BENDRASIS PROGRAMAVIMO DOKUMENTAS

EUROPOS SĄJUNGA

NACIONALINĖ ŽEMĖS TARNYBA
PRIE ŽEMĖS ŪKIO MINISTERIJOS

THE BASIS OF AN EU CO-FINANCED PROJECT "QUALIFICATION IMPROVEMENT OF THE LITHUANIAN GEOGRAPHICAL INFORMATION MANAGERS" (NO BPD2004-ESF-2.2.0.-02-05/0143)

Civil servants learning programme

DISTANCE LEARNING OF GEOGRAPHIC INFORMATION INFRASTRUCTURE

Training Material

APPLICATIONS OF GEOGRAPHIC INFORMATION INFRASTRUCTURE

GII-04

Vilnius, 2008

MALASPINA
UNIVERSITY-COLLEGE

HNIT - BALTIC

Training material „Applications of Geographic information infrastructure“ (GII-04)

Authors

1, 2, 3, 5 Modules - Brad Maguire

4 Module - dr. Gennady Gienko

Reviewed and edited by

doc. dr. Gintautas Mozgeris (Lithuanian University of Agriculture)

Reviewed by

doc. dr. Žilvinas Stankevičius (Vilnius Gediminas Technical University)

From English translated and edited by

Astraneta UAB

Study material for civil servants distance learning programme “GII Distance Learning”, which has been approved by a law “RE Approval of Study Programmes for Civil Servants” (No 27V-72; Apr 17, 2007) by the State Service Department under the Ministry of the Interior has been prepared on the basis of an EU co-financed project “Qualification Improvement of the Lithuanian Geographical Information Managers” (No BPD2004-ESF-2.2.0.-02-05/0143)

Beneficiary of the project National Land Service under the Ministry of Agriculture.

Contractor of the project HNIT-BALTIC UAB, Vilnius Gediminas Technical University

Table of Contents

1 Introduction to Geographic Information Infrastructures

Outline:

1. Geographic Information Infrastructures (GII)
2. Components of GII
3. Investment in GII
4. Development of GII
5. Examples of GII
6. Applications of GII

1.1 Geographic Information Infrastructures (GIIs)

1.1.1 Introduction

One of the foundations of modern societies is the accessibility of services to make people's lives easier and more efficient. Such services may be human or information-based. Human-based services include transportation, garbage disposal, medical services, ambulance and police services, national defence, and education. Information-based services include telephone services, television, and the provision of data on the Internet through websites, e-mail, or communications programs.

These services require an infrastructure in order to run. Infrastructure may take the form of roads and railways, ports, hospitals, schools, police and fire stations, other public buildings, the electrical grid, the telephone network, water and sewer systems, and the physical aspect of the Internet (i.e. fibre-optic, ADSL, cable networks for data transmission, equipment for wireless networks, and computer servers to support data transmission), among others.

These services may be provided by the government, by private industry, or by non-governmental organizations. The fundamental economic reality underlying the provision of services by government is that it is less expensive per capita to provide the services for all of society then it is for private or non-governmental organizations to provide the same services. Clearly, governments can provide some of these services less expensively because they have the capacity to obtain economies of scale that would be unavailable to smaller organizations.

However, this is not true for all services, so the degree to which governments provide services has both an economic and a political dimension. The people of the nation may decide that it is worth providing services such as advanced education, not because it is necessarily less expensive for a government to provide, but because this is an expression of the type of country that the people want to live in. There may be services such as military defence, which cannot be made "profitable" in any way, that governments undertake to provide because they help to ensure a society's survival.

When governments provide services, the people who benefit from the services are able to be more productive. Take the obvious example of the provision of sanitary sewers. If the government did not provide sanitary sewers, individuals would have to go to great lengths to provide such services for themselves (for example, to construct septic fields), or alternately, they might succumb to serious diseases such as cholera. On the other hand, by pooling the resources of millions of people, governments are able to provide this service relatively inexpensively, and enable people to use their time more productively for other pursuits.

For centuries, governments have produced topographic maps, primarily to support defensive military operations (Figure 1). Such general-purpose maps have proven to be very useful for a large variety of other users. Efforts to coordinate map production internationally have been going on since 1891, when in the International Map of the World project was initiated. Such efforts have intensified since the 1960s, and now, in the 21st Century, mapping programs are being extended into the digital domain to create Geographic Information Infrastructures (GII). GIIs are also known as Geospatial Information Infrastructures or Geospatial Data Infrastructures (GDI or GSDI). As with many new technologies, there are a number of names currently in use, but we will use the term Geographic Information Infrastructure (GII) throughout this course.

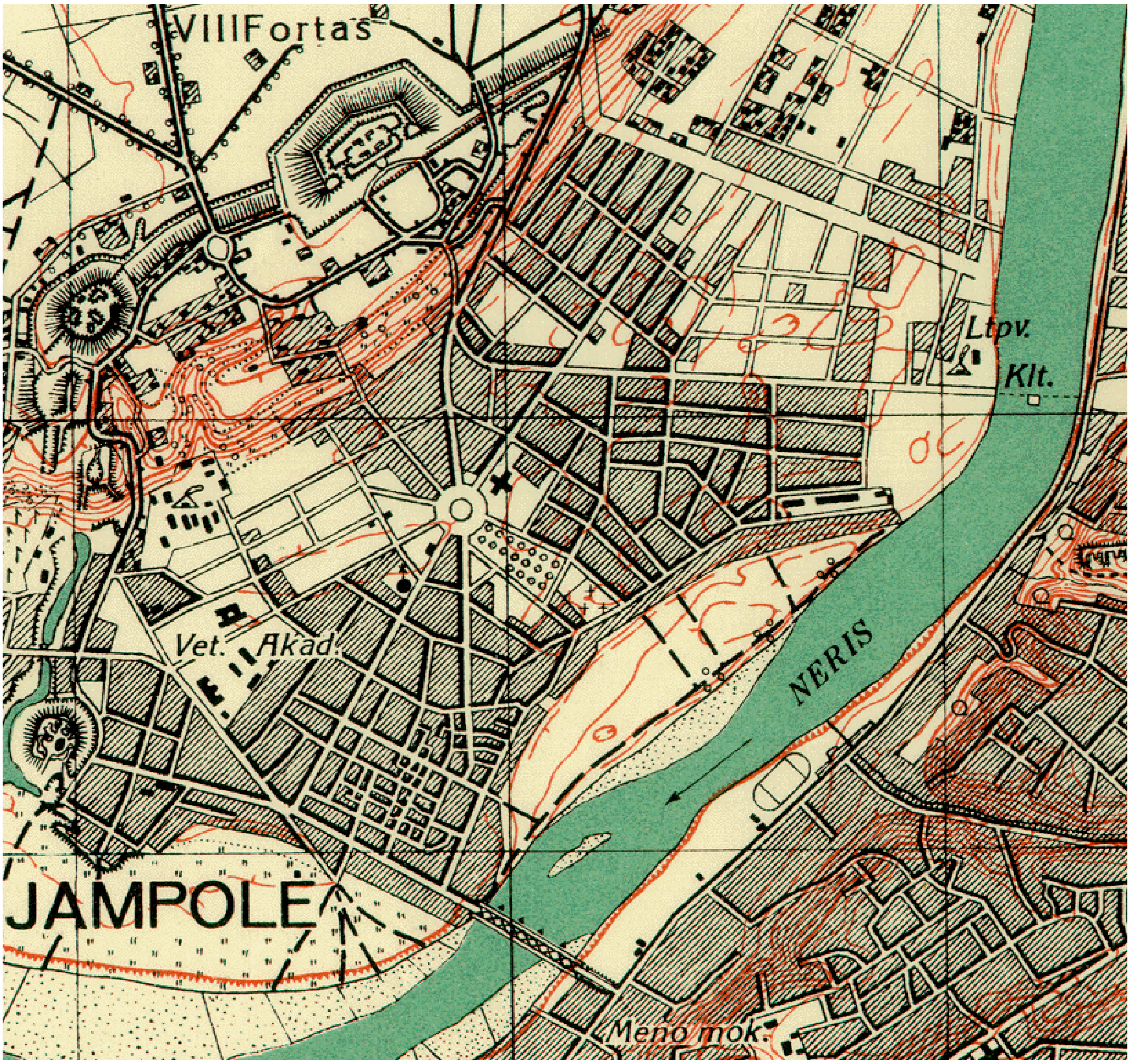


Figure 1. Detail of a Lithuanian topographic map from about 1930 showing a portion of Kaunas (Source: http://www.lib.berkeley.edu/EART/tour/topo.html)

Geographic Information Infrastructure is an infrastructure of spatial data including maps and geographically referenced databases that can be used by all of society. By taking the time to organize spatial data and make it widely available, Lithuania can reduce the cost of government operations, but more importantly, it can set the stage for the development of an Information Society, in which spatial data are available to everyone. In such a society, governments, organizations and individuals are able to make better decisions because they have the spatial information that they require.

Not only is information more freely available, but the quality of that information is better as well. Because duplication of effort and inefficiencies are reduced, more resources are available for the production of better spatial data. In addition, because many different agencies and individuals are involved, the spatial data that is produced is better designed for general use.

1.1.2 Purpose of GII

As with topographic maps, large amounts of spatial data can be shared by many different users, within government, private industry, or for personal use. If these data were standardized and made available to all, it would be relatively inexpensive, and could be used for a wide variety of purposes. The structuring and distribution of digital spatial data is a relatively expensive process, so it makes sense for individuals and organizations to share data when possible. Of course, sharing the data adds its own level of complexity, since standards must be developed, complied with, and documented if everybody is to be able to use the collective data.

Spatial information can be divided into three types: Foundation Data, Framework Data, and Application-Specific Data. Foundation Data is analogous to the information collected for topographic map sheets. This is the very general base data, and can be used by a wide variety of users throughout society. Foundation Data can be used so widely that it should be collected with the intention of sharing the data. Framework Data may also be used by a number of organizations, although it is probably not worth the effort to specifically collect this data with data sharing in mind. Application-Specific Data is only useful to the organization that created it, and investing time and money to make it shareable within a GII is not a good idea. (Groot and McLaughlin, 2000)

If a GII is implemented properly, spatial data can be used throughout society at many different levels, just as are topographic maps. When geographic data becomes widely available, it can lead to new applications that were not previously possible. Take, for example, the capabilities provided by Google Maps, which requires national road networks and worldwide satellite and air photo coverage to operate. Without these fundamental sources of spatial information, the application would not be possible (Figure 2).

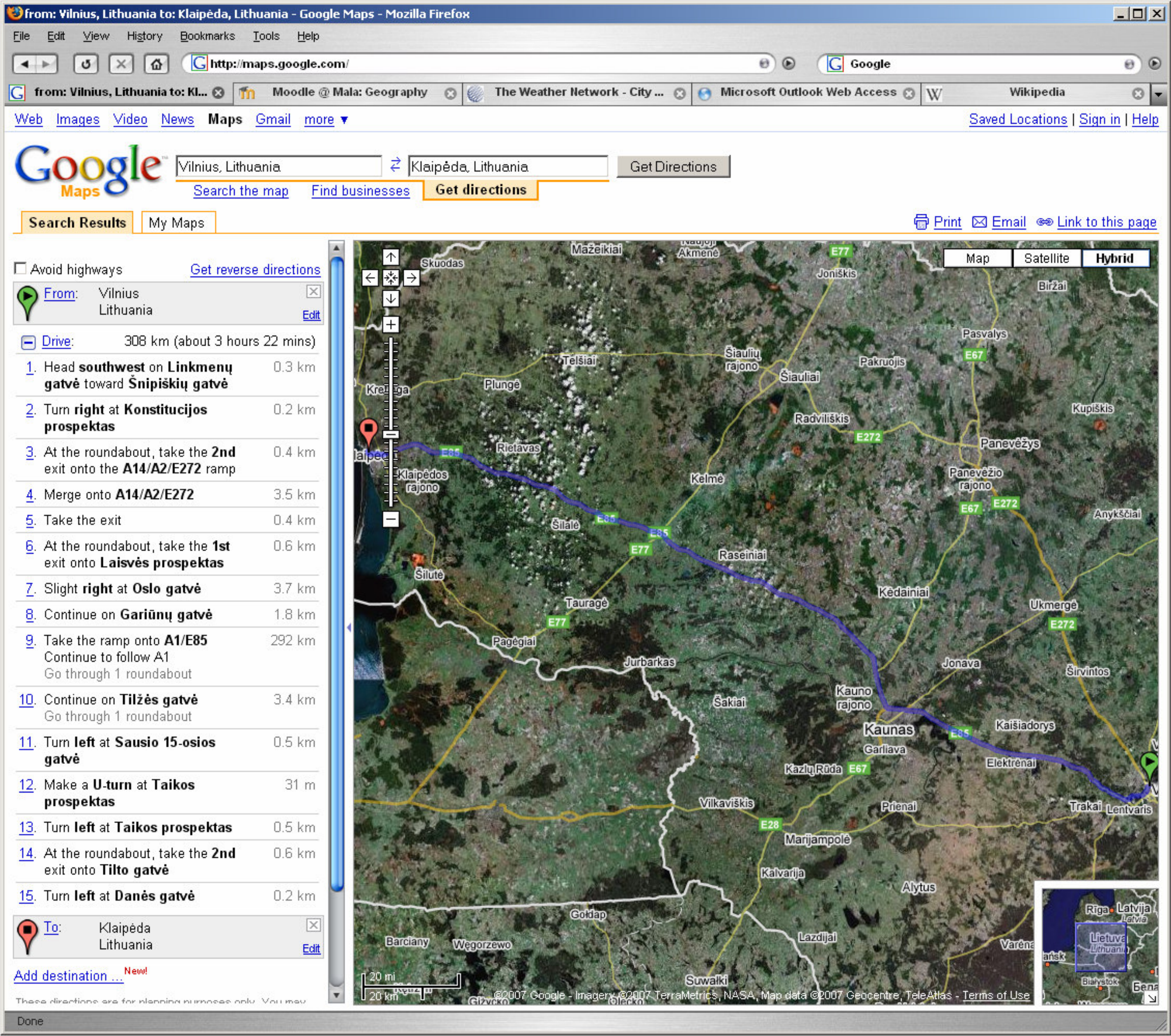


Figure 2. Directions from Vilnius to Klaipėda in Google Maps

The increasing availability of Global Positioning System (GPS) receivers means that people will come to expect to know exactly where they are at all times. Now, just as we expect to know the time whenever we glance at our watch, it seems likely that in future, people will become accustomed to knowing their exact position. The value of this information is enhanced significantly if this positional information is supplemented with accurate spatial data that describes surrounding features. In the words of Lance McKee, (Groot and McLaughlin, p. 15) Geospatial technologies will eventually become "mundane," meaning that they are part of an individual's environment just like the knowledge of time.

The combination of GPS and accurate Digital Spatial Data within a small, hand-held unit such as a cellular telephone, promises a revolution in the way that people use and interact with space. In a few years, it will be possible, using Location-Based Services (LBS), to foresee a situation such as the following:

> Imagine a future where you can say to your personal appliance (maybe on your wrist) 'I'm hungry, show me the nearest pizzeria.' Imagine that the appliance performs speech recognition (or wirelessly connects

> to a service that does). By wireless networks, a trader in restaurant information is found and queried, and through it a database of pizzerias is accessed. The personal appliance reports your GPS position to the database, and the nearest three pizzerias are located. These locations, and your location (and perhaps other information, such as the map projection and scale, and a symbol library you understand) are handed to a map generation service (somewhere on the web), which creates a digital map showing three routes to three restaurants. This map is sent to your appliance where it is displayed for you. Perhaps the display also gives information on today's specials, and the approximate waiting times at each pizzeria. (**Klaida! Nerastas nuorodos šaltinis.**) (Clifford Kottman, quoted in Groot and McLaughlin, p. 18)

Figure 3. Location Based Services to find nearest pizzeria

Although this example is somewhat trivial, it shows what might be possible through the combination of GII and GPS. The ability to be able to find medical help, the nearest police officer, or for military officers to be able to know the location of all soldiers on the battlefield at all times (http://science.howstuffworks.com/ffw2.htm) are far more significant, and are likely to be as great a contributor to feelings of "security" as are cellular telephones, automobiles, and enhanced 911 services today.

So in addition to the immediate benefits of GII to governmental and private users of spatial data, there are a number of near- and long-term benefits as well. Investment in GII is justified based on the immediate benefits, but the future promise lies in the new investments that will arise from the construction of the Information Society. Near-term benefits, such as LBS are a virtual certainty, given the advances in computer and communications technology that are underway. The longer-term benefits involve a great deal of speculation, since we have never lived in an Information Society, but there are likely to be one or two significant applications that nobody has yet envisioned.

1.2 Components of GII

Unlike other infrastructure products, Geographic Information Infrastructure is in many ways invisible. This is because GII is an infrastructure of *information*, and the computers on which this information is stored are located in climate-controlled server rooms in non-descript government buildings. Even if we include the Internet, which is a critical component for distributing the information stored in a GII, there is very little to see, except for telephone wires and buried fibre-optic cables. Contrast this with road infrastructure, which is clearly visible.

One component of GII is a group of networked computer servers known as a Geospatial Clearinghouse. The function of a clearinghouse is to make spatial data available to the public across the Internet. Typically, a clearinghouse is used to network existing computer servers in those agencies that are the primary custodian for particular types of spatial data. Thus, other than some new servers that maintain a continually updated database of metadata describing the content of the existing servers, we only need to ensure that the computers are properly programmed and networked to create a clearinghouse.

Typically, there are two types of network servers in a clearinghouse. A global server operates a database that contains descriptions of information that is available in the GII, and indexes where the data can be found. The global server has four possible functions:

- Control of transactions
- Analysis of requests
- Reception of update information from local servers, and
- Maintenance and updating of data directories

When a user selects information, the location of that information is provided to the user. The user then queries a database on a local server where the data itself is stored. This then allows the data to be downloaded by the user. The reason for this arrangement is that GII data may be held by different agencies on different computers. Since the volume of information is so large, having multiple server computers for different portions of the data makes it possible for the GII to be scaled up when the demand for information increases (Figure 4). One advantage in using computers located in the offices of the data custodians is that local control of the data is maintained. Confidential data remains confidential, and public access data is made publicly available on a regular basis. As soon as data updates have been approved for public release, they are placed on the local server for distribution. The global server is automatically notified that data has been updated, and its database of metadata is updated to reflect the changes on the local server.

It is also possible to combine the functions of the global and local servers in a single computer, but this solution is only suitable for relatively small amounts of data being accessed by few people, since a single computer has to perform many functions.

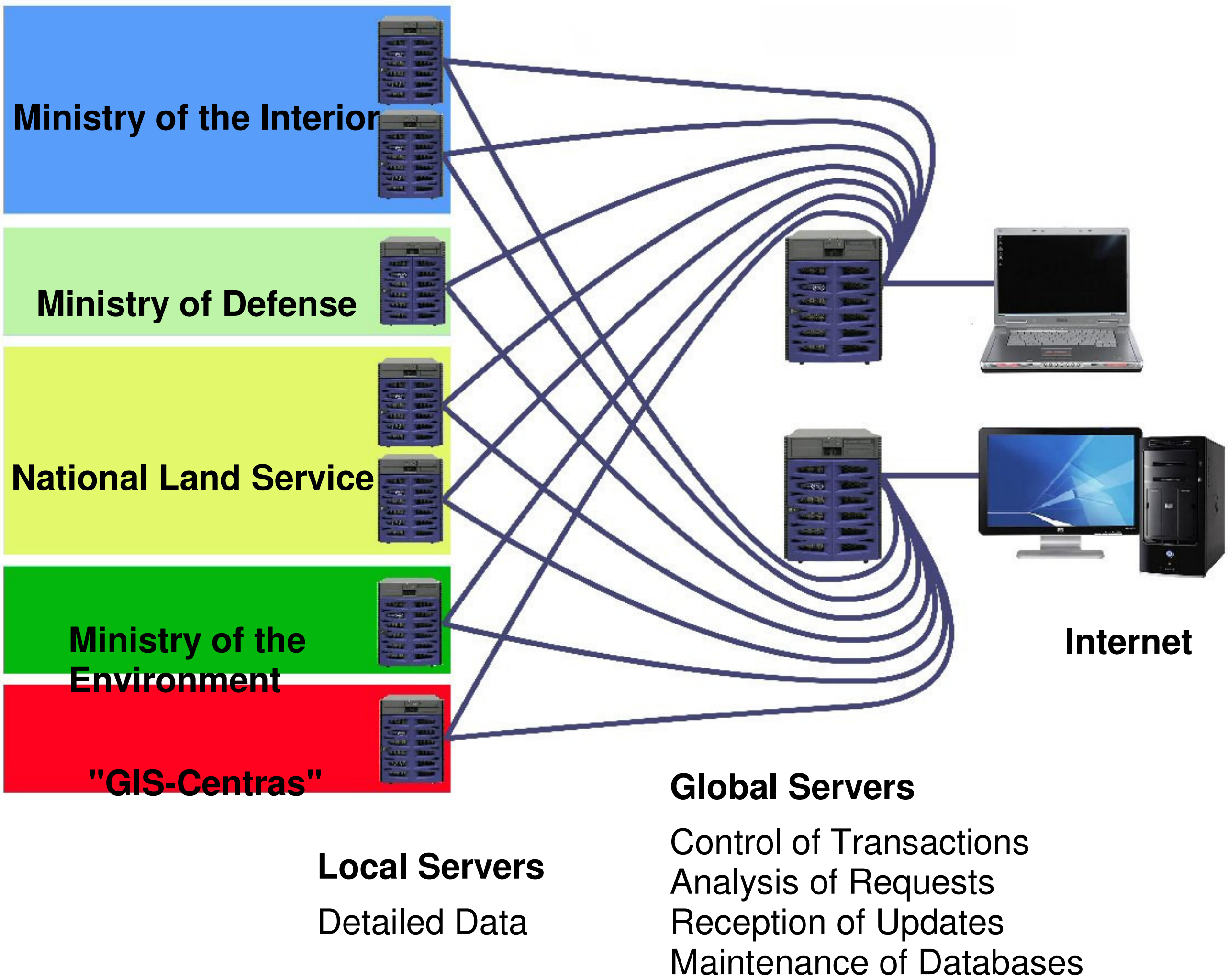


Figure 4. Local and Global Servers are used to distribute spatial data to the Internet in the GII. Note: Relative size of each agency and number of servers is only for illustrative purposes. Many other groups involved with GII are not shown.

GII has a large software component, and consists of many spatial and non-spatial databases running on server computers. These databases interact with one another to provide information such as descriptive metadata, the locations of stored spatial data, and the spatial data itself.

Perhaps the least visible, but most important component of GII is simply *organization.* To create a GII requires the efforts of hundreds of individuals working together to standardize spatial data, and to ensure that new data meets GII specifications. This planning and coordination consumes enormous amounts of time initially, but reduces the workload for many people throughout society once it is complete.

1.3 Investment in GII

Over the past few decades, a number of analogies have been used to explain the value of information technologies to the population at large. Since GII is so closely related to, and requires the existence of the Internet, we should have a close look at these models to see how they both help to inform and cloud our discussion of the value of GII. We can then look more clearly at some of the reasons to invest in GII, and some of the costs of this endeavor.

1.3.1 Analogies that Sell

Such analogies as the "Information Highway," the "Information Marketplace," or likening the Internet to the electric power grid have been used to sell the development of infrastructure for the Internet to the public. Although such analogies are helpful in helping people understand the value of complex technological investments, it is important to realize the limitations of these analogies as well.

Take for example, the "Information Highway" analogy. While it is true that Internet backbones can supply vast amounts of information, this analogy fails to explain that there is essentially no distance involved in Internet transmissions. The user never notices the millionth of a second difference between information coming from Latvia as opposed to information coming from Australia. Furthermore, because of the way that the Internet is organized, it may be that information from Latvia actually travels through Australia on its way to Lithuania!

The concepts behind the "Information Highway" are partially related to the development of a GII, but there are other factors specific to a GII that this analogy ignores. One assumption is that the most important factor is the transmission of data from point to point. But sending bits along a wire is the easy part. The analogy ignores the fact that there is a great deal of effort involved in the collection and organization of the data before it can be transmitted. Indeed, now that a great deal of spatial information exists in digital form, the main technological hurdle behind the development of a Geographic Information Infrastructure is the organization of the data. Also, the "Information Highway" analogy implies a two-dimensional arrangement of roads on the Earths surface. With the Internet, information may move horizontally (between agencies and disciplines), or vertically (between different levels of government) (Groot and McLaughlin, 2000).

The "Information Marketplace" analogy looks at the Internet as a place to do business, much like a glorified shopping mall. Although this analogy is also correct to some degree (witness the success of Amazon.com and eBay), it fails to account for the differences in the way that Economics work in the digital world versus in the real world. The two major differences are that there are no shipping costs, since it costs virtually nothing to transport digital information using the Internet, and the fact that once created, digital information can be reproduced at no cost, with no loss in quality.

In this respect, the electric power grid analogy is more accurate in that it reflects extremely rapid transmission of information from source to destination through one of many different possible paths. This analogy also helps us to understand that what is important is the availability, usability, reliability, and low cost of the information provided through the Internet (Groot and McLaughlin, 2000). This analogy, however does not distinguish between the different types of data that comes through the Internet, the different levels of quality of the data received, or the billions of different possible sources of data that exist in today's Internet. Electricity is electricity -- there is only one product provided by the electrical power grid.

1.3.2 Funding GII Development

The development of a Geographic Information Infrastructure is unquestionably an expensive proposition. Without a clear return on the investment, such an endeavor makes little sense. However, as this new technology develops, the costs continue to drop, the efforts of other nations help to clarify what are best practices, and it becomes easier to obtain funding once those groups who stand to gain the most from investment in GII are identified.

Because of the details the intergovernmental nature of GII development, it has been very difficult to obtain a precise cost for the development of GII. Much of the work has been done informally by GIS professionals and other governmental employees as part of their daily duties. In the United States, we can at least obtain an accurate value of the federal expenditure for GII, because of the line-budget approach of the Federal Government. The annual cost to the US federal government for data capture and maintenance is approximately US $4 billion, and an additional US $6 billion is estimated to be spent by state and municipal governments (Rhind in Groot and McLaughlin, p. 43). The total cost for GII is estimated to be US $5-$6 billion per year, according to a narrow definition, and more than US $15 billion per year using a broader definition (Rhind in Groot and McLaughlin, p. 44). These are, of course, highly qualified estimates, and are based on the development of GII for one of the world's largest countries.

McKee (2000) points out that there are two forces at play in the development of a GII. These are the "push" of technology, and the "pull" of markets (Groot and McLaughlin, p. 13). The decreasing costs of computer and information processing technology mean that it is less costly to develop GII as time goes on. The development of LBS will lead to a new class of non-professional spatial data user. These people will be accustomed to obtaining information from their cell phones, but they will not be experts in geospatial technology. They will have little patience for arguments surrounding metadata, data formats, custodianship, or pricing models for spatial data.

Furthermore, increasing public demands for the online delivery of government services mean that there will be a stronger case in favour of the development of GII as time goes on. Both of these trends going towards the development of a "geospatial marketplace," in which appropriate data can be sold to many different users. The more appropriate the data to a particular users needs, the higher the price that the data can fetch.

Of course, if highly valuable data has been created, it is important that this data be protected so that it remains as a source of revenue for the agency or agencies that created it. It is very expensive to create these data, so if these data are lost through theft, then the loss of these data can make it unprofitable to produce them, which nullifies one of the key reasons for building a GII.

There are a number of reasons why a society would want to invest in a Geographic Information Infrastructure. The first of these is simply that there will be a return on the investment, although it will take as much as a decade for the investment to pay off. The payoff will be economic, in the form of data sales and increased competitiveness, political, in the form of increased prestige and ability to work effectively with neighboring countries, environmental, in the form of better decision-making, and social, in the form of increased spatial awareness and the development of an Information Society.

There are some interesting trade-offs involved in the pricing of geospatial data. If the data are too expensive, then only the wealthiest individuals and companies will be able to make use of the data. However, remember that the marginal cost of making additional copies of data is nothing, so high prices encourage the theft of data. On the other hand, if the data is priced too low, and insufficient revenues will return to the agencies to sustain the GII.

In the former case, there is no development of an "Information Society," and in the latter case, there is an initial rapid development and use of spatial data. The former case is what happened in Canada (and, in my opinion, led to the withering away of the Canadian GIS Industry in the early 1990s), and the latter case is what happened in the United States. In the United States, there was a deliberate policy to release federally collected data at no cost. The argument was that the taxpayers had already paid federal agencies to collect the data, and should not have to pay again to use the data.

So an appropriate pricing policy is necessary to maintain sufficient government revenues, while still promoting the general development of society. Given the right policies, a revolution can occur in the way that people make use of spatial data, and the Information Society will come into existence. Lance McKee provides some insight as to how individuals in an Information Society might handle and make use of spatial data:

> Through GIS, everything looks like overlay thematic maps that can be combined to provide new thematic maps. This technology instills a mental habit of analyzing spatially, temporarily, visually, above the fray, non-linearly, aware boundaries but unconstrained by them, and aware of the overlaid human and natural elements of the planet. It is a kind of analysis that synthesizes and integrates rather than divides, and thus may influence our values in a direction which is more in line with the need for a sustainable life on Earth (Groot and McLaughlin, p. 23).

To get to this point will not be easy. Because of constant changes in the legal, organizational, and technological environment, the path to a complete GII will not be a straight line. The Lithuanian Geographic Information Infrastructure, when complete, will be a reflection of the society and institutions that built it. While it may have features in common with the Geographic Information Infrastructures in other nations, it will be unique. It will never be perfect, but as long as it is reasonably functional, then it will produce significant returns to the nation in decades to come.

The ultimate goal is to create a situation where everybody in society who needs spatial information to make a decision has access to it. McKee once again summarizes the point nicely. "GIS extends our mental ability to put together an understanding of our world so that we can shape not only our world but also our behavior, because ultimately, our survival depends on our directed individual and collective behavior" (McKee in Groot and McLaughlin, p. 23).

1.4 Development of GII

Geographic Information Infrastructure is based on both technological capabilities and social needs. The social side of GII development has come about because many individuals in different nations have realized that Geographic Information Systems have limited capacity without high-quality data. These individuals were the pioneers of the GII movement, and worked to strengthen ties between organizations until a threshold was reached in which the importance of GII became evident both domestically and internationally.

This "bottom-up" approach, as happened in North Carolina, USA (Groot and McLaughlin, p. 251) was difficult and slow, with practitioners often making up rules and procedures as they went along. It is important to note that the "bottom-up" developers of GII eventually arrived at a situation where government took a central role in the development and maintenance of GII; the "bottom-up" early development led to later on "top-down" coordination by the central government. For this reason, it makes little sense to consider the "bottom-up" approach at this time, since all players have now adopted more formal approaches.

Once a number of GIIs had been developed, not only did the advantages of GII become apparent, but also it was also easier to learn from the mistakes of others. This has helped to reduce the enormous cost of setting up a GII for those nations that were not the initial adopters. Now that international standards are being developed, it makes more sense to begin with accepted international practice, and see how it can be applied in Lithuania.

1.4.1 Role of Government

Depending on the political philosophy of the government, its budget for GII, and the way he that the GII is structured, government may have a large or a small role in GII. At a minimum, it will want to be involved in establishing appropriate base layers and metadata formats. The central government may also wish to establish one or more central organizations to coordinate GII activities among the different parties involved, as Lithuania has done with the National Land Service.

Although the Lithuanian Government will be both the largest investor in the largest beneficiary in the Lithuanian GII, at least initially, it is important to remember that a GII has society-wide effects. GII affects the following groups, in addition to the central government:

- Local Government
- Information providers in private industry
- Private software vendors
- Industrial conglomerates
- Nonprofit organizations
- Academic sector
- International Agencies, and
- Individual citizens (Rhind in Groot and McLaughlin, p. 41)

In many ways, every individual in Lithuania will benefit from GII, either directly or indirectly, now or in the future.

Of course, it is also possible that GII could be harmful, if, for example, private information were released publicly. So, not only does government stand to benefit from GII, but it must also exercise control, so that the benefits of GII are maximized and the potential harm is minimized. EC Directive 95/46, which relates to the collection, retention, and distribution of personal data by government agencies, has helped to clarify what personal information must be kept confidential.

The development of GII by the central government is important, but it is also important to realize that the central government also has a role in *getting out of the way* of GII development. For example, Copyright laws that were created to support the publishing industry may need to be revisited to consider the special demands that are made by the online publishing of digital data. The European Union has already made a number of legislative changes in this area, in particular EC Directive 96/9/EC, the right of extraction provision for databases, which provides 15 years of protection from the date of creation of a database.

Given that the technological, social, political, and legal environments surrounding GII are gradually changing, it is fair to say that the role of government will also have to change in coming years.

1.5 Examples of GII

Around the world, Geographic Information Infrastructure projects have been developed to meet the needs of national data users. The following examples show that each nation involved has unique characteristics that need to be considered for the successful development of GII.

1.5.1 Australian Census Mapping Project

Australia had no Geographic Information Infrastructure to speak of until 1996, when the Australian Bureau of Statistics requested tenders for the Census Mapping Project, which was to create a national series of maps to support the Australian census. Each State and Territory had been pursuing independent and incompatible mapping programs. The Public Sector Mapping Agencies of Australia (PSMA) was created at that time to coordinate the activities of the different levels of government to create the maps for the Census. Within PSMA, the New South Wales Surveyor General's Department became the lead agency to coordinate the nine separate agencies involved.

A multiresolution data set was created by combining and standardizing topographic and cadastral information that was derived from Commonwealth, State, Territorial, and Municipal mapping programs. The source documents had incompatible coordinate systems, formats, and specifications.

The Census Mapping Program created a de facto authoritative national data set, which has become the base map for a new series of national mapping initiatives. A National Cadastral Database and a National Digital Road Network Mapping Project are now in the works based on the Census Mapping Program framework. More than 15 organizations have now enquired about reselling the Census Mapping Project data.

In this example, we see that a single high priority project was necessary to initiate a development of the Australian Geographic Information Infrastructure, but once the project was initiated, its utility quickly became apparent to all parties involved (Grant in Groot and McLaughlin, p. 255).

1.5.2 Netherlands National Geospatial Data Infrastructure

In the Netherlands, the Clearinghouse for Geospatial Information has served as the nexus for a National Geospatial Data Infrastructure. A clearinghouse is a website or a series of websites that allow digital data to be located, reviewed, and downloaded. This project was recommended in a 1992 RAVI (Netherlands Council for Geographic Information) report, recommending the development of a nationwide metadata service and a national clearinghouse for geospatial data.

The project began as a grassroots initiative around 1995, with no direct financial support, except for one-time funding of €100,000 for coordination, and €500,000 for a pilot project. Under an informal management scheme, a number of initial studies were conducted, and recommendations were made for a clearinghouse and the format of metadata. A prototype metadata service was created on the Internet using the CEN (European Committee for Standardization) TC 287 standard for metadata.

By 1996, the success of the clearinghouse prototype lead to support for the development of a professional clearinghouse system that could support more users and which featured a more professional User Interface. The project became more formalized, and obtained direct financial support of €500,000 funded by the government of the Netherlands and participating organizations. At this time, a maturity model was developed to help assess the capability for different organizations to contribute data to the clearinghouse project.

In 1998, the Clearinghouse for Geospatial Information was subsumed into the NGDI Institute, which had an annual budget of €250,000 per year, guaranteed until the year 2000. In future it appears that the clearinghouse will become self-supporting, based on value added or fee-based services. It is anticipated that this will set the stage for the development of a marketplace for geospatial data in the Netherlands.

In this example, we see that the grassroots initiatives of a number of individuals gained the support of government agencies, which recognized the importance of this GII project. The grassroots approach to this project lead to wide support among many different organizations, and helped to guarantee the success of the project.

1.5.3 Service New Brunswick

The Canadian province of New Brunswick has developed its own Geographic Information Infrastructure to meet the needs of its citizens. This province, which has roughly the same area as Lithuania, but a population of only 700,000 people, has developed a sophisticated, web-based system to support its population.

Until the 1980s, government agencies had constructed separate individual databases for the management of natural resources in New Brunswick. Beginning in the 1980s, these databases were gradually unified into a single province-wide network. In the 1990s, with the advent of the Internet, the public was given a way to easily access these databases.

In 1989 in the Provincial Land Information Policy attempted to standardize the use and collection of land information in the province. This policy led to the creation of Service New Brunswick (SNB), the Crown Corporation that became the lead agency in the creation of the provincial GII. In 1996, Service New Brunswick unveiled a commercial online Land Registry system, which made cadastral information available to everyone in the province. Since the introduction of the system, environmental data have been added, and additional data sets will be made available as they are completed.

Because SNB is a Crown Corporation, it operates as a business, with the profits returned to the Government of New Brunswick. This ensures that the effort to establish a GII in New Brunswick is self-supporting, and it also ensures that the needs of the government and people of New Brunswick are met. Under its charter, SNB was intended to become financially self-sufficient within its first five years of operation.

This example shows that a top-down approach to the creation of a GII is possible under the right conditions. Part of this may be due to the small size of New Brunswick's government. Having high-level support within the provincial government was important factor contributing to the success of SNB. Its creation allowed a lead agency to undertake the standardization of all land records and resource mapping within the province. As a Crown Corporation, SNB has focused on those aspects of GII that were most immediately profitable in order to ensure its survival, and only after the initial geographic framework was established did it look towards expanding the scope of the data that were offered in its Land Registry system.

1.6 Applications of GII

Why invest large sums of money in GII? Because a properly established GII benefits all individuals in a society either directly, or indirectly. The way that it does this is by permitting a wide range of applications that would not be possible otherwise. The following is a list of some applications that a complete GII permits.

- Data Storage and Retrieval
  - Efficiently storing and retrieving information about land parcels and other spatial data

- Location and Allocation Analysis
  - Determining the best location for a new facility or business
  - Determining which areas are best served by existing facilities or businesses
  - Mathematically modeling the appropriateness of all locations on the map using many input data sets

- Network Analysis
  - Determining vehicle travel time based on a network of roads
  - Determining optimal paths based on networks (vehicles on roads, water from a municipal water supply, sewage in sewers, electricity in a power grid)
  - Determination of approximate addresses along networks
  - Storage and retrieval of information relating to maintenance of networks

- Modeling
  - Optimizing spatial model results, based on varying one or more parameters
  - Optimizing networks for most efficient travel of vehicles, water, sewage, electricity...

- Weighted Distance Functions
  - Determination of travel costs on a surface (cross-country vehicle travel, forest fire growth, wildlife migration)
  - Determination of a least cost route across a surface between two points

- Remote Sensing
  - Use of Remote Sensing data to update existing information or collect new information

- Digital Terrain Models
  - Analysis of terrain
  - Delineation of watersheds
  - Classification of streams
  - Prediction of flow and floods

- Analysis of groundwater flow

With the standardized data at the regional and national level, applications such as these can be applied to multiple scales, with little reconfiguration. Without a GII, the amount of customization and reconfiguration that is required makes it difficult to implement these applications as any more than prototypes.

The real strength of the GII comes from the ability to combine applications. Suppose a police call is received -- an enhanced 911 system provides the name of the person and their address, this address is then automatically located on the map (determination of addresses), the nearest police station is identified (allocation problem), and the fastest route is determined from the police station to the location of emergency (fastest route determination). In an emergency situation such as this, when seconds count, a GII has the potential to save lives and property. Individually, the technologies are important, but not world changing. Together, they can change the entire way that Lithuanian society works.

In this course, we will examine how the above applications can be made routine through the implementation of a Geographic Information Infrastructure. This course will be divided into four modules:

1. Optimization of Locations
2. Routing and Determination of Travel Time
3. Land Use Analysis
4. Terrain and Hydrological Analysis

By the end of this course, you will have a good understanding of why Geographic Information Infrastructure is being built in Lithuania, and the benefits that it can bring to society once implemented.

Module Self-Study Questions

1. The concept of a Geographic Information Infrastructure is difficult to grasp because it is new way of organizing information that is quite abstract in nature. Can you think of any other information frameworks that you make use of in daily life?

   Answer: spoken and written languages, culture, mathematics, symbols that you understand, music

2. In Figure 2, we saw that Google Maps already provides limited GII for Lithuania. Can you think of any reasons why we should not simply wait and rely on private corporations to create a complete GII for the country?

   Answer:
   a) because Google is probably already making use of data from Lithuania in their map products
   b) because it is unlikely that Google will ever provide the level of detail required for government use
   c) because there may be privileged/sensitive information that the government does not want a private company being the custodian of.

Required Readings

Groot, Richard & McLaughlin, John (2000). Geospatial Data Infrastructure: Concepts, Cases, and Good Practise. Oxford: Oxford University Press. pp. 13-24, 245-251.

ESRI Virtual Campus Module

- Introduction to Urban and Regional Planning Using ArcGIS 9 Module 1: Introduction to Urban and Regional Planning Concepts

References

Groot, Richard & McLaughlin, John (2000). *Geospatial Data Infrastructure: Concepts, Cases, and Good Practice*. Oxford: Oxford University Press.

How Stuff Works (2007). *How the Future Force Warrior Will Work: Battlefield Awareness*, http://science.howstuffworks.com/ffw2.htm (June 25, 2007)

University of California, Berkeley (2007). *Earth Sciences and Map Library Topographic Maps*, http://www.lib.berkeley.edu/EART/tour/topo.html (June 25, 2007)

Terms Used

- Allocation Analysis
- Application-Specific Data
- "Bottom-Up" Approach to GII
- Commercial Off-the-Shelf (COTS)
- Digital Terrain Models (DTM)
- Framework Data
- Foundation Data
- Geocoding
- Geographic Information Infrastructure (GII)
- Geospatial Clearinghouse (Clearinghouse)
- Geospatial Data Infrastructure (GDI, GSDI)
- Global Data Infrastructure
- Global Positioning System (GPS)
- Human-Based Services
- Information Highway
- Information Marketplace
- Information Society
- Information-Based Services
- Infrastructure
- Location Analysis
- Location-Based Services (LBS)
- Network Analysis
- Server Computer
- "Top-Down" Approach to GII
- Weighted Distance Functions

2 Optimization of Locations

Outline:

7. Introduction
8. Foundation Technologies
9. Key Technologies
10. Optimization
11. Areas of Application
12. Examples

2.1 Introduction

Virtually all data has a spatial component. Consider that "Everything happens somewhere, and everything and everyone is somewhere" (Groot and McLaughlin, p. 14). Until very recently, however, it has been difficult to collect spatial information about where things and people are, and where events occur.

Although spatial information is very important, the tools to collect, analyze, and distribute that information have only recently become available. Only large governmental organizations and wealthy individuals have had the ability to collect and use spatial data. Furthermore, the tools to analyze these data were not available.

Geographic Information Systems are a new technology, and have only recently become available to the public. The development of Geographic Information Infrastructures has meant that quality spatial data is becoming increasingly available at a relatively low cost, and these data, plus information on how to use them, are now available over the Internet. With the advent of new geospatial technologies and inexpensive Global Positioning System receivers, spatial data is becoming increasingly common, and with that there is an increasing awareness of the importance of space in people's daily lives.

Our increasing ability to measure and analyze space means that it is now possible to optimize the location of new facilities so as to maximize their utilization. In virtually all sectors of society, the ability to choose the best place has the promise to gradually reorganize the spatial basis of society, so that activities are done in a more efficient, cost effective manner.

In this chapter, we will discuss issues surrounding the allocation of people or objects to facilities and the optimal location of new facilities. We will begin with a discussion of the Multipurpose Cadastre, which is a specialized type of GIS that contains information on land parcels, the construction of roads networks in a GIS, and we will review the tools available in Map Algebra. These technologies are required in order to solve at location and location problems. After we have discussed how to locate new facilities manually, we will discuss techniques to optimize the location of these facilities.

2.2 Foundation Technologies

The tools that are used to analyze and calculate optimal locations do not operate by themselves. The key technologies used for location and allocation are based on a number of foundation technologies; without which, location and allocation analyses could not be performed. The foundation technologies are used to preprocess data and convert it into a form that is amenable to location and allocation analysis.

The foundation technologies are critical because they make location and allocation operations practical. For example, data in a Multipurpose Cadastre can be used to determine where people live (approximately) for determining the zone of allocation around a particular facility. Map Algebra is a set of tools, without which, it would be impossible to perform Index Modelling to create a shortlist of potential sites during location analysis.

2.2.1 Multipurpose Cadastre

A Multipurpose Cadastre is a type of specialized GIS that is used for the management of land parcels (Figure 5). The Multipurpose Cadastre has two major functions, first to manage the legal records of land tenure (judicial cadastre), and second to manage records of property valuation (fiscal cadastre). Although the Multipurpose Cadastre provides information on the number of buildings and the occupancy allowed, many other types of information are stored, including:

- Household Income
- Land Use
- Building Type
- Mortgage
- Land Value (for Tax Assessment)
- Land Ownership
- Parcel Boundaries (adjacent properties can be determined through topological relationships)
- Land Rights and Restrictions
- Administrative Districts and Boundaries
- Population and Census Data
- Cultural and Archaeological Information
- Zoning

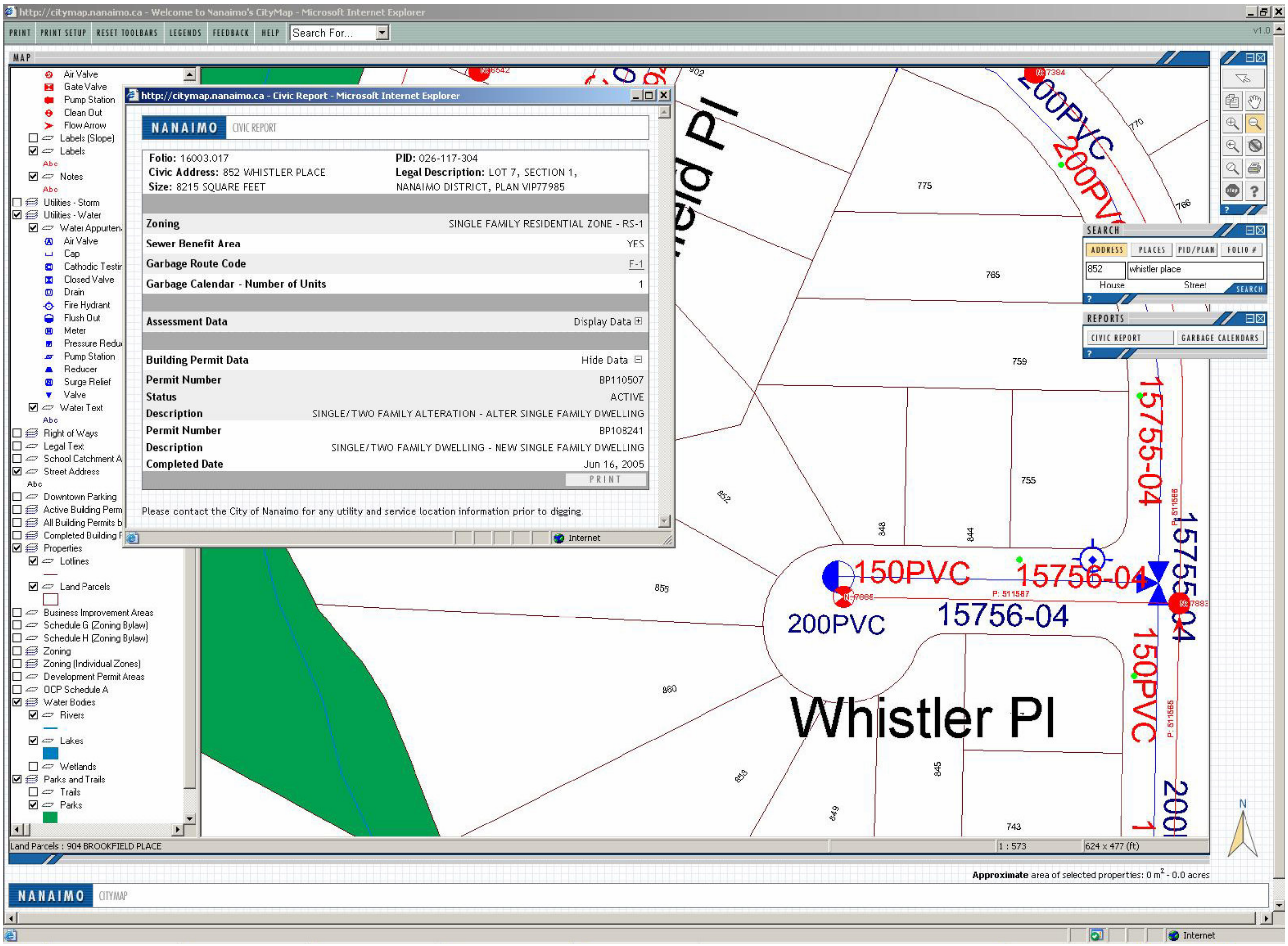


Figure 5. Example of an online Multipurpose Cadastre showing water and sewer line, and a text report. Property information from the City of Nanaimo (http://citymap.nanaimo.ca/)

In Lithuania, the State Enterprise Centre of Registers (SECR) maintains a Multipurpose Cadastre containing 33 layers, including information such as administrative boundaries, centre points of buildings, addresses, and real property value zones. Portions of the Multipurpose Cadastre are made available over the Internet to all Lithuanian citizens.

The Multipurpose Cadastre is one of the fundamental sources of the data that is used in solving location and allocation problems. Having a complete record of all land parcels enables us to determine which parcels of land are available for the development of new facilities. Furthermore, information in the Multipurpose Cadastre, such as the location of sewers, water lines, and other utilities can be very valuable in helping us determine the suitability of different parcels of land for a particular type of business or facility.

2.2.2 Map Algebra

Map Algebra is a technology that allows us to combine multiple raster layers using algebraic notation. The fundamental assumption behind Map Algebra is that raster layers are overlapping, and so we can perform overlay operations between different layers. In ArcGIS, Map Algebra is entered through the Raster Calculator tool.

Map Algebra is an ideal way to combine and to weight multiple raster layers, which is how we create Index Models (see Section 0). Suppose, for example, you wanted to determine all built-up land that was susceptible to landslides. To do this, you would want to combine a raster layer that represented the urbanized area with a raster layer representing slope and another raster layer representing annual precipitation. The areas susceptible to landslides would have a slope greater than 50%, an annual precipitation of more than 2000 mm, and would be within the urban boundary. Once our rasters had entered into the GIS, we can use the Raster Calculator to enter the expression shown in Figure 6 to determine the areas of concern.

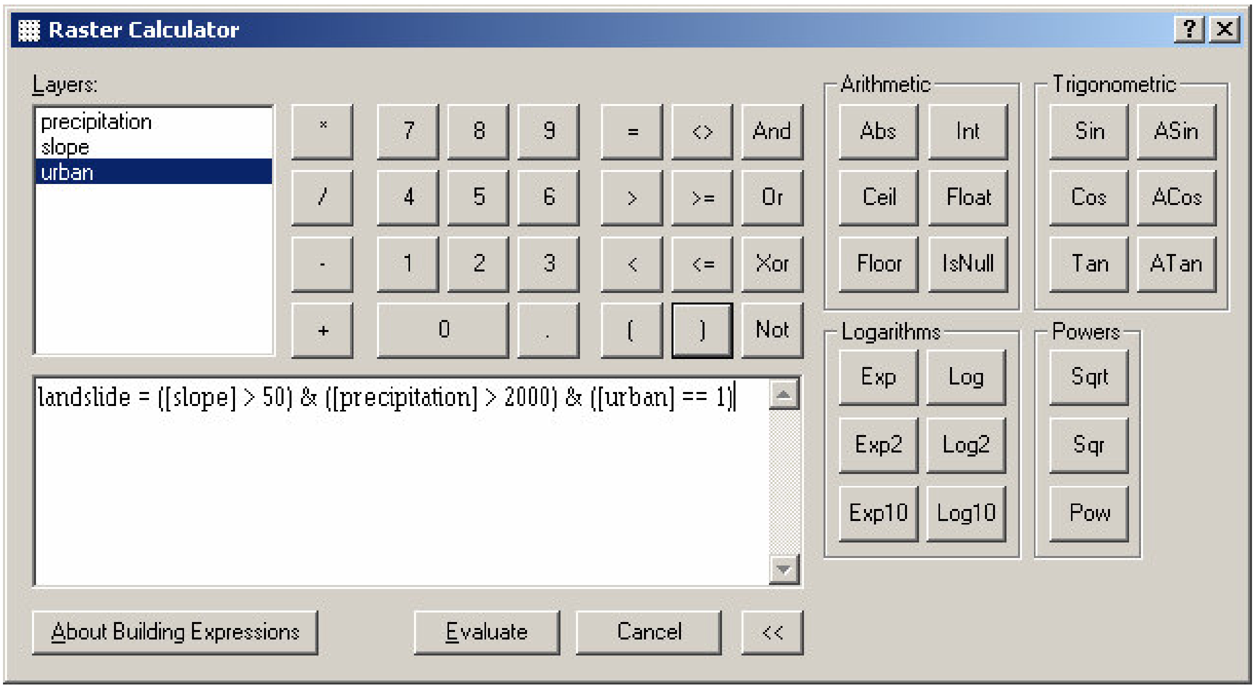


Figure 6. The Raster Calculator, which is used to evaluate Map Algebra expressions.

The concepts behind Map Algebra may seem quite familiar if you have learned Matrix Algebra. In many cases, each input grid can be thought of as matrix, and matrix operations such as addition and subtraction work exactly the same way on matrices as they do in Map Algebra on rasters. However, multiplication and division work differently in Map Algebra, in that each cell is multiplied or divided by the overlapping cell; there is no transposition during multiplication and division as there is in Matrix Algebra.

Another important concept is the difference between Zero and NoData. In Map Algebra, zero is a value, whereas NoData literally represents "no data" -- the absence of information. Consider a Digital Elevation Model of the entire Earth. Some locations have an elevation above sea level (positive), some locations have an elevation below sea level (negative), and some locations have an elevation that is exactly sea level (0). Here, an elevation of 0 is perfectly valid, and provides us information about the elevation. NoData represents an absence of information. Any operation that calculates on a cell with a value of NoData returns a value of NoData. Therefore 1 + NoData = NoData, 1 / NoData = NoData, NoData / NoData = NoData and so forth.

Map Algebra offers a rich set of functions and mathematical operators (Table 1), the most common of which can be seen in Figure 6.

There are also many different functions that are available with Map Algebra, including Local, Focal, Block, Zonal, and Global, functions. Also, Map Algebra has a complete set of operators for managing precedence and flow control within a Map Algebra equation. Taken together, the operators and functions in Map Algebra for a complete modelling language that can be used to develop sophisticated models. For a more thorough discussion of Map Algebra and its many functions, please consult GII-01, Elements of Geographic Information, Module 4, Methods of Data Analysis in GIS.

Table 1. Map Algebra Operators

Operator Type	**Function**	**Examples**
Arithmetic	Cell-by-cell operations between two rasters	+, -, /, *
Relational	Cell-by-cell comparison between two rasters	==, <>, >, >=, <, <=
Bitwise	Performs calculations on the binary representation of the grid values	&& (Bitwise And), || (Bitwise Or), !! (Bitwise XOr), << (Bitwise left shift), >> (Bitwise right shift), ^^ (Bitwise complement)
Combinatorial	Used to preserve all input values when combining rasters	Cand (Combinatorial And), Cor (Combinatorial Or), CXOr (Combinatorial XOr)
Logical	Boolean operations between rasters	& (And), | (Or), ! (Xor), ^ (Not), Diff, In, Over, IsNull, Test, InList
Accumulative	Used to sum all cell values	+=
Assignment	Used to replace values with new ones	=

2.3 Key Technologies

GIS applications that are used to solve location or allocation problems may be based on a vector or a raster data structure. As you may recall, the vector data structure is used to represent discrete data, in the form of points, lines, or areas. The raster data structure is used to represent continuous data, using a regular grid of cells. For location and allocation problems, each of these data structures as advantages and disadvantages, and your choice of data structure will depend upon the nature of the location or allocation problems that you wish to solve.

Location and allocation problems are essentially problems of economics, in that we are dealing with problems of supply, demand, and the cost of providing services. Determining the supply of services is relatively easy, since they are tend to be only a few sources of supply, for example, police stations, fire stations, ambulance stations, repair depots, or businesses. Determining the demand is much more involved, since there may be thousands of people who use or might potentially use a service. Connecting the supply and demand can be done by creating a network of transportation routes.

If we do not have the addresses of service centres, a number of different techniques are available for identifying their locations. If a list of addresses is available, these addresses can simply be address geocoded to produce a series of points to represent sources. Address Geocoding matches addresses in a table with road names and address ranges in a road layer to determine the approximate position of the address. This will be discussed in more detail in Module 3. Alternately, the location of services can be easily collected by using a GPS receiver and by visiting each service centre and determining its location.

Determining the demand for a particular service requires detailed information about the location of potential users of that service. Generally, the best such information can be obtained through the use of a Multipurpose Cadastre, which records information about land parcels. If a Multipurpose Cadastre is not available, it may be possible to use Census Data or other aggregate data sources, however the results of location and allocation analyses will obviously be less detailed than when using more precise data to represent demand.

Connecting the source and demand for services can be done using straight-line (Euclidean) distances, or by using a network. Using a network allows us to determine the approximate travel time and distance along the shortest path along the network from start to finish, and produces a more accurate results than by simply assuming straight-line distances.

2.3.1 Modelling

Spatial Models allow us to use the data in the GII to better understand existing conditions and to make verifiable predictions based on our data. When people think of models, they tend to think of their predictive capacity of models, which is found in a class of models known as Prescriptive Models. While this is a common use of models, it is important not to forget that models can also be a representation of existing conditions, which is known as a Descriptive Model. Descriptive Models help us to understand trends and patterns in existing data (Chang, 2006).

An example of a predictive model might use a Digital Elevation Model of the Lithuanian coast, land use and population layers to predict the effects of the sea level rise as expected to result from Global Warming. With this model, we can predict the number of people that will be displaced, and the economic impact of sea level rise on the Lithuanian Economy, based on particular amounts that sea level increases by. This allows us to create a series of predictions based on low, medium, and high estimates of sea level rise in the next 50 years. Furthermore, we can model the impact of sea level rise in 20-year increments over the next century.

If we were trying to estimate agricultural productivity, we might be able to develop a descriptive model based on layers showing soil type, average annual precipitation, and climate. Once this model has been verified and has been shown to produce a valid result in places where the agricultural productivity is known, it can then be used to create a model of the entire country, giving predictions of agricultural productivity in locations where it has not actually been measured.

Most models of the "real world" have some degree of uncertainty in them. For example a model of deer habitat might be dependent on the abundance of vegetation, which is dependent on rainfall. These models, in which uncertainty plays a part, are referred to as stochastic models. In the human world, and in the world of physics, deterministic models are also found. In a deterministic model, a given outcome always occurs when the same inputs are provided. There is no uncertainty involved. A model of incoming solar radiation to the Earth's atmosphere is an example of a deterministic model, because the amount of incoming solar radiation at a particular location can be determined exactly, depending on the position of the sun on a particular day and time. Of course, the amount of solar radiation reaching the Earth's *surface* must be modeled stochastically, since the amount of cloud cover is stochastic (Chang, 2006).

Models may also vary in time. A dynamic model is responsive to changes in input conditions. An example of a dynamic model would be a model of traffic flows, which is dependent on a number of monitoring sensors installed in freeways. Typically, however, GIS models are static unless they have been custom programmed to accept inputs directly from outside data sources without human intervention (Chang, 2006).

Another way to look at modeling is how a model determines results. If a model is *data-dependent*, it is known as an inductive model -- the conclusions stem from the input data. On the other hand, if a model is dependent on a theoretical underpinning, it is known as a deductive model. Most GIS models tend to be inductive: we might look at areas that are suitable for agriculture based on the conditions (soil, solar radiation, precipitation) that are found where agriculture is most productive. If you flip this model around, you could develop a conceptual model of the most effective conditions for agriculture, and then find locations where these conditions are met (Chang, 2006).

Model Design

Many textbooks teach that the beginning of any modeling process should be a flow chart. Unfortunately, this perspective ignores the creative process that occurs when designing new models. Only for straightforward models does this approach hold merit; however, this is not to say that flowcharting has no value. When designing a complex model, it may be necessary to refine flowchart several times, or to start a new flowchart from scratch. The real value of the flowchart is as a *documentation* tool, which helps the modeler keep track of his or her thoughts and explain them to other people. In most cases, a final flowchart cannot be created until the model is nearly complete. This flowchart then forms the basis for documentation of the model, and makes it easier to explain the operation of the model to other people. For those people who are being taught the operation of the model, it may appear that the flowchart preceded the model, when in fact, the flowchart only comes first in the teaching of the model.

Unlike the top-down, flowchart-based methodology that is used to teach the operation of existing models to new users, creation of a model is a much more bottom-up, inductive process. Often, the process begins in a very broad task-based method, where useful commands and existing models that can help in the accomplishment of the modelling task are identified and catalogued (Chang, 2006).

At this point, it is important to determine whether the model is best based on a vector data structure, a raster data structure, or a hybrid of the two. Clearly, some types of analysis work best using a particular data model, but to obtain all necessary functionality, it may be necessary to employ both raster and vector models. Not all GIS support all vector and raster functions, so if a particular type of functionality is not available, it may be necessary to consider writing a model for another type of GIS. There may be cases where the strengths of two different GIS packages must be combined, or where a GIS package must be combined with other types of software, such as statistical software, image processing software, or custom programs written in languages such as Visual Basic or C++.

When he model is large and complex, a "divide-and-conquer" approach is often appropriate. The model is broken down into sub-models, which then can be individually built within the overall framework. Here, flowcharts are essential for keeping track of tasks in the position of subtasks within the overall model.

Once a model is sufficiently developed to produce the results (not necessarily correct results yet), it needs to go through a process of calibration, in which model parameters and assumptions are adjusted until an acceptable result is produced. This can be a long, iterative process.

Once the model is producing what appear to be reasonable results, the model must then be validated. Validation is essentially a scientific test in which the model is used with data that it has never encountered, to see whether the results are reasonable and represent reality. Only when a model can be shown to produce reasonable results on new data that are not used for model development, can it be said to be effective. It is much easier to create a model that works only for a particular set of data, than it is to create a model that is generally applicable (Chang, 2006).

Classes of Models

There are four different kinds of spatial models with which GIS can help. These are Binary, Index, Process, and Regression models.

Binary Models

Binary models are based on the overlay of layers that have only two possible values. For example, we might want to overlay a layer showing riparian buffers with a layer showing logging activity to see whether any logging has encroached on riparian buffers. For each layer, there are two possible values: riparian and non-riparian areas in the first layer, and logging activity and no logging activity in the second layer. The resulting layer contains either logging in a riparian area (infraction) or no logging in a riparian area (no infraction).

By performing a series of overlays, we can build up the complexity of the analysis, since each overlay adds a new level of complexity to the model. In addition, there are different types of binary overlay operations, such as intersection, union, erase, identity, symmetrical difference, and update.

Each of these operations is based on a logical (Boolean) operator operating between points on each layer. For example, if we are using the Union operator, if the condition is present at a particular point on one layer, and it is present at the same point on another layer, then the Union operator returns a value of True (Figure 7). Because of the use of Boolean logic, Binary models are essentially the extension of database queries into space (Chang, 2006).

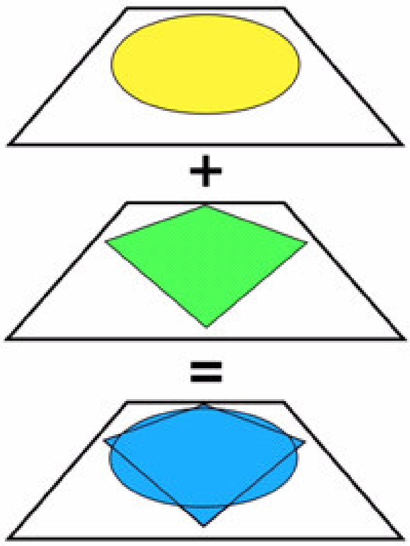


Figure 7. The Union Operator

Of course, the assumption behind Binary models is that each layer represents only the presence of one condition or its absence. For this reason, Binary Models are ideally suited to discrete data types, and Binary operations are supported for both vector and raster data models.

In many cases, however, we are dealing with continuous data, which allows for a range of possible values. In cases such as this, it is necessary to create an Index Model to allow for the processing of continuous data.

Index Models

In Index Models, the overlay operator is extended to support arithmetic operations as well as binary operations. This allows us to use continuous data as well as binary data in the modeling process. This gives rise to the Map Algebra modeling environment, in which we can use arithmetic (Layer 1 + Layer 2) as well as Boolean (Layer 1 OR Layer 2) operators. Index Models are some of the most powerful available in GIS, since they allow many layers to be processed in quick succession using a formula-like approach. For example, we can apply a 1.3 metre vertical datum correction, and then identify all areas below 25 metres elevation using the following formula:

result = (DEM + 1.3) < 25

In this case, we have a single Digital Elevation Model (DEM) as an input layer, which is immediately raised by 1.3 metres elevation, and then all areas less than 25 m are identified. The result is a binary raster layer called result showing the affected areas in the corrected DEM.

It is easy to see how multiple rasters can be easily combined into a single mathematical expression that produces a result. The importance of many different factors can be considered in a single mathematical expression, provided that each input layer is rated in a similar fashion. For example, if we wanted to model forest fire risk, we might consider the amount of precipitation in the past month, the combustibility of vegetation species that exist in the forest, and the amount of unburnt material on the forest floor. One difficulty in creating this model is that there is no way of comparing our three layers of data directly, because they are in different units on different scales. Index Models give us the ability to normalize data so that layers that are organized differently can be combined in some fashion. In this example, we would make all of our data ratio, and map low, medium, or high combustibility values to 1,2, and 3 so that we could combine precipitation, vegetation, and unburnt material together to create a meaningful fire risk prediction. Next, we would choose an appropriate multiplier, so that each risk factor is scaled to the range (0.0, 1.0). Finally, the normalized variables are multiplied by a factor that represents the importance of each variable. In Table 2, below, the most important factor in fire risk is the amount of precipitation, so this is assigned a Variable Importance Multiplier of 3, and the other two variables are assigned a multiplier of 2 to indicate their lesser importance.

Table 2. Normalization and Assignment of Variable Importance to produce an Index Modeling equation for forest fire risk.

Layer	Range	Multiplier to Normalize to (0.0,1.0) range	Variable Importance Multiplier
PRECIP (Precipitation)	0-1200 mm	1/1200 = 0.00083	3
VEGBURN (Vegetation Combustibility)	L,M,H --> 1,2,3	0.333	2
FLOORAMT (Amount of Material on Forest Floor)	0-10 cm	0.1	2

Risk = (PRECIP*0.00083) * 3 + (VEGBURN * 0.333) * 2 + (FLOORAMT * 0.1) * 2

Of course, it is difficult to normalize and weight each layer in a scientifically defensible manner without backup studies, which may not be available. Again, the issue of model validation is very important to ensure that the weighting and normalization of the input variables in an Index Model produces reasonable results that reflect reality. It's very easy to produce a model that *looks* right, but which is a poor model of reality.

At an Index model forest fire risk can be very valuable, but it only shows us in the risk of fire starting at any particular point. Of course, fire is dynamic, and in addition to Index Models showing the susceptibility of particular areas to fire, we also wish to model the predicted spread of fire under certain conditions. Modeling the movement of fire is an example of a Process Model, in which we model a dynamic change through time.

Process Models

A Process Model is a model of a dynamically changing situation. Process Models contain both a predictive element, in that they can predict the outcome of a dynamic process, and an explanatory component, in that the model must be a reasonable simulation of reality. For the model to even remotely approximate a natural process, it is necessary that it behave in accordance with many natural phenomena. The prediction of a wildfire, for example, must behave the same way a real fire does when exposed to changes in slope, when encountering water bodies of different sizes, when encountering different types of fuel of different dryness levels, and when wind blows the fire in a particular direction. In order to be able to model the processed effectively, you need to have a good understanding of how the various parameters that affect the model work. When the model does not perform as expected, it is possible to learn something about the process as you adjust the parameters to ensure that the model behaves as expected.

In addition to modeling wildfires, Process Models have been used to model soil erosion using the Revised Universal Soil Loss Equation (RUSLE), agricultural runoff using the Agricultural Nonpoint

Source model (AGNPS) (Young et al., 1987), and soil and water quality with the Soil and Water Assessment Tool (SWAT) (Srinivasan and Arnold, 1994) (Chang, 2006).

Regression Models

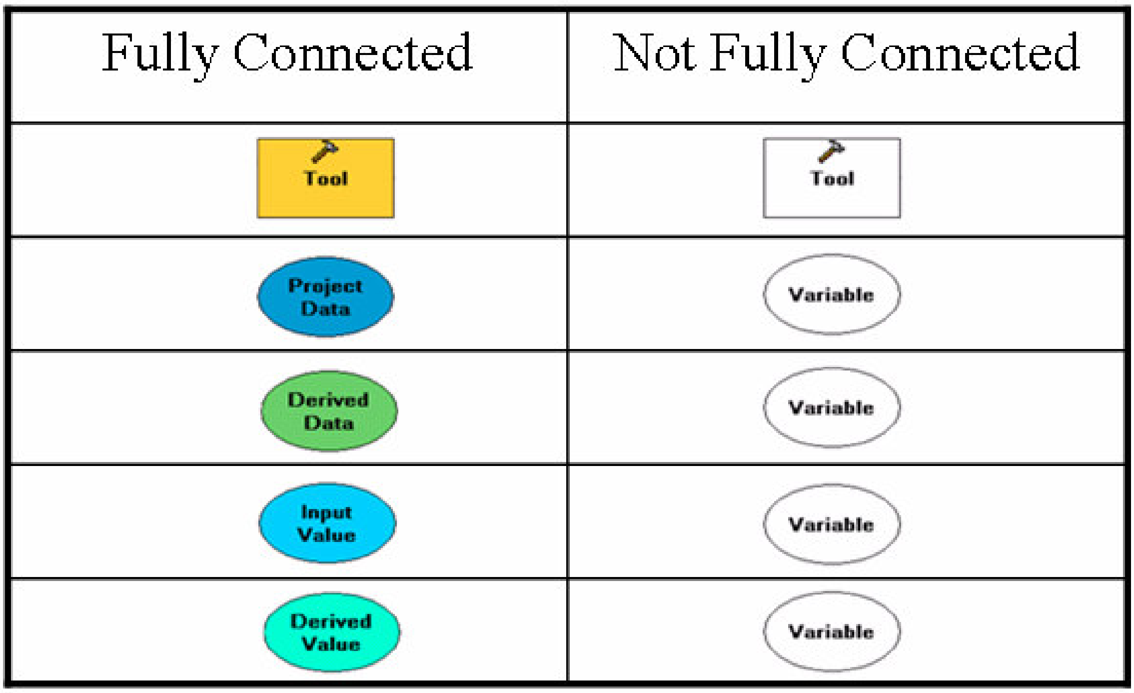
In Statistics, a Regression Model aims to predict the level of a dependent variable based on the level of a related independent variable. For example, there might be a relationship between soil depth and tree height in a forest. The deeper the soil, the higher the trees are in an area. A Regression Model applies in equation to determine a dependent variable at every point on a map where an independent variable has been mapped. This would allow us, for example, to create a predictive map of tree heights, based on a map of soil depths.

2.3.2 Using Model Builder

ArcGIS 9.2 has an interactive model design environment called Model Builder, which allows users to create models based on a flow chart metaphor. This environment works well because it uses the flowchart to explain the operation of the model, while maintaining the ability for the user to build and adjust the model interactively.

Model Builder can be used to create a raster, vector, or mixed raster and vector models. All of the operations in ArcToolbox, including Map Algebra operations, can be incorporated into models. I

Model Builder is "self-documenting," in that the user changes the operation of the model by altering the appearance of the flowchart. Each element in the flowchart, including Tools, Project Data, Derived data, Variables, Input Values, and Derived Values



(

Figure 8) can be configured and modified interactively, by right-clicking on the object and opening it (Figure 9). Elements that have been successfully connected to other elements are drawn with a background colour, whereas those that are missing connections are drawn in white.

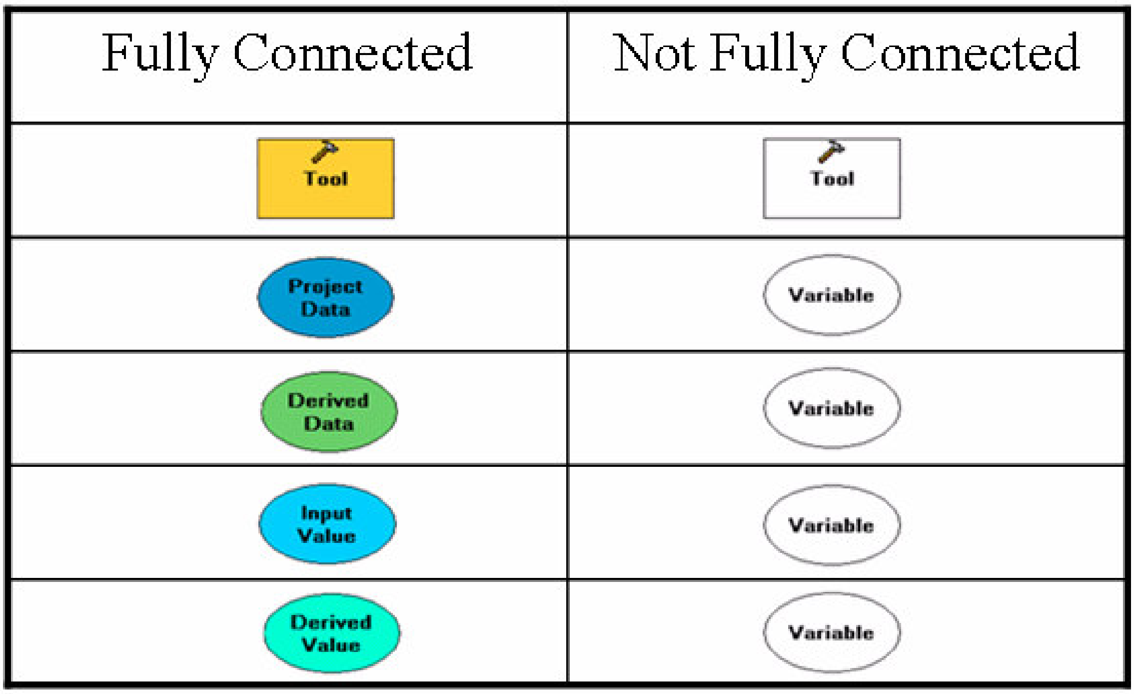


Figure 8. Appearance of Model Elements.

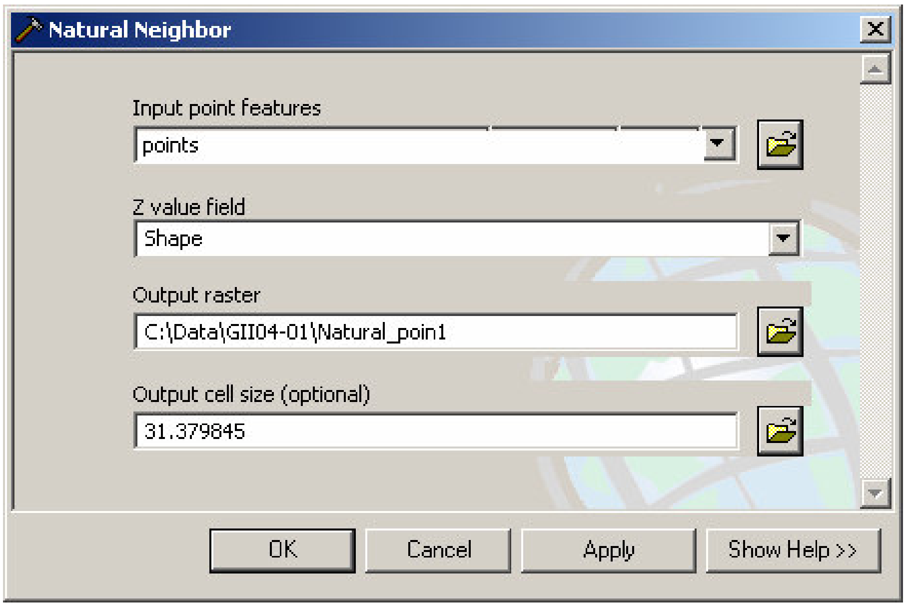


Figure 9. Connections for the Natural Neighbour Tool are displayed when the tool is opened.

Creating a model is a simple as creating a new model, dragging a number of layers into ArcMap to create some elements, and connecting the elements together. In practice, however, more complex models require the configuration of multiple connections for complex commands, the addition of variables, and the setting of model parameters, which requires a higher level of expertise.

We will discuss the construction of models by showing the example of how to construct a simple model to create buffers around a series of industrial sites.

To create a model, you first need to create a Toolbox in which to place it. Toolboxes are folders that contain tools within ArcToolbox. In ArcToolbox, simply right-click on an empty space and select "New Toolbox" (Figure 10), and then rename the Toolbox to something appropriate. Note: user-created models cannot be placed into the default Toolboxes in ArcMap.

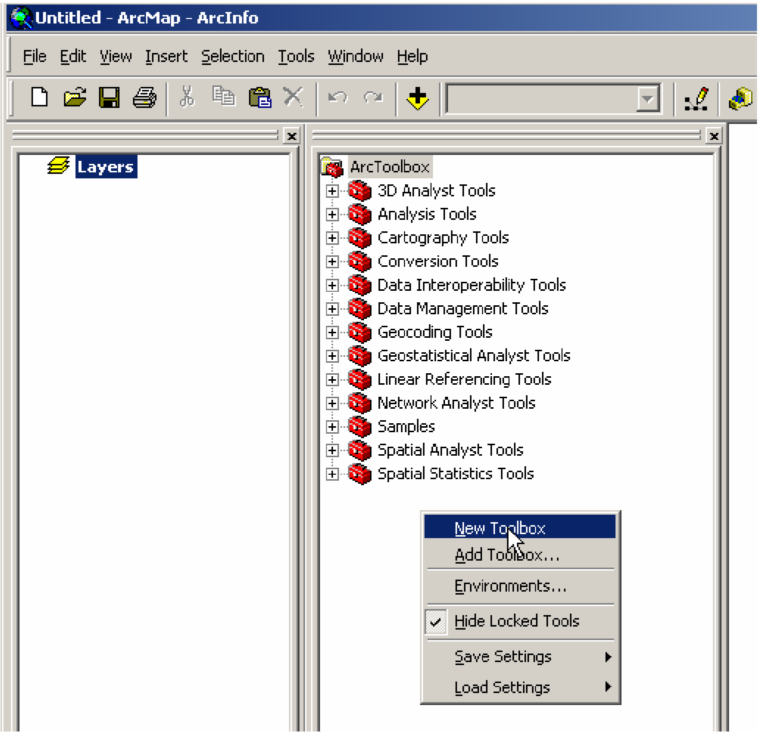


Figure 10. Creating a new toolbox within ArcToolbox in ArcMap

Creating a new model is similar in approach. Right-click on your new Toolbox, and select "New" and "Model" to create a new model within it (Figure 11). Once the model has been created, edit it by right-clicking on the model name and select "Edit" (Figure 12).

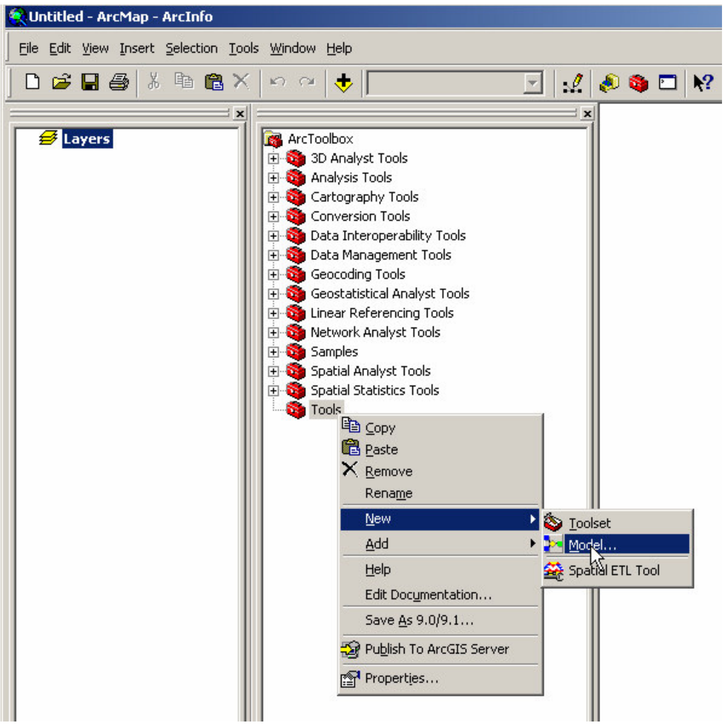


Figure 11. Creating a new model within the new toolbox, which has been renamed "Tools"

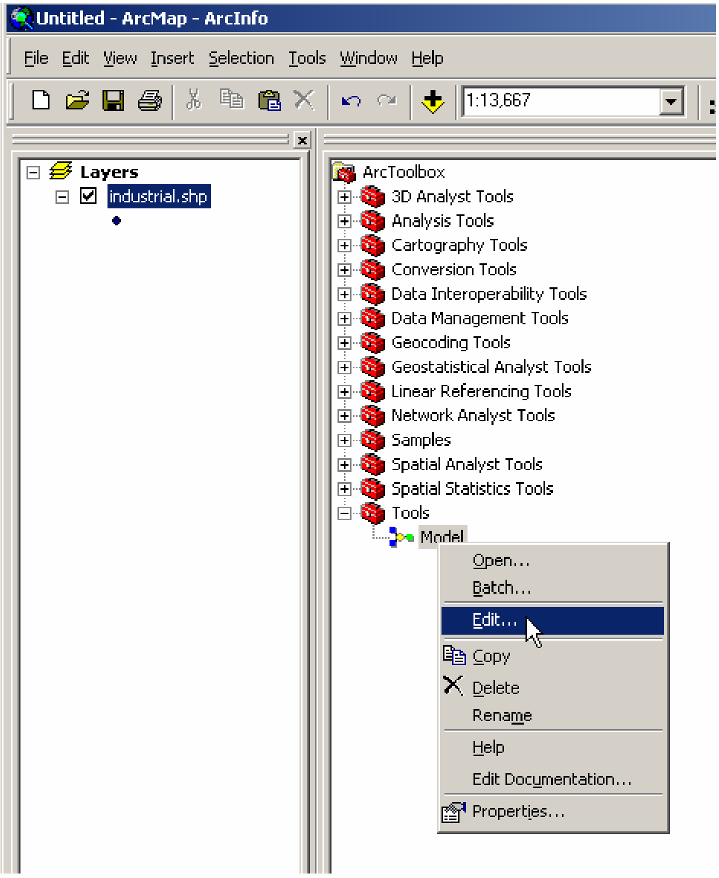


Figure 12. Editing a new model.

Layers can be dragged from the Table of Contents, and tools can be dragged from ArcToolbox onto the Model Builder window as well (Figure 13). Adding a variable requires you to right-click on the background window for the model and select "Create Variable…." You may then add one of many different types of variables (Figure 14).

Once all of the elements are in place, use the Add Connection ( ) button to draw connections between the elements of the model.

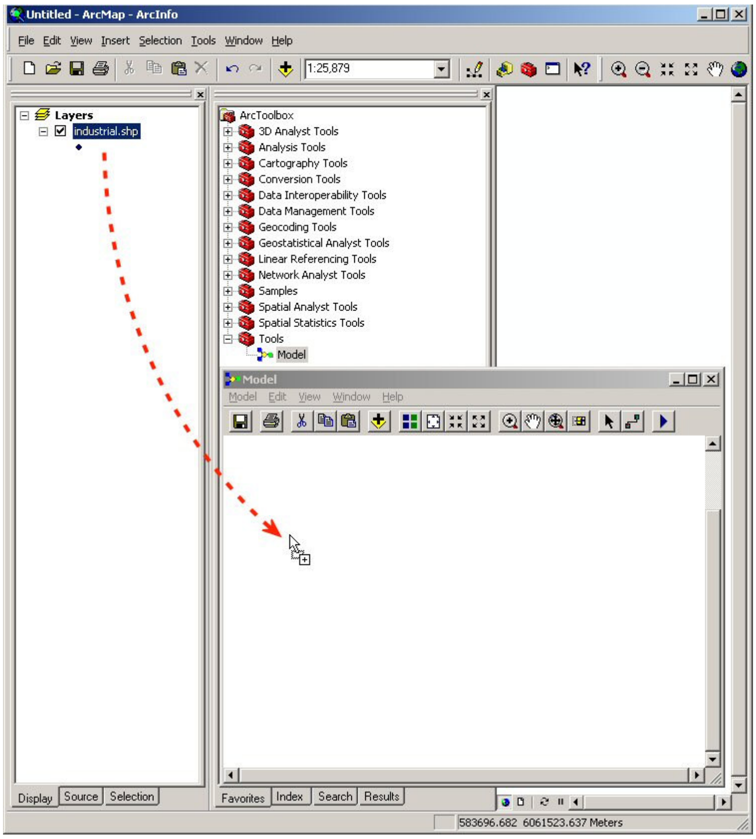


Figure 13. Left-click on the "industrial.shp" layer and drag it onto the Model Builder window to add it to your model.

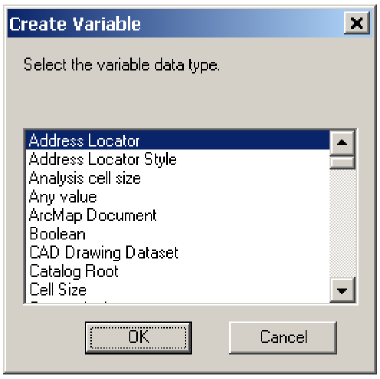


Figure 14. Adding a Variable to Model builder

Once the elements have been dragged onto the model and connected, they form an incomplete model. For the model to work, the elements that are drawn in white need to be modified (Figure 15).

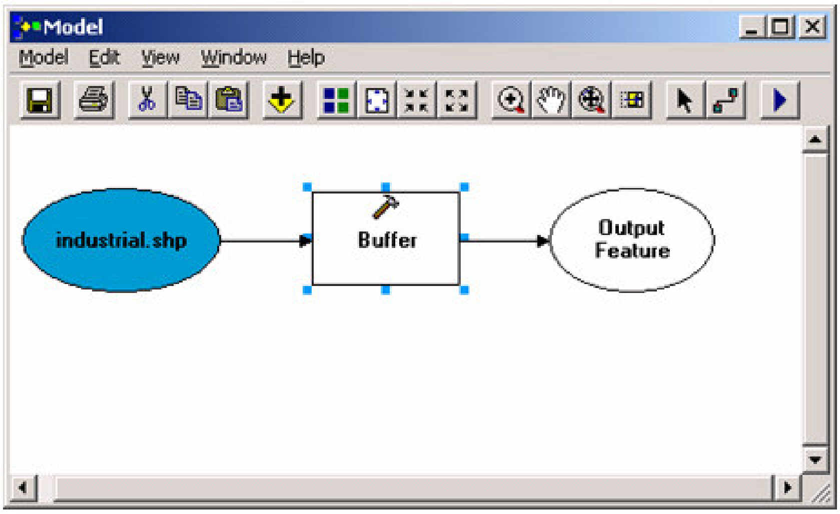


Figure 15. Initial model appearance (not all parameters have been set)

Once we have filled in the missing parameters (in this case the buffer distance parameter and the output file name), all of the elements become coloured to indicate that the model is ready to run. Click on the run ( ) button in the Model Builder window to see the results (Figure 16).

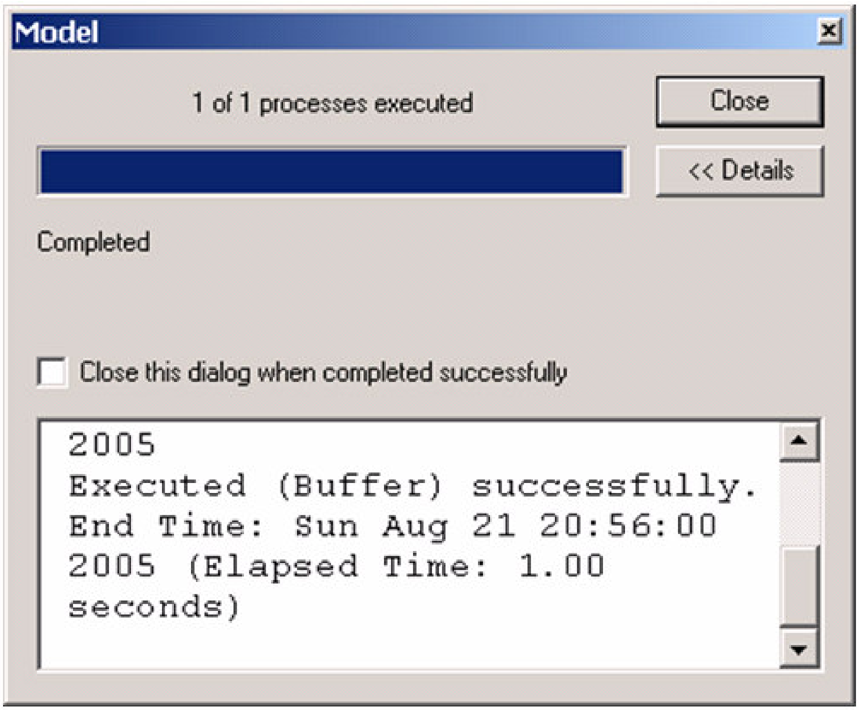


Figure 16. Running the Model for the first time

Unfortunately, at this point, the model has a fixed output file name and buffer size, so it is not particularly useful. To make it more useful, we need to extract the buffer distance to a separate variable, which can then be turned into a parameter, together with the output filename. Creating parameters causes them to be automatically added to a User Interface, which is displayed when a user double-clicks on the model name in the Toolbox. To turn the upper distance into a variable in the model, right-click on the Buffer element and choose Make Variable/From Parameter/Distance [value or field]. This creates a separate variable object for the buffer width (Figure 17).

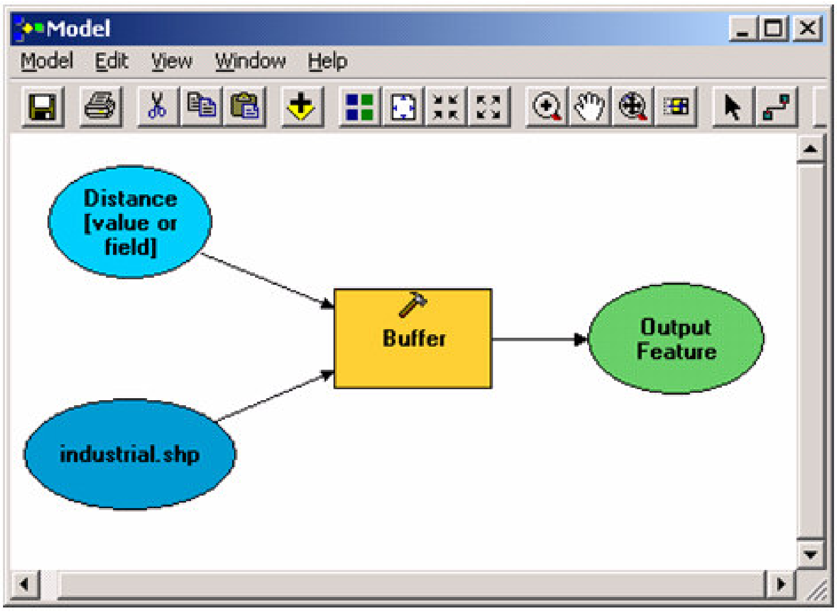


Figure 17. The model with the buffer distance extracted into a separate variable element

Converting the buffer distance elements and the output feature into a parameter can be accomplished by simply right-clicking on the element and selecting Make Parameter (Figure 18).

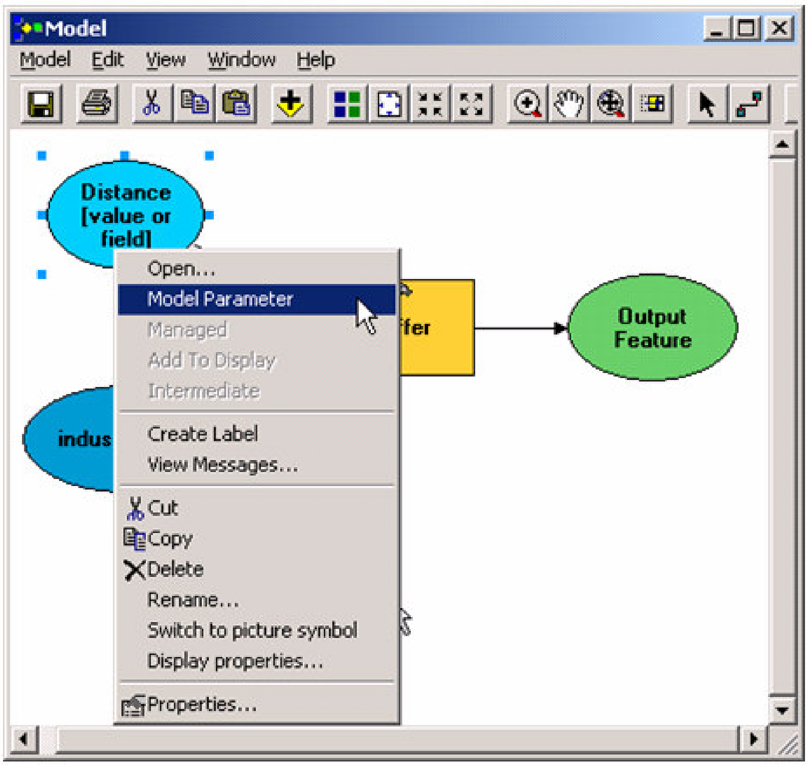


Figure 18. Creating a Model Parameter

Now that we have Parameters set, when we double-click on the model name in ArcToolbox, we see the following User Interface presented (Figure 19).

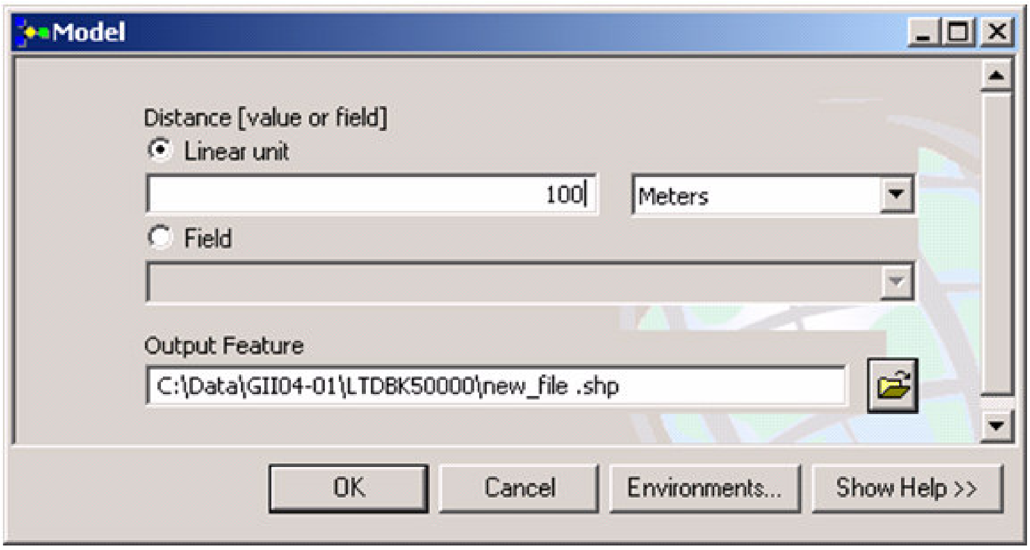


Figure 19. The User Interface for the completed buffer model.

Model Builder will support relatively large models, with hundreds of operations. If a model is going to repeat the same functions a number of times, then small models can be imported as tools into an overriding model. This gives Model Builder the ability to have subroutines to simplify model operation.

Model Builder gives non-programmers to capability to create reasonably complex models. The Flow Chart metaphor makes it reasonably easy to build, modify, and document models. The ability to convert models into a formal scripting language means that it can also help programmers to quickly lay out many lines of code, which can then be modified using an editor.

2.3.3 Allocation

Related to the idea of optimizing location is the concept of determining which people are best served by which facility. This is best exemplified by determining which ambulance station best serves which residents. This "zone of allocation" will be irregularly shaped, and will depend on road conditions and the configuration of roads in the local area.

It is important to note the difference between Natural and Mandated Service Areas. Using GIS, we attempt to determine the Natural Service Area, which identifies the facility to which people would normally go, based on their needs. Mandated Service Areas are the areas that are assigned to hospitals, ambulances, police stations, or individual police officers. In an ideal situation, the Mandated Service Areas are equal to the Natural Service areas, but more often than not there will always be areas that fall outside a particular Natural Service area, which somebody needs to be responsible for. The use of Allocation Analysis should help to rationalize Mandated Service Areas, so that they are as close to Natural Service Areas as is possible (Cromley and McLafferty, 2002).

Allocation problems can be solved using both vector and raster data structure. In urban environments, where virtually all travel is done all road networks, the vector technique is preferred. When overland travel is a factor, and vehicles are not necessarily limited to road travel, the raster technique has advantages.

Vector Technique

In the Vector Technique of Allocation analysis, we make use of a transportation network to determine the number of houses were people that are allocated to a particular facility. The use of a transportation network means that "... the capacity of a given service is distributed throughout the net[work]" (DeMers, 2005, p. 297).

To solve allocation problems, we first need to know the facility to which people are being allocated, and we need to know the locations of people. Generally, the location of the facility is well known, but the distribution of people surrounding the facility may not be as well known. The best solution is to know how many people live on each segment of street surrounding facility, but this information is often unavailable.

There are several sources of data that we can use to estimate the number of people surrounding the facility for which we wish to perform allocation analysis. Parcel-level data may be available in a Multipurpose Cadastre. By determining the approximate population per parcel based on the zoning of that parcel, we can come up with a fairly accurate estimate of the number of people on each parcel.

Another possibility is to use Census data to determine the populations of the census districts surrounding the facility, however these data tend to be highly agglomerated, and may not be suitable for allocation of analysis for small facilities, such as elementary schools that may only have 100 or 200 students. One important consideration about Census data is that it typically records populations at midnight, local time, so as to minimize the number of people who are in transit. If are trying to allocate based on daytime populations, it is important to realize that many people will be at work at this time, and a population distributions will be significantly different than what is recorded in the census, particularly if people perform long-distance commuting to get to work.

Once we have the location of the facility and the approximate distribution of population around that facility, we can begin to allocate people. Beginning at the location of the facility, groups of people (based on parcels or census tracts) are allocated to the facility at increasing distances. Every time a group of people is added to the facility, the allocation zone for that facility increases to incorporate the boundary of the area from which the people came. This process continues until the capacity of the facility is reached, at which point he allocation process stops, and the final boundary of the zone of allocation is determined (DeMers, 2005) (Figure 20).

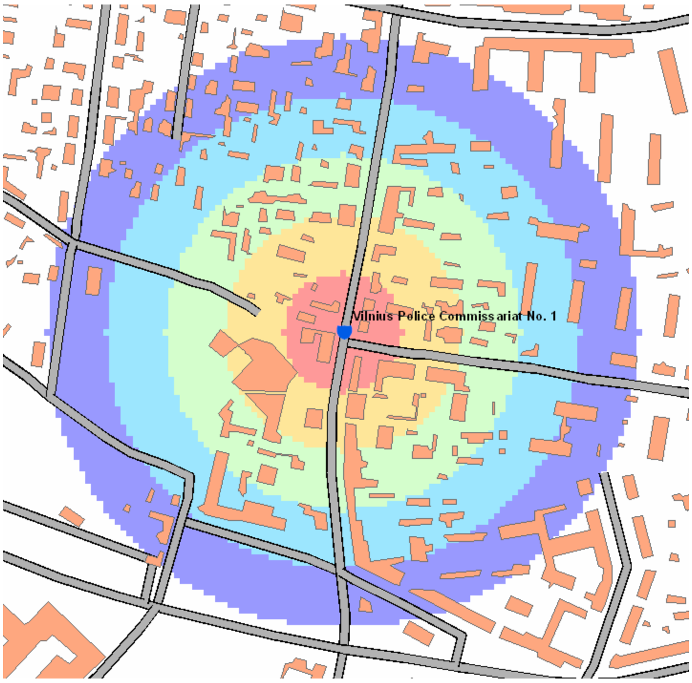


Figure 20. Allocation of buildings to the Vilnius Police Commissariat No. 1. Buildings are assigned based on distance, so the buildings in the red zone are assigned first (based on building centre -- 8 buildings), then in the yellow zone (a further 23 buildings) and so on until the capacity of the station has been reached. Allocation may also be done based on a maximum distance (the outer edge of the blue area is 250 metres from the police station).

Allocation can also be done on a network based on travel time. In this case, we can see the amount of time it will take a police car to reach particular location in 1-minute increments from some of the police stations in Vilnius (Figure 21).

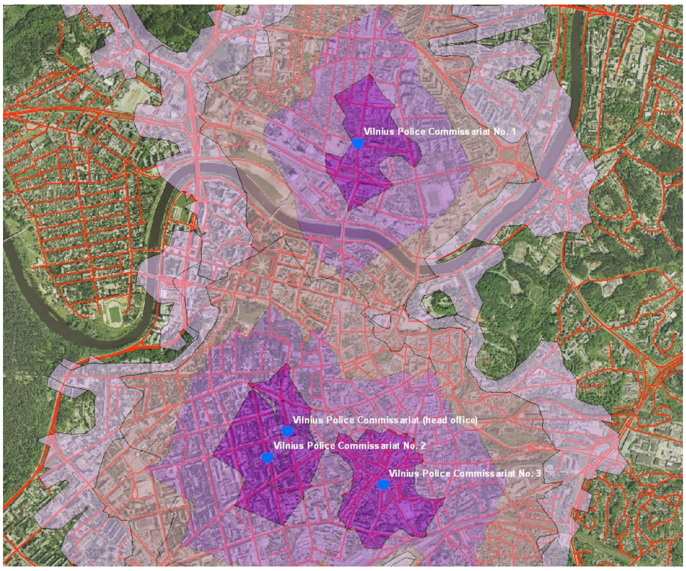


Figure 21. Allocation analysis of police stations in Vilnius using the Vector Technique.

In addition to capacity limits, we can also place limitations on a maximum distance to be travelled (for example, when young children must walk to school), the maximum travel time (for example, ambulance or police response times), or time limits on service (for example, that couriers must deliver overnight packages by 9 a.m.) (DeMers, 2005). Minimum distance is calculated using the p-median location model, and the minimum travel time is calculated using the Maximum Covering Model (Chang, 2006).

Of course, it is not necessary to limit the service center size; it may also be worthwhile to determine the absolute capacity of each service center, if all people that were close to it made use of it. Analysis such as this can help us to determine how much over capacity a particular service center is, which can be an important tool in the planning of new service centres.

Allocation analysis may indicate that there are sufficient facilities to support the population (with an acceptable level of overcapacity in case of emergency), that there are too few facilities, which commonly occurs in areas that are growing rapidly, or too many facilities, which can occur as a result of population shrinkage. Of course, a location-allocation model cannot account for every factor, and it may be necessary to consider factors that are not modelled, such as higher needs for services in particular populations

Raster Technique

It is also possible to solve Allocation problems using raster data. In this case, a capacity value is assigned to each cell in a raster, and cells are allocated at increasing distances from the service centre until the capacity of the service centre is reached. However, it should be noted that for most urban applications, the use of a road network in a vector model is the best way to represent allocation problems; raster allocation problems may be more appropriate when travel is not confined to a network of roads, as in wildlife movement studies and analysis of overland transport (DeMers, 2005).

With the Vector Technique for Allocation, we assume that travel will be along roads. The allocation is extended outwards from the facility, and the allocation limits along the road are used to "connect the dots" to include those areas that are not on a road. The Raster Technique for Allocation works somewhat differently. Every cell in an area has a cost value. Along roads, the cost is very low, and through buildings, the cost is so high as to be essentially infinite, since it is virtually impossible to travel through a building. The more difficult it is to travel over a surface, the higher the travel cost.

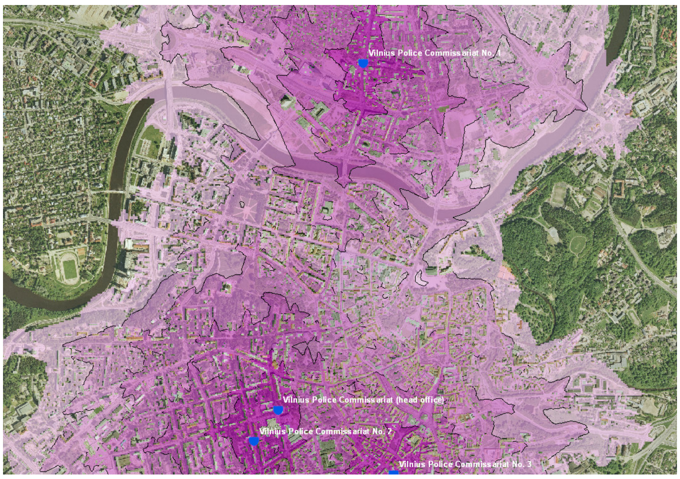


Figure 22. Raster Technique for Allocation. Allocation based on a maximum of 10000 units allocated and cost surface with roads costing 5 units per cell, bare land costing 15 units per cell, and buildings costing 9999 units per cell (impenetrable). Note: land use was not explicitly modelled, so movement over water is possible!

2.3.4 Location

One basic truth of the real estate business is that the three most important factors in choosing a property are "location, location, and location." Although this is a cliché, for private businesses, choosing an effective location may mean the difference between success and failure. Retail businesses must be located in areas where their product is in high demand, where they are accessible, and where they have high visual exposure. It is not surprising that the business community was the first breakthrough market when GIS software became available to optimize the location of new facilities.

Although the implications of facility siting are less severe for public sector agencies, there is still a great deal of money involved. A poor location may mean the difference between having to construct one facility or two, at a cost of millions of Litai. Furthermore, a facility may be underutilized if it is not placed in an optimal position. This is clearly a waste of government finances, and should be avoided at all costs.

This section will discuss three different techniques for location modeling, beginning with Index Modeling, which can be used for course site analysis, then continuing with Gravity Modeling, which allows more sophisticated analysis of individual locations with respect to other operations, and finishing with Regression Analysis, which can be used to predict the sales that can be expected for a store at a particular site.

Index Modeling

This was discussed above in general terms in Section 0. Index Modeling to determine location involves the selection and weighting of factors such as zoning, buildings or parcels of land that are available, demographics, distance from other facilities, accessibility, and cost to buy or lease. The normalization, weighting, and combination of these layers, using vector overlay techniques or Map Algebra can create a sophisticated model that allows prospective locations to be shortlisted for further evaluation. Through techniques such as these, millions of possible sites can be narrowed down to perhaps a dozen, which makes it possible to evaluate sites using more sophisticated techniques such as Gravity Modeling.

Gravity Modeling

Gravity Modeling is a way of evaluating potential service locations to determine the impact of the new services in relationship to the existing surrounding services. For example, if a new store was proposed for a particular location, Gravity Modeling used to determine whether the new store serves a new population, or whether it competes with existing stores in the area. If the store is found to be in competition, the expected revenues from the new store location can be predicted based on the distance of the new location from existing stores.

In its most basic (non-spatial) form, gravity modeling calculates the probability of a person using a particular facility versus other facilities. This can be calculated by determining the utility the proposed facility and dividing it by the utility of all other facilities.

$$P_{ij} = \frac{U_j}{\sum_{j=1}^{n} U_j}$$

where P_{ij} represents the probability of user i using facility j, U represents the utility of each facility, and j represents the facilities.

In its geographic form, the Gravity Model incorporates the floor area of each facility and a distance between the user and each facility. In this case, the floor area of each facility using easily measured proxy for the utility of the facility, but it may be better to calculate this based on the different types of merchandise that are available, or the appeal that the products have to the users.

$$P_{ij} = \frac{S_j^{\alpha} D_{ij}^{\beta}}{\sum S_j^{\alpha} D_{ij}^{\beta}}$$

where P_{ij} represents the probability of user i using facility j, S_j is the floor area of each facility, D_{ij} represents the distance from the user to the facility, and α and β are parameters that control the relative effect of floor area and distance to customer based on the type of service is being modelled. For convenience products, such as a meal at a fast-food restaurant, α and β are larger, whereas for specialty items, such as purchasing a particular brand of watch, these factors are smaller, because people are willing to travel greater distances, and are willing to seek out these items.

The method for calculating distance is a critical factor in determining the probability of using a particular facility. Although it might initially make sense to consider straight-line distance, quite often the actual route taken along the road deviates from the straight-line distance by a large amount. In urban centres that are laid out on a grid, it may be better to conceptualize distance using the Manhattan Distance, which assumes that diagonal travel is not possible, but that any combination of streets and avenues will result in the same distance. Although Manhattan Distance is easy to calculate (the difference in X-coordinates plus the difference in Y-Coordinates between source and destination), it does not work well when roads are not laid out on a regular grid. In this case, the best conceptualization of distance is the Network Distance, which is calculated by following the network from source to destination, and takes into account winding roads.

Euclidean Distance Formula:

$$d_{ij} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}$$

Manhattan Distance Formula:

$$d_{ij} = \left|X_i - X_j\right| + \left|Y_i - Y_j\right|$$

Finally, the expected revenues and each facility can be modelled by multiplying the probability of using a service by the amount that has been budgeted for the purchase of that service.

$$E_{ij} = P_{ij} B_{ik}$$

Where E_{ij} is the expected revenue for user i at facility j, P_{ij} represