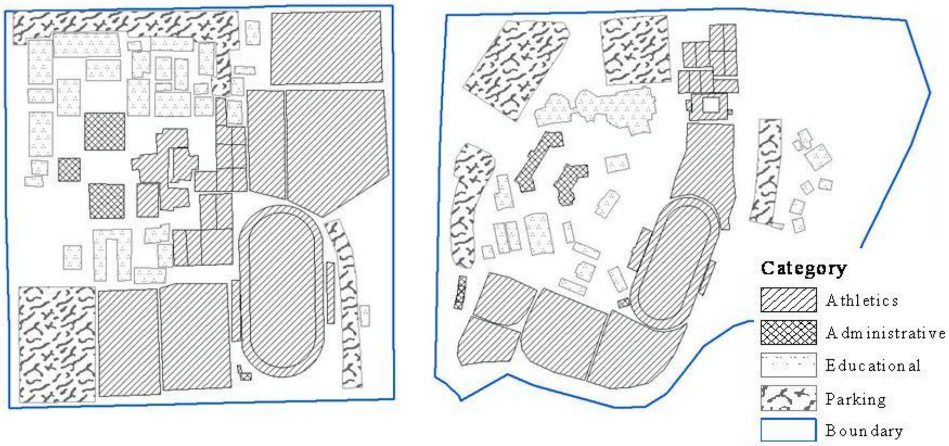


Figure 4-12 Variance in High School Appearance

Building size, shape, and quantity vary with location and construction age. Administrative buildings are not always centralized or closest to drive-through parking areas. Blocks of portable structures often augment permanent educational buildings.

Sports facilities exhibit similar differences in the numbers of gymnasiums, tennis and basketball courts, and baseball and softball fields. These variances preclude the existence of an individual organic universal ideal. A group of high schools was therefore selected in order to establish a universal ideal against which to compare other facilities. The training sites were chosen as they provide a well-rounded variety of style and composition. The creation of training data is facilitated within the site similarity user interface discussed within the next few sections.

4.3.2. Site Similarity Assessment User Interface

The custom Site Similarity Assessment toolbar shown in Figure 4-13 was programmed with Model Builder, as well as Python and Visual Basic for Applications (VBA) scripting, to guide users through the assessment and reporting processes. Model Builder was initially used to implement the functionality behind the *Conduct Analysis* button. This approach enabled tools available within ArcToolbox to be easily combined and customized to meet the needs of this project. Python scripting was used to replace and streamline the model processes. The scripting consolidated several forms in to one *K-Score* form, thus simplifying user interaction. The Python script calls on several ArcGIS system tools in addition to incorporating the K max and K-Score statistical equations discussed later in this chapter. The *Reporting* button was created using VBA and ArcObjects to access a reporting capability within ArcGIS.

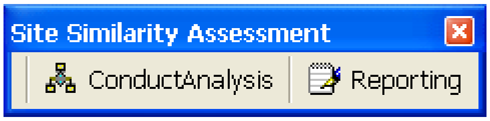
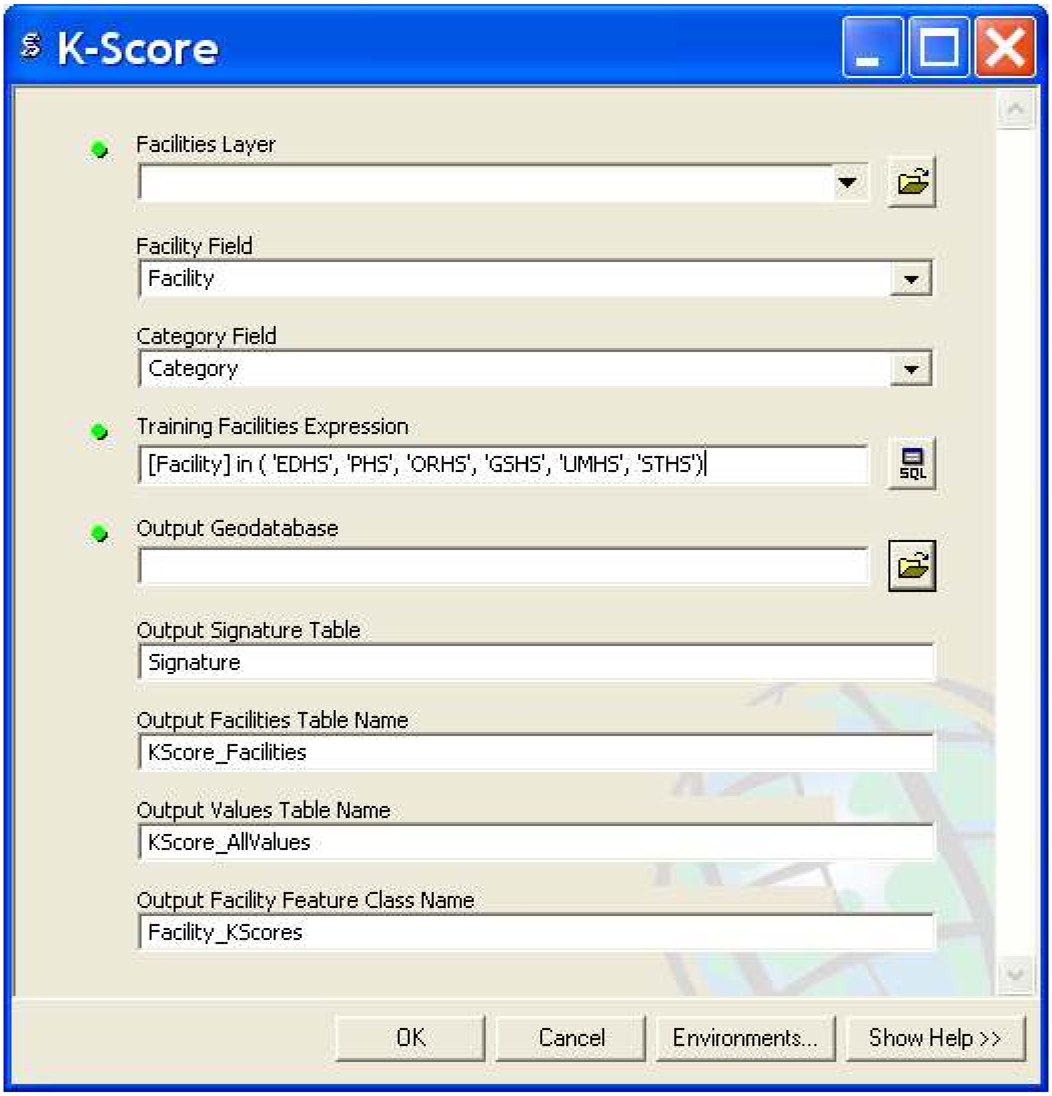


Figure 4-13 Site Similarity Assessment Toolbar

Selecting *Conduct Analysis* on the toolbar opens the K-Score form (Figure 4-14), facilitating user input for selecting training facilities, and conducting three different site similarity analyses.



K-Score

Facilities Layer

Facility Field
Facility

Category Field
Category

Training Facilities Expression
[Facility] in ('EDHS', 'PHS', 'ORHS', 'GSHS', 'UMHS', 'STHS')

Output Geodatabase

Output Signature Table
Signature

Output Facilities Table Name
KScore_Facilities

Output Values Table Name
KScore_AllValues

Output Facility Feature Class Name
Facility_KScores

Figure 4-14 K-Score Form

This form provides the means for the user to select the input feature class containing multi-component facilities, to identify the facility and category fields, to select the facilities that will comprise the training set, and to identify the output location and the layer names for the feature layers resulting from the process. This process required that the training facilities and test facilities be within the same feature layer.

4.3.3. Site Similarity Assessment Algorithm

The SSA methodology enables users to select one or more facilities to create a universal ideal, or prototypical, high school. The process starts by selecting participating facilities by name. Dissolving the attributes of the facility components then results in a table totaling the area each component category occupies within each facility. Next, the Python script conducts the K-Score statistical equations (Equations 4.1 – 4.3) that summarizes each component category by facility, and provides the frequency and area of the components found within each category. The resultant tables from the dissolve and frequency functions then become the input data for the calculation of the K max and K-Score calculations. The final results creates a new feature layer containing the facility name, K max, K-Scores, and normalized K-Scores for both frequency and area are saved as feature layers within the geodatabase.

The raw K-Score calculated utilizing equation 4-1 is normalized against the K-Score of the training data set by the formula in equation 4-3. Normalization has the effect of establishing the training set K-Score as absolute zero against which all other scores are calibrated.

$$\text{Normalized K-Score} = \frac{fo - fe}{1 - fe} \qquad (4\text{-}3)$$

The algorithm shown in (Figure 4-15) is used to extract facility training data, compute the K max and K-Scores, and the resultant K-Score feature layers.

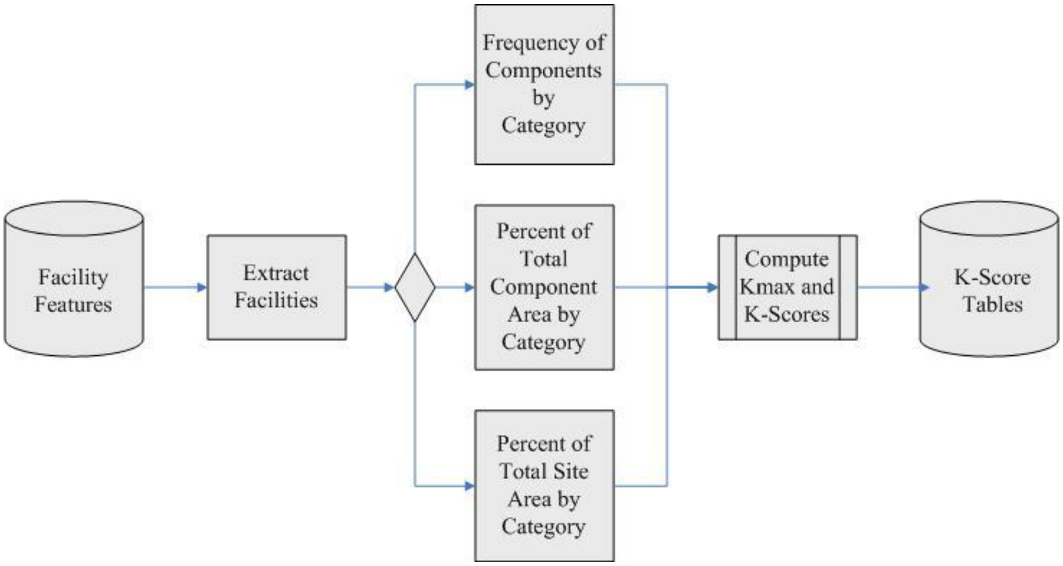
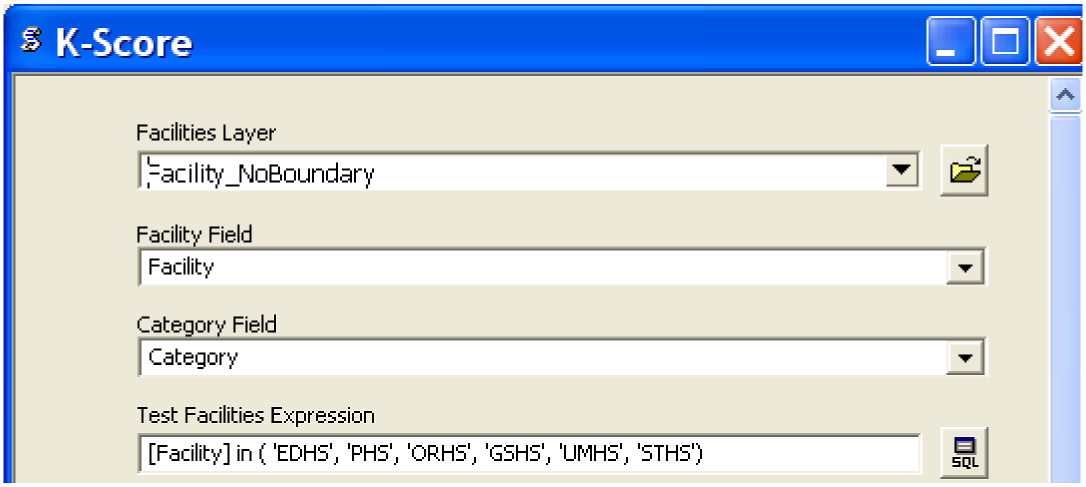

Facility Features
Extract Facilities
Frequency of Components by Category
Percent of Total Component Area by Category
Percent of Total Site Area by Category
Compute Kmax and K-Scores
K-Score Tables

Figure 4-15 Site Similarity Assessment Algorithm

4.3.4. Frequency of Components by Category

The first approach to assessing similarities of nominal data was to analyze categorical frequency distributions, by calculating the observed versus expected categorical frequencies. Frequency statistics derived from the training data (the expected set) are compared against the observed frequencies from the test data. This process begins with selecting the Conduct Analysis button on the SSA toolbar that opens the K-Score form (Figure 4-14). The user identifies the desired facilities layer from a drop-down list of feature layers within the ArcMap project table of contents, which in turn automatically fills in the ‘Facility’ and ‘Category’ fields. Secondly, the user creates an SQL expression to select the facilities that will compose the training data (Figure 4-16). Finally, the output location is set for the resultant K-Score feature layers created as part of the analysis process. The analysis begins when the user selects ‘OK’ OK at the bottom of the form.


K-Score

Facilities Layer
Facility_NoBoundary

Facility Field
Facility

Category Field
Category

Test Facilities Expression
[Facility] in ('EDHS', 'PHS', 'ORHS', 'GSHS', 'UMHS', 'STHS')

Figure 4-16 K-Score Field Parameters

The frequency distribution statistical analysis begins by dissolving (aggregating) the features within each facility by category, creating a new database table, adding a *Count* numerical attribute column, and providing a count sum. For example, a facility having twenty educational building polygons and corresponding records in a database table will be reduced to a single record within a new table with the *Count* column reading ’20’. This process is repeated for each category within every facility.

The second step calculates the mean observed values within each category of the training facilities identified by the user in the SQL statement. These mean values become the *expected* (f_e) values against which all observed values (f_o) within all the facilities are compared, resulting in the K max score (Equation 4-1). The K-Score is then calculated per Equation 4-2 and normalized by Equation 4-3, resulting in a new feature layer containing each facility and the resultant K-Score. Finally, the user selects *Reporting* Reporting on the SSA toolbar to obtain a report itemizing each facility sorted in ascending order by K-Score.

4.3.5. Percent of Total Component Area by Category

The second approach to assessing similarity was to evaluate the percentage of categorical component occupation across the facility, exclusive of all unoccupied space. This approach takes into consideration that a facility boundary may be unknown, and the possibility that the site extent may be unimportant. This approach excludes the unoccupied category. As depicted in Figure 4-17, the boundary of a facility may not be intuitive. Therefore, without ancillary information, such as a parcel layer or a priori knowledge, the extent of the site cannot be accurately determined. An assessment exclusive of a boundary then seeks to determine similarity solely on the utilization of occupied space. This analysis assesses only the total categorical occupation of all the site elements exclusive of all unoccupied space.

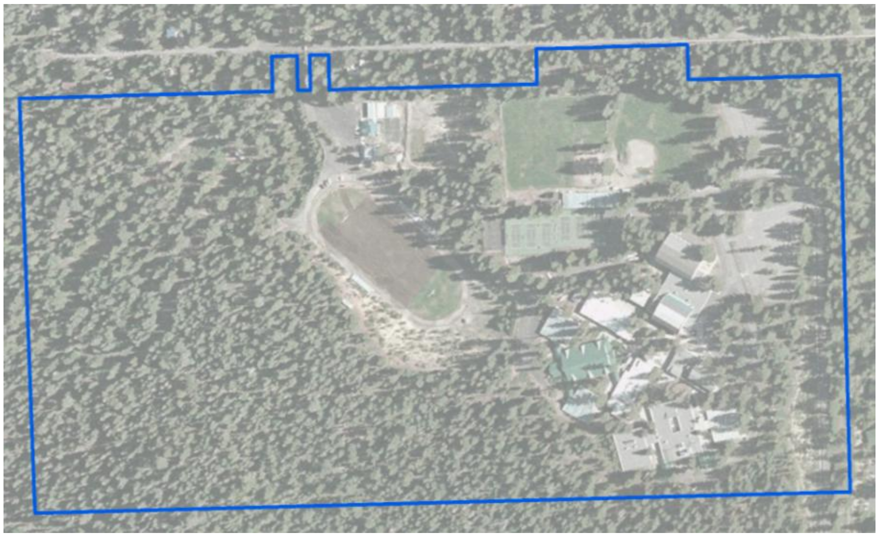


Figure 4-17 Indiscernible Site Boundary in El Dorado County, CA

(Image: ESRI's ArcGIS Image Server accessed 24 Oct 2007)

The total categorical occupation analysis process is quite similar to that of the frequency analysis. In the first step, all the categories are aggregated for each facility, resulting in a new feature layer with one row per category per facility. This process sums the total square footage of component occupation per category per facility. For example, a site having 20 educational buildings each occupying 4,000 square feet, is recorded as a single record having an area occupation of 80,000 square feet. The second step utilizes the same algorithms, equations, and processes as discussed in section 4.3.4. However, the K-Scores in this instance describe the percentage of total component area by category.

Finally, the user selects *Reporting*

on the SSA toolbar to obtain a report itemizing each facility sorted in ascending order by K-Score.

4.3.6. Percent of Total Site Area by Category

The approach used for this analysis considers unoccupied space within the site and seeks to determine whether total area is an important element. The methodology employed for this analysis is the same as described in sections 4.3.4 and 4.3.5, the only difference being the inclusion of the 'unoccupied' category. Since unoccupied space within a facility is the inverse of the occupied space, this area was derived by erasing all occupied areas from the boundary polygons. Finally, the user selects Reporting

on the SSA toolbar to obtain a report itemizing each facility sorted in ascending order by K-Score.

4.4. Methodology Summary

The SSA methodology utilizes vector data to allow the user to select one or more facilities to compose a training data set, against which test facilities are assessed for similarity. Three different similarity assessments are measured with this proof-of-concept prototype. First, component categorical frequencies measure the categorical distribution within and among the facilities. Second, only the occupied space within a site is analyzed with the percent categorical component occupation, exclusive of boundary method. Third, percent categorical component occupation, inclusive of boundary, expands upon the previous analysis to consider all unoccupied space within a sites' known extent.

5. Results

The methodology developed in this project assists in classifying target facilities based on a site definition derived from a training data set. The project data consisted of nineteen facilities: nine high schools, three elementary schools, three middle schools, and four professional sports facilities. From these nineteen, six sites were selected as training sites. The training sites represented a range of suburban and rural high schools. The goal for this project was to develop a methodology to assist in the identification of target facilities, which was accomplished.

Three measures of the attribute *site occupation by components* were developed: 1) frequency of components by category; 2) percent of total component area by category; and 3) percent of total site area by category. In each case, the relative similarity of the site in question to the training sties was measured using the k-score statistic (Equation 4-1) (Flewelling, 1997). This compares the observed frequency (f_o) for a component to its expected frequency (f_e) summed over all of the categories. The sum of the differences is normalized by the maximum possible difference (K max) (Equation 4-2). A score of zero implies no difference (100% similarity) and a score of 1.0 implies a maximum difference.

5.1. Derived Signatures

Three signatures were derived from the training set of facilities, one for each measure of categorical occupation (Table 5-1). These formed the basis against which facility similarity was assessed in the test dataset.

Table 5-1 Derived Signatures

Category	Frequency of Components by Category	Percent Total Component Area	Percent Total Site Area
Admin	3.6667	0.0269	0.0101
Athletics	19.3333	0.6333	0.2753
Education	19.5000	0.1427	0.0623
Parking	4.3333	0.1971	0.0800
Unoccupied	NA	NA	0.5359

5.2. Frequency of Components by Category

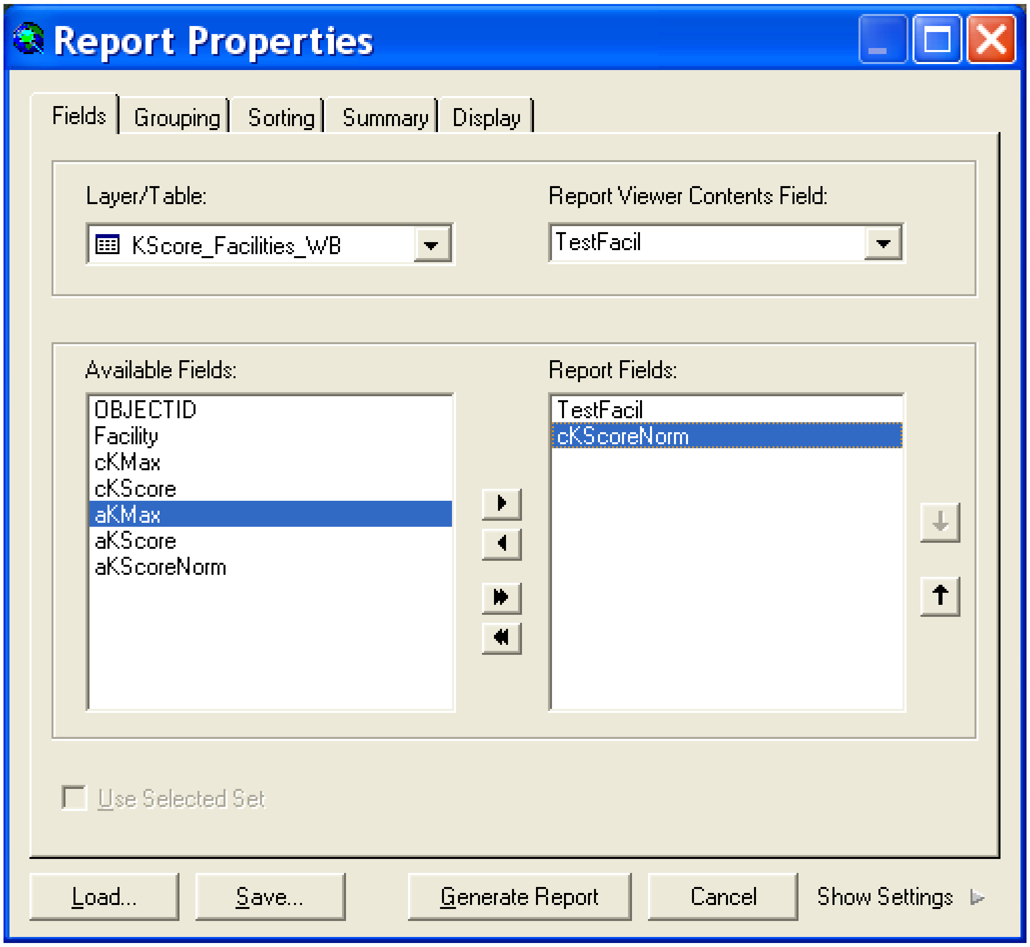
Initially, total component area occupation showed significance in assessing similarity. However, frequency of component categories ultimately proved to be the strongest similarity indicator for this project. The preliminary analysis focused on establishing a training set of six representative high school facilities. All nine high schools were then evaluated against the ideal for each of the three methodological approaches. These preliminary results revealed percent total component area occupation to be most significant. The introduction of lower echelon schools and non-school facilities into the test set rendered this an insignificant measure of similarity. The frequency at which the categorical components were observed became a strong indicator of similarity between sites with heterogeneous functions and echelons.

Initially, the extracted feature data sets were constructed inconsistently. A review of the feature extraction methodology revealed a difference in extraction method used for different types of athletic facilities. For example, the boundary of an entire set of tennis courts was extracted as a single feature. This resulted in a school with ten courts having the same count (one) as a school having three courts. The utilization of similar methodology when extracting basketball courts perpetuated the problem. In some instances, individual basketball courts were very difficult to count because court markings were faded, image resolution made it difficult to discern markings, or exposure and shadowing impeded identification. However, by consulting multiple project image sources, as well as Google Earth, an accurate count was finally determined. These changes resulted in a consistent extraction methodology that enables an objective analysis. Table 5-2 shows the resultant frequency of components by category.

Table 5-2 Frequency of Components by Category

Facility	Admin	Athletics	Education	Parking
SIGNATURE (*expected*)	**3.6667**	**19.3333**	**19.5000**	**4.3333**
Anaheim Angels Stadium	0	1	0	1
Candlestick Park	0	1	0	1
El Dorado High School	2	18	20	5
Golden Sierra High School	2	12	18	6
Gold Trail Middle School	3	3	8	3
Herbert Green Middle School	3	4	13	1
Highland High School	3	36	16	4
Hangtown Speedway	7	5	3	2
Hollywood Park	1	4	18	3
Indian Creek Elementary School	2	3	8	2
Louisiana Schnell Elementary School	1	2	12	1
Markum Middle School	2	3	12	1
Oak Ridge High School	4	23	20	4
Ponderosa High School	5	25	29	3
Redlands High School	4	25	32	3
Sutter's Mill Elementary School	1	3	10	3
South Tahoe High School	4	16	9	4
Union Mine High School	5	22	21	4
William S. Hart High School	2	20	29	4

Selecting the *Reporting* button on the SSA toolbar brings up the *Report Properties* dialog box (Figure 5-1) that guides the user through selecting the desired information to include in the results report. Figure 5-2 shows the resultant final report sorted in ascending K-Score order. Full facility names are listed at the beginning of the document in the List of Acronyms and Abbreviations.



Layer/Table: KScore_Facilities_WB

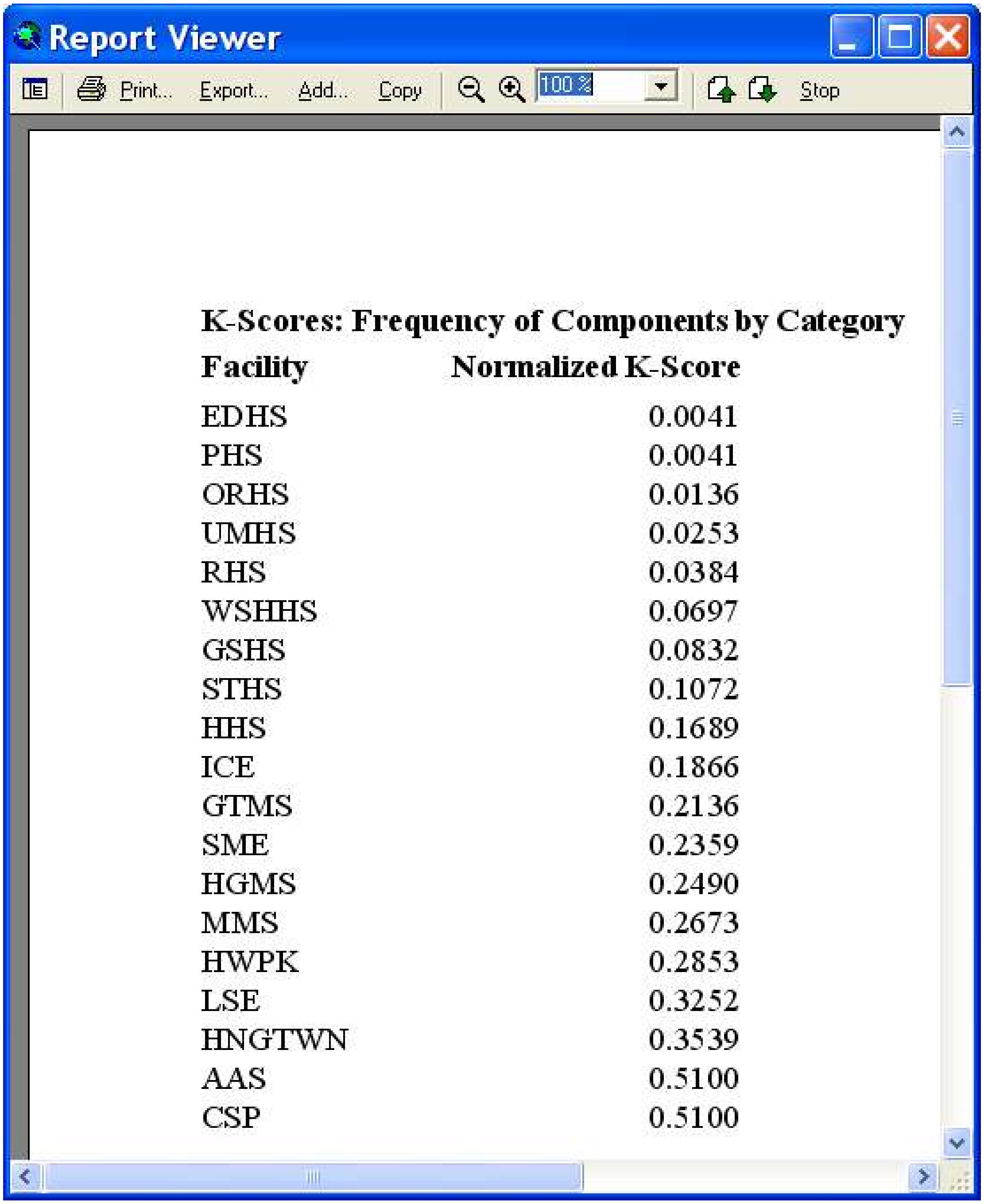
Report Viewer Contents Field: TestFacil

Available Fields:
OBJECTID
Facility
cKMax
cKScore
aKMax
aKScore
aKScoreNorm

Report Fields:
TestFacil
cKScoreNorm

□ Use Selected Set

Figure 5-1 Report Properties Dialog Box



K-Scores: Frequency of Components by Category

Facility	Normalized K-Score
EDHS	0.0041
PHS	0.0041
ORHS	0.0136
UMHS	0.0253
RHS	0.0384
WSHHS	0.0697
GSHS	0.0832
STHS	0.1072
HHS	0.1689
ICE	0.1866
GTMS	0.2136
SME	0.2359
HGMS	0.2490
MMS	0.2673
HWPK	0.2853
LSE	0.3252
HNGTWN	0.3539
AAS	0.5100
CSP	0.5100

Figure 5-2 K-Scores: Frequency of Components by Category Report

Table 5-3 and the corresponding chart (Figure 5-3) depict the final analysis results. Facilities within the table are sorted in ascending order by K-Score and grouped into three sections: A, B, and C, based on the delta of K-Scores between facilities, with significant deltas forming the sections. Different colors within the chart distinguish facility type, with sections A – C corresponding to the table. Similarity values range from 0.0041 (highly similar) to 0.5100 (dissimilar).

Table 5-3 K-Scores: Frequency of Components by Category

	Facility	K-Score
A	EDHS	0.0041
	PHS	0.0041
	ORHS	0.0136
	UMHS	0.0253
	RHS	0.0384
	WSHHS	0.0697
	GSHS	0.0832
	STHS	0.1072
B	HHS	0.1689
	ICE	0.1866
	GTMS	0.2136
	SME	0.2359
	HGMS	0.2490
	MMS	0.2673
	HWPK	0.2853
C	LSE	0.3252
	HNGTWN	0.3539
	AAS	0.5100
	CSP	0.5100

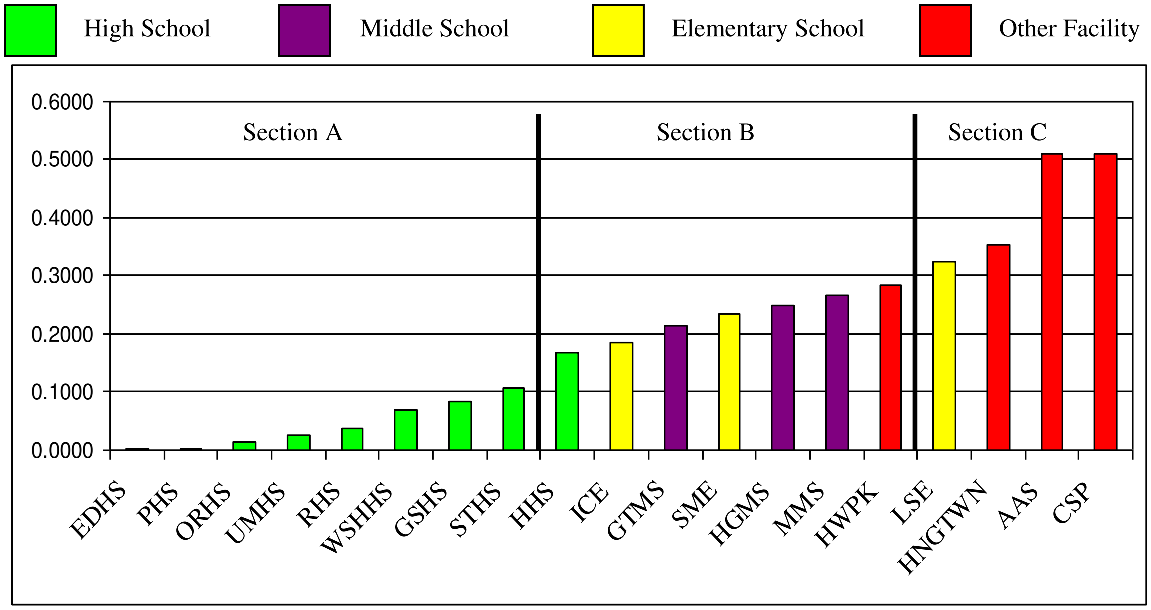


Figure 5-3 K-Scores: Frequency of Components by Category

In section A, eight of the nine (88%) high schools within the test data — and 100% in this section — were evaluated as being highly similar to the training data. South Tahoe High School (STHS), evaluated at the upper level of this section due to its low number of educational buildings. This school is located in a rural area, at an elevation exceeding 5,000 feet, and has an enrollment of 1,588 students. The buildings are multi-storied and designed to minimize the need for students to be exposed to snow and frigid temperatures during the winter.

Six of the seven facilities within Section B are schools, including one high school, in addition to one professional sports facility. The dissimilarity of Highland High School (HHS) to the training set is based on the presence of 36 athletic components compared to an expected 19.3. Upon further inspection, the high school has 13 basketball courts compared to four and eight for all the other high schools. Conversely, Hollywood Park, evaluated into the top of this section based on its low number of athletic components (four) compared to the expected 19. Two middle and three elementary schools round out this section.

Section C contains four facilities (one elementary, three professional sports) with the greatest dissimilarity to the training data. The dissimilarity of Louisiana Schnell Elementary (LSE) is based on having only two (2) athletic facilities compared to the prototypical 19. The lack of educational components at all three professional sports facilities result in their high dissimilarity assessments. Furthermore, all the facilities in this category have a single parking component compared to the four expected.

These scores accurately reflect what can be seen when looking at different facilities in imagery (Figure 5-4). There is a high degree of similarity between the number and types of components within high schools, with athletic facilities being quite prominent. Lower echelon schools have fewer components and different types of athletic facilities. Furthermore, the professional sport facilities exhibit extremely different patterns than do the schools.

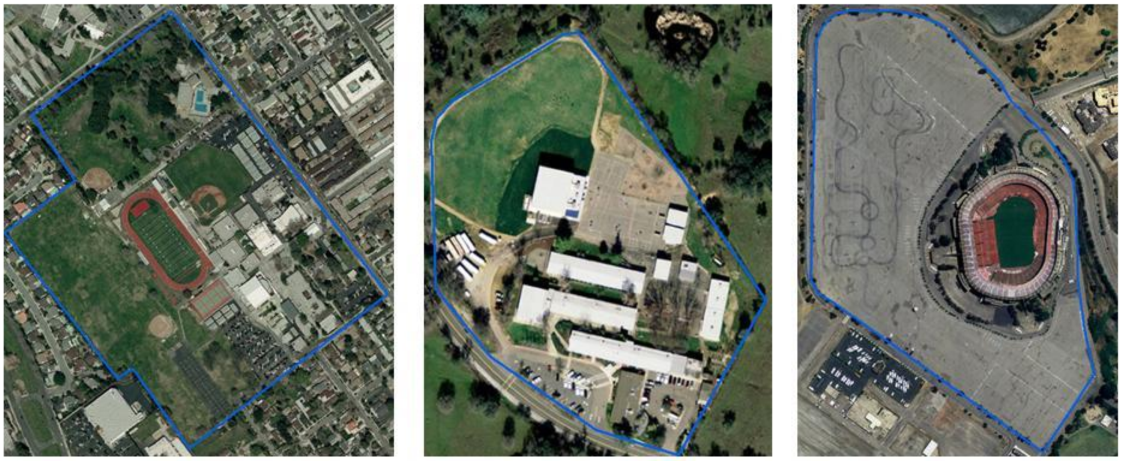


Figure 5-4 Facility Type Visual Differences

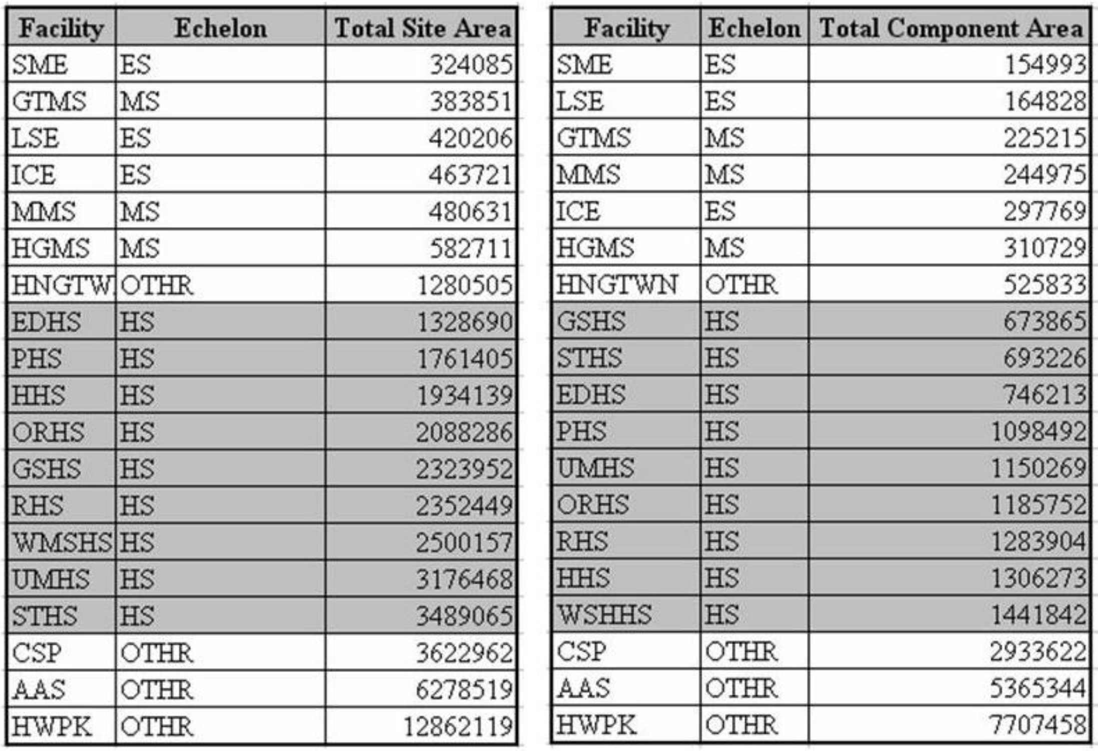
(Left to Right: High School, Middle School, and Professional Baseball Park)

(Image: ESRI's ArcGIS Image Server accessed 18 Nov 2007)

5.3. Raw Total Site and Component Area

The type of facility (high schools) selected for this project have a significant set of regulations and practices that constrain the number of students per class, student-to-teacher ratios, and the total number of students within the school itself. These criteria have an effect on the ultimate total area needed for different echelons of schools.

A simple study of site square footage shows a strong demarcation between elementary or middle schools (ES, MS) and high schools (HS). Elementary and middle schools require less area than do high schools, which in turn require less area than do professional sports facilities (OTHR). The only outlier HNGTWN – OTHR is primarily a county run fairgrounds, however it does host professional sprint car races on its quarter-mile dirt track. Total site and component square footage for all the facilities are shown in Figure 5-5 with the shading drawing attention to the fact that high schools require more area than do lower echelon schools but less than professional sports facilities.



Facility	Echelon	Total Site Area
SME	ES	324085
GTMS	MS	383851
LSE	ES	420206
ICE	ES	463721
MMS	MS	480631
HGMS	MS	582711
HNGTW	OTHR	1280505
EDHS	HS	1328690
PHS	HS	1761405
HHS	HS	1934139
ORHS	HS	2088286
GSHS	HS	2323952
RHS	HS	2352449
WMSHS	HS	2500157
UMHS	HS	3176468
STHS	HS	3489065
CSP	OTHR	3622962
AAS	OTHR	6278519
HWPK	OTHR	12862119

Facility	Echelon	Total Component Area
SME	ES	154993
LSE	ES	164828
GTMS	MS	225215
MMS	MS	244975
ICE	ES	297769
HGMS	MS	310729
HNGTWN	OTHR	525833
GSHS	HS	673865
STHS	HS	693226
EDHS	HS	746213
PHS	HS	1098492
UMHS	HS	1150269
ORHS	HS	1185752
RHS	HS	1283904
HHS	HS	1306273
WSHHS	HS	1441842
CSP	OTHR	2933622
AAS	OTHR	5365344
HWPK	OTHR	7707458

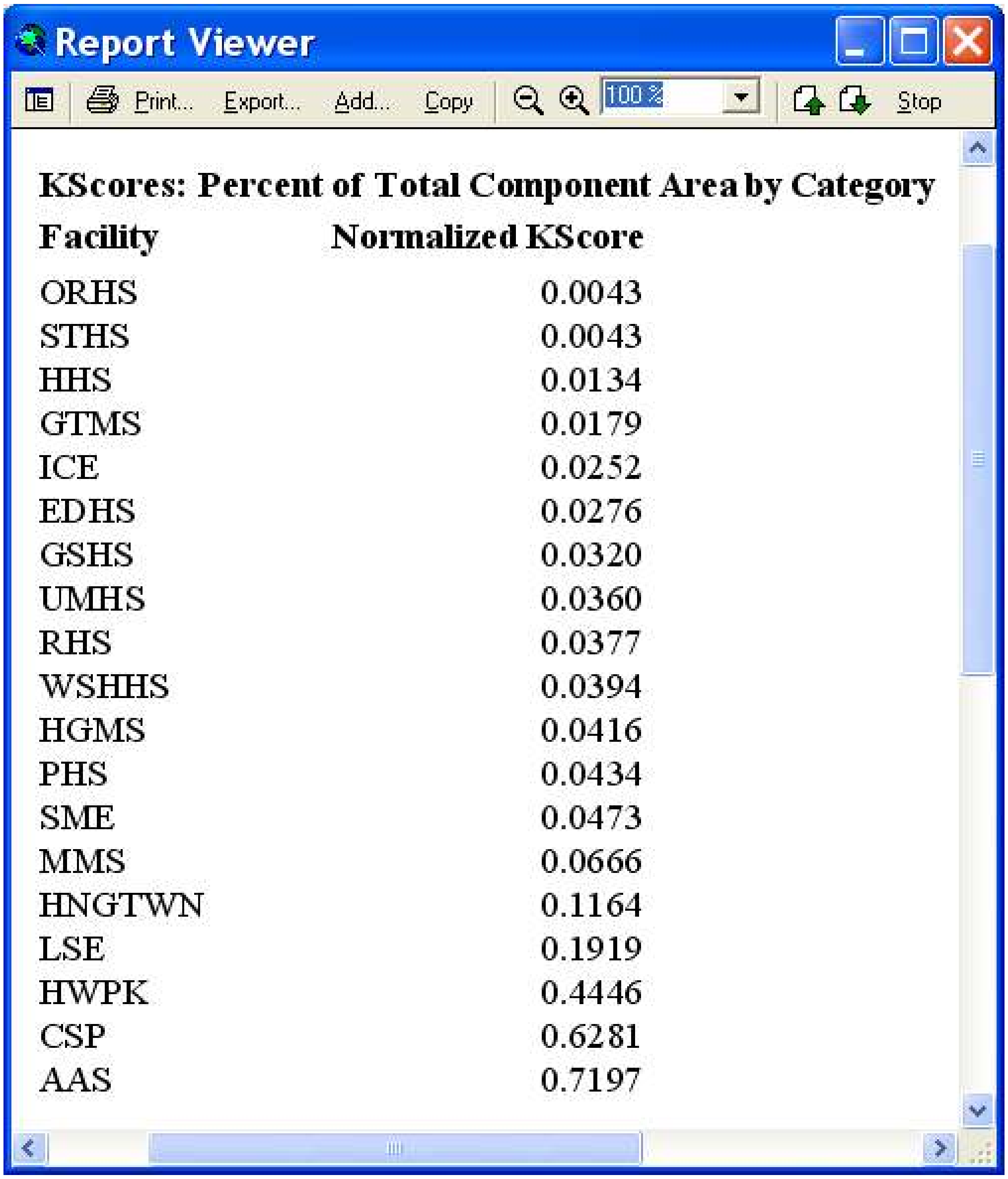
Figure 5-5 Total Site and Component Area

5.4. Percent of Total Component Area by Category

Data extraction methodology also proved crucial when examining component area. Athletic facilities, such as baseball and softball fields, and tennis courts, were originally digitized as multi-part features. This was so that they would visually represent their associated features. For example, baseball and softball field boundaries were digitized as one part and the infields were digitized separately. The resulting polygons depicted the outline of the entire field with a discernible infield. While this was useful for visual recognition, it had the effect of cutting out the infield from the entire ball field, reducing the total area for these components. In a similar way, tennis court boundaries were captured as one part of the polygon, while the actual courts, delineated by different surface color and court markings, were captured as additional parts of the polygon, having the effect of misrepresenting the actual total area as well.

The significance was not discovered until lower echelon schools were added to the test set. It became evident when the totals of areas represented by athletic facilities were questionably close between echelons. Elementary and middle schools typically have multi-purpose playgrounds and athletic fields without strongly delineated regulation sports fields and were therefore collected as a single component. When they were analyzed against high school facility K-Scores, the results showed no significant difference between the facility types. In order to determine if this was indeed an accurate result, each category was assessed individually. This analysis determined that the problem lay within the 'athletic' category. At this point, the extraction methodology was reviewed, and it became apparent that the discrepancy between the extraction methodology for high schools and the other schools was skewing the results. All the multi-part features within the high schools were then recollected as single-part features.

Table 5-4, Figure 5-6, and Figure 5-7 depict the final results. The table is sorted in ascending order by K-Score, per facility and has been grouped into three sections: A, B, and C. These groups are based on the delta of K-Scores between facilities, with significant deltas forming the sections. Each facility type is identified by a different color within the chart for visual clarity, with sections A – C corresponding to the table.



KScores: Percent of Total Component Area by Category

Facility	Normalized KScore
ORHS	0.0043
STHS	0.0043
HHS	0.0134
GTMS	0.0179
ICE	0.0252
EDHS	0.0276
GSHS	0.0320
UMHS	0.0360
RHS	0.0377
WSHHS	0.0394
HGMS	0.0416
PHS	0.0434
SME	0.0473
MMS	0.0666
HNGTWN	0.1164
LSE	0.1919
HWPK	0.4446
CSP	0.6281
AAS	0.7197

Figure 5-6 K-Score Percent Total Component Area by Category Report

Table 5-4 K-Scores: Percent of Total Component Area by Category

	Facility	K-Score
A	ORHS	0.0043
	STHS	0.0043
	HHS	0.0134
	GTMS	0.0179
	ICE	0.0252
	EDHS	0.0276
	GSHS	0.0320
	UMHS	0.0360
	RHS	0.0377
	WSHHS	0.0394
	HGMS	0.0416
	PHS	0.0434
	SME	0.0473
	MMS	0.0666
B	HNGTWN	0.1164
	LSE	0.1919
C	HWPK	0.4446
	CSP	0.6281
	AAS	0.7197

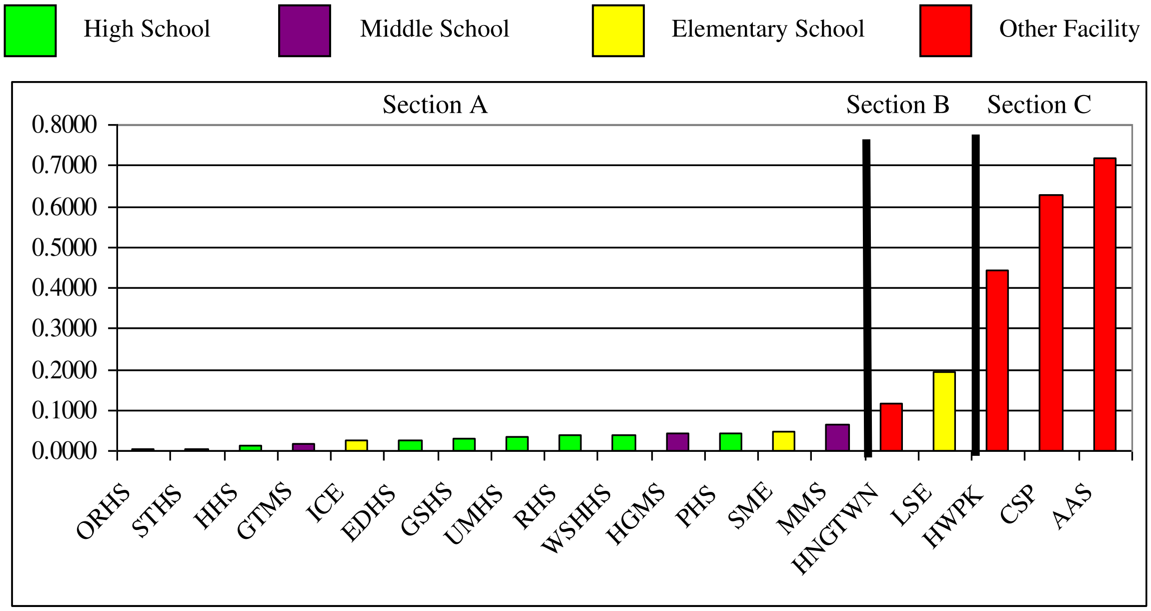


Figure 5-7 K-Scores: Percent of Total Component Area by Category

All fourteen of the facilities within section A are schools, therefore it might be concluded that schools generally tend to utilize occupied space in a similar manner, regardless of echelon.

Reasoning that the total enrollment might correlate to echelon, the enrollment for each school was compared. Enrollment varies for these 14 facilities between 213 (elementary school) to 3,452 (high school) students. The only outlier in this group is Louisiana Schnell elementary school (LSE), which has an unusually high percentage of its occupied space (30%) devoted to educational buildings compared with the mean (14%). The architecture of educational buildings within this site is substantially different from the other schools, especially within the lower echelon facilities. Figure 5-8 depicts these differences, with LSE having considerably larger buildings (all images have a nominal scale of 1:2000). LSE has 393 students enrolled, falling between the 213 enrolled at Indian Creek Elementary School (ICE) and 481 enrolled at Sutter's Mill Elementary School (SME), therefore its Charter School designation may be a contributing factor to its difference from its peers.

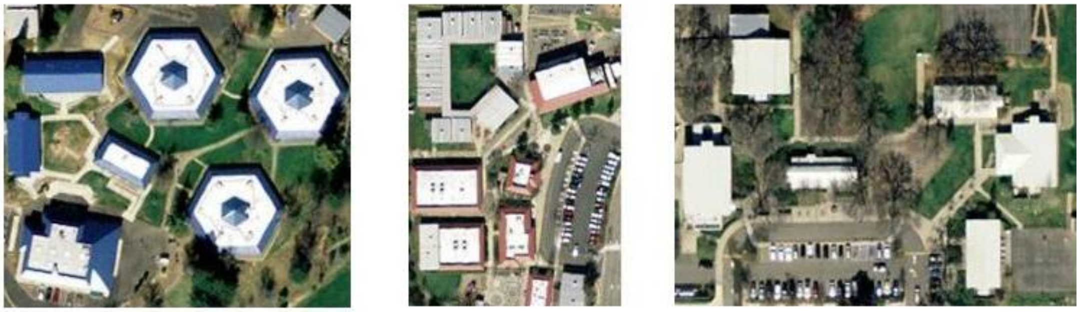


Figure 5-8 Elementary School Architectural Differences

(Left – Right: LSE, ICE, and SME)

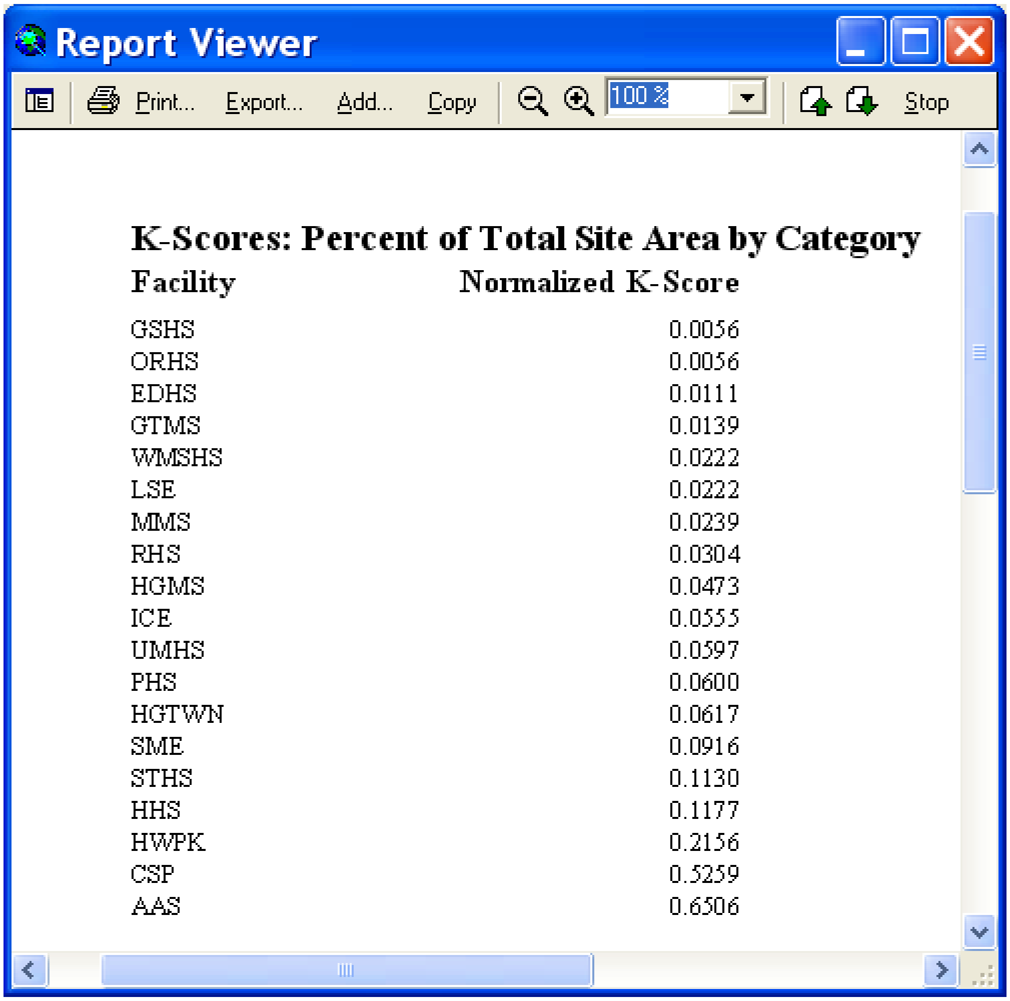
(Image: ESRI's ArcGIS Image Server accessed 08 Nov 2007)

Section B contains one professional sports facility (Hangtown Speedway, HNGTWN) and one elementary school (LSE). The dissimilarity of Hangtown Speedway is due to 34% (15% excess) of its component area being dedicated to parking rather than the 20% expected. Conversely, Louisiana Schnell's dissimilarity is based on a deficit within the athletic category of 23%, having only 40% dedicated to athletics rather than the expected 63%.

Section C is composed of three non-high school facilities. Hollywood Park's dissimilarity is based on having 34% component allocation for parking, exceeding the expected 20% by 15%. Both Candlestick Park and Anaheim Angels Stadium, professional baseball facilities, as expected are the most dissimilar in this evaluation due largely to having no educational facilities (100% deficit).

5.5. Percent of Total Site Area by Category

This assessment builds upon the previous categorical area assessment by considering the total site area. This was accomplished by adding the unoccupied, thus uncategorized, land on a site as a categorical component. The corresponding report is shown in Figure 5-9.



K-Scores: Percent of Total Site Area by Category

Facility	Normalized K-Score
GSHS	0.0056
ORHS	0.0056
EDHS	0.0111
GTMS	0.0139
WMSHS	0.0222
LSE	0.0222
MMS	0.0239
RHS	0.0304
HGMS	0.0473
ICE	0.0555
UMHS	0.0597
PHS	0.0600
HGTWN	0.0617
SME	0.0916
STHS	0.1130
HHS	0.1177
HWPK	0.2156
CSP	0.5259
AAS	0.6506

Figure 5-9 K-Scores: Percent of Total Site Area by Category Report

Analyzing the effects of including all unoccupied space within the site boundary again proved to be insignificant in measuring high school similarity (Table 5-5 and Figure 5-10).

Table 5-5 K-Scores: Percent of Total Site Area by Category

	Facility	KScore
A	GSHS	0.0056
	ORHS	0.0056
	EDHS	0.0111
	GTMS	0.0139
	WMSHS	0.0222
	LSE	0.0222
	MMS	0.0239
B	RHS	0.0304
	HGMS	0.0473
	ICE	0.0555
	UMHS	0.0597
	PHS	0.0600
	HGTWN	0.0617
C	SME	0.0916
	STHS	0.1130
	HHS	0.1177
D	HWPK	0.2156
E	CSP	0.5259
	AAS	0.6506

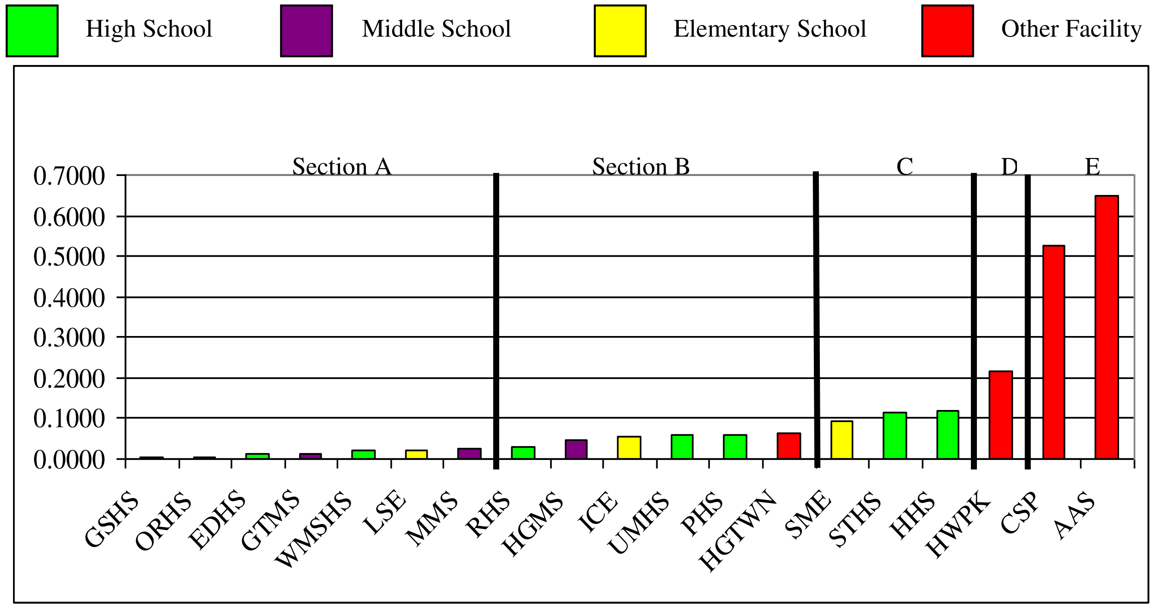


Figure 5-10 K-Score: Percent of Total Site Area by Category

This method also proved to be statistically insignificant. School facilities, regardless of echelon, were not distinguishable from non-school facilities. In addition, greater variance between the deltas of the K-Score broke the facilities into five groups. Area devoted to athletic components only contributed to four facilities (two high schools, one middle, and one elementary school) being evaluated as dissimilar to a prototypical high school. A deficit in area allocated to parking contributed to the dissimilarity of two professional sports facilities and one middle school. Hollywood Park exceeded the expected parking allocation of 8% by 31%. Unoccupied areas had the most significant impact on this measure, with 11 facilities ranging between seven and 31% below the expected value.

5.6. Results Discussion

Assessing frequency of components by category effectively identified high schools, with the exception of one outlier. This assessment cannot discriminate between lower echelons of schools indicated by the mix of elementary and middle schools in section B. Based on these results, the logical next step was to determine whether the raw total area of the sites or the components could discriminate echelons more effectively.

Based on the results of the different measures employed for this project, it appears that frequency of components, being the strongest indicator of similarity, may relate to the function of the facility. It is therefore logical that high schools would have the greatest number of components. Greater student enrollment at high schools necessitates a greater number of educational buildings. In addition, high schools participate in a variety of competitive sports, thus have athletic components not found at lower echelon schools. Finally, student-parking areas are located only at the high school level.

The simplistic area analysis readily discriminated between high schools, lower echelon schools, and non-school facilities, with only a single outlier. However, this is a gross metric. In order to determine if area occupation could provide a more refined result, categorical analysis of component and total site area was analyzed.

Assessing categorical component occupation discriminated between only school and non-school facilities, with one elementary school (LSE) being an outlier to these results. Additionally, this measurement introduced uncertainty between school echelons, therefore discounting it as an effective measurement tool for this project.

Assessing percent of total site area by category introduced the highest degree of statistical insignificant resulting in five clusters of facilities with high schools being spread throughout all five. This approach was not capable of distinguishing schools, regardless of echelon, from non-school facilities.

5.7. Results Summary

The site similarity assessment interface is able to accept user input to establish a training data set from which to calculate an ideal expected value for the type of categorical analysis desired. Additionally, it provides an algorithm to assess three different similarity measures, as well as creating new associated feature layers and reports of the results. The resulting feature layers contain a K-Score, which quantifies a facility's similarity to a derived ideal, while the reports provide the results textually. Frequency of component by category proved to be the strongest indicator to assess similarity, while percent of component area by category and percent of total site area by category produced inconclusive results.

6. Summary

The objectives for this project were to develop an assessment for classifying sites using site-based relationships to include:

- A formal site description including related subcomponents.
- Creation of a geodatabase to organize and manage facility data, site description, and analysis.
- Generation of a profile from control sites.
- Mechanism to assess test data against the control data profile.
- Capability to communicate results.

All of these objectives were met. The formal site description identified five categories into which high school components may be classified, as well as the variety of components that may be classified within those categories. The facility description was implemented with domains and subtypes. Domains constrain the category to which the components are classified. Components were restricted to administrative, athletic, educational, parking, or unoccupied categories. Subtypes constrain the value choices of the component type.

The personal geodatabase created for the project manages all the input data, such as domains, subtypes, facility data, and raster catalogs, in addition to all output data, such as facility signatures, K-Scores, and K-Score reports.

A site similarity assessment toolbar with customized buttons was created to facilitate analysis and report creation. A combination of model builder and Python scripting was used to create an interface that guides users through the process necessary for creating training data against which test facilities are assessed for similarity to the derived signature. Visual Basic for Applications and ArcObjects code was used to create reporting functionality from the site assessment toolbar.

Three measures of the attribute *site occupation by components* were developed: 1) frequency of components by category; 2) percent of total component area by category; and 3) percent of total site area by category. In each case, the relative similarity of the site in question to the training sites was measured using the K-Score, comparing the observed values to its expected frequency. Frequency of categorical by category proved to be the strongest indicator to assess similarity, correctly identifying 88 % of the high school facilities.

7. Future Work

This project has identified three research areas that may improve this methodology. The first is to incorporate contextual information as an additional measure of similarity. For example, environmental conditions, such as proximity to lines of communication, hydrologic features, or terrain features, such slope and aspect, may constrain where a facility under consideration may optimally be located. Certain types of facilities, depending on their nature, may need to be close to a water source, rail lines, or to terrain that is amenable to building tunnels. Conversely, context can also exclude areas for consideration. For example, sensitive facilities, such as power generation or chemical plants, may be less likely to be located near earthquake fault lines.

The second area that may be beneficial to investigate may be dependent on the complexity of the multi-component facility being evaluated. Blaser's (2000) work with scenes, which are conceptually the same as the sites discussed in this project, proved the concept of utilizing topological relationships, spatial location, orientation, and direction to evaluate similarity. Sites containing multi-component facilities that are composed of a fewer number of components, as well as a limited number of possible shapes associated with those components, may more reasonably be evaluated based upon topological relationships. This methodology would be able to assess similarity of sites even if the configuration of the components were quite different. Used in tandem with the previously mentioned contextual analysis may be quite valuable when trying to locate a facility that has proven difficult to find visually on imagery. In addition, metric refinements to the 9-Intersection model, such as outer and inner closeness, splitting, and alongness, introduced by Shariff (1996) would provide greater detail of the qualitative spatial relationships between features in a site.

A third area, which may ultimately prove to be the most promising, is object based image analysis. This approach may also be the most complex because it incorporates contextual feature identification and similarity rules within the extraction process. In addition, incorporating elevation data into the process would enable extracted features to have an elevation value, which enables 3-D representation and knowledge of feature heights. This could be an effective way to monitor sites over time to determine if there have been changes to the heights of buildings and thus to the potential capacity of the site. While the current software that is capable of this is extremely expensive and purportedly requires a significant learning curve to create accurate training date, the cost-benefit ratio may ultimately prove to be worth the investment.

8. References

Blaser, A. D. (2000). An Efficient and Representative Model to Capture Spatial Neighborhoods. *The First International Conference on Geographic Information Science* (*GIScience 2000),* Savannah, Georgia, October 28-31, 2000.

Blaser, A. D. (2000). *Sketching Spatial Queries.* Unpublished Dissertation, University of Maine, Orono, ME.

Bruns, T., & Egenhofer, M. J. (1996). Similarity of Spatial Scenes. *Proceedings of the 7th International Symposium on Spatial Data Handling (SDH 1996)* Delft, The Netherlands, August 1996.

Egenhofer, M. J. (1989). A formal definition of binary topological relationships. *Lecture Notes in Computer Science, 367*, 457-472.

Egenhofer, M. J. (1997). Query Processing in Spatial-Query-by-Sketch. *Journal of Visual Languages and Computing, 8*(4), 403-424.

Egenhofer, M. J., & Al-Taha, K. K. (1992). Reasoning about Gradual Changes of Topological Relationships. *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, 196-219.

Egenhofer, M. J., & Franzosa, R. (1991). Point-set topological spatial relations. *International Journal of Geographical Information Systems, 5*(2), 161-174.

Egenhofer, M. J., & Herring, J. R. (1994). Categorizing Binary Topological Relations Between Regions, Lines, and Points in Geographic Databases. Unpublished Technical Report. Department of Surveying and Engineering, University of Maine, Orono, ME.

Egenhofer, M. J., Sharma, J., & Mark, D. (1993). A Critical Comparison of the 4-Intersection and 9-Intersection Models for Spatial Relations: Formal Analysis. *Auto Carto 11*, 1-11 Minneapolis, MN, October, 1993.

Flewelling, D. M. (1997). *Comparing Subsets from Digital Spatial Archives: Point Set Similarity.* Unpublished Dissertation, University of Maine, Orono, ME.

Frank, A. U. (1996). Qualitative spatial reasoning: cardinal directions as an example. *International Journal of Geographical Information Science, 10*(3), 269-290.

Hong, J. H., Egenhofer, M. J., & Frank, A. (1995). On the Robustness of Qualitative Distance-and Direction Reasoning. *Autocarto, 12, 301-310*, Charlotte, NC, February 1995.

Hornsby, K., & Egenhofer, M. J. (1998). Identity-based change operations for composite objects. *Eighth International Symposium on Spatial Data Handling*, 202-213.

Longley, P., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). *Geographical Information Systems and Science*: West Sussex, England: Wiley.

O'Sullivan, D., & Unwin, D. J. (2002). *Geographic Information Analysis*: Hoboken, New Jersey: Wiley.

Papadias, D., Egenhofer, M. J., & Sharma, J. (1996). *4th ACM Workshop on Advances in Geographic Information Systems (ACM Press), Hierarchical reasoning about direction relations*. Rockville, Maryland, November 1996.

Shariff, A. (1996). *Natural-language spatial relations: Metric refinements of topological properties.* in Spatial Information Science and Engineering, University of Maine.

Shariff, A., Egenhofer, M. J., & Mark, D. M. (1998). Natural-Language Spatial Relations Between Linear and Areal Objects: The Topology and Metric of English-Language Terms. *International Journal of Geographical Information Science, 12*(3), 215-245.

Sharma, J., Flewelling, D. M., & Egenhofer, M. J. (1994). A Qualitative Spatial Reasoner. *Sixth International Symposium on Spatial Data Handling*, Edinburgh, Scotland, 665-681.

Stevens, S. S. (1946). On the theory of scales of measurements. *Science, 103*, 667-380.

Tversky, A. (1977). Features of Similarity. *Psychological Review, 84*(4), 327-352.

Wong, D. W. S., & Lee, J. (2005). *Statistical Analysis of Geographic Information with ArcView GIS and ArcGIS*: Hoboken, New Jersey: Wiley.

Appendix A – Geodatabase Contents

The site similarity assessment methodology was designed to work with a personal geodatabase. Refer to Appendix A Table 2 for a list of the data contained within the geodatabase.

Appendix A Table 1 Database Contents

Dataset	Description	Source and Date	Data Model	Attributes
Facility_ NoBoundary	Feature Extracted Facility Components excludes unoccupied area	Derived	Polygon (vector)	OBJECTID, SHAPE, FacilityName, Feature, Category, FacilCat, Shape_Length, Shape_Area
Facility_ WBoundary	Feature Extracted Facility Components includes unoccupied area	Derived	Polygon (vector)	OBJECTID, SHAPE, FacilityName, Feature, Category, FacilCat, Shape_Length, Shape_Area
Boundaries	Total Area Occupied by Facility	Derived	Polygon (vector)	OBJECTID, SHAPE, FacName, Shape_Length, Shape_Area
edcSchools03	Aerial Photography, 2 foot resolution	AirPhoto, USA, 2003	Raster Catalog	OBJECTID, Shape, Raster, Name, Shape_Length, Shape_Area
edcSchools04	Aerial Photography, 2 foot resolution	AirPhoto, USA, 2004	Raster Catalog	OBJECTID, Shape, Raster, Name, Shape_Length, Shape_Area

Dataset	Description	Source and Date	Data Model	Attributes
Signature_NB	Frequency and % Component Area Signatures	Derived	Database Table	OBJECTID, Category, cExp, pExp, aExp
Signature_WB	Total Site Area Signature	Derived	Database Table	OBJECTID, Category, cExp, pExp, aExp
K-Score_Facility _NB	K-Scores for Frequency and Component Area	Derived	Feature Class	OBJECITD, Facility, TestFacil, cKMax, cK-Score, cK-ScoreNorm, aKMax, aK-Score, aK-ScoreNorm
K-Score_Facility _WB	K-Scores for Percent Total Site Area	Derived	Feature Class	OBJECITD, Facility, TestFacil, cKMax, cK-Score, cK-ScoreNorm, aKMax, aK-Score, aK-ScoreNorm

The domains and values used by the Facilities feature class are listed in Appendix A Table 3. The domains were implemented to ensure data integrity. The component categories and associated subtypes used by the Facilities feature class are listed in Appendix A Table 4. The component categories were implemented to conduct statistical analysis from and the subtypes were implemented to enforce data integrity.

Appendix A Table 2 Facility Domain

Domain Name	**Coded Values**
Category	ATHL – Athletic ADMIN – Administrative EDUC – Educational PRKG – Parking UNOCC - Unoccupied
FacilityName	AAS – Anaheim Angels Stadium CSP – Candlestick Park EDHS – El Dorado High School GSHS – Golden Sierra High School GTMS – Gold Trail Middle School HGMS – Herbert Green Middle School HGTWN – Hangtown Speedway HHS – Highland High School HWPK – Hollywood Park ICE – Indian Creek Elementary School LSE – Louisiana Schnell Elementary School MMS – Markum Middle School ORHS – Oak Ridge High School PHS – Ponderosa High School RHS – Redlands High School SME – Sutter's Mill Elementary School STHS – South Tahoe High School UMHS – Union Mine High School WMSHS – William S. Hart High School

Appendix A Table 3 Facility_NoBoundary Feature Class Attributes

Field Name	**Description**	**Data Type**
OBJECTID	ArcGIS system generated	Object ID
SHAPE	ArcGIS system generated	Geometry
FacilityName	Name of facility	Text
Feature	Type of component	Short Integer
Category	Component category	Text
FacilCat	Concatenation of FacilityName and Category used to derive component frequency counts per facility and total area per category per facility	Text
SHAPE_Length	ArcGIS system generated	Double
SHAPE_Area	ArcGIS system generated	Double

Appendix A Table 4 Facility_WBoundary Feature Class Attributes

Field Name	**Description**	**Data Type**
OBJECTID	ArcGIS system generated	Object ID
SHAPE	ArcGIS system generated	Geometry
FacilityName	Name of facility	Text
Feature	Type of component	Short Integer
Category	Component category	Text
FacilCat	Concatenation of FacilityName and Category used to derive component frequency counts per facility and total area per category per facility	Text
SHAPE_Length	ArcGIS system generated	Double
SHAPE_Area	ArcGIS system generated	Double

Appendix A Table 5 Boundaries Feature Class Attributes

Field Name	Description	Data Type
OBJECTID	ArcGIS system generated	Object ID
Shape	ArcGIS system generated	Geometry
FacilityName	Name of facility	Text
SHAPE_Length	ArcGIS system generated	Double
SHAPE_Area	ArcGIS system generated	Double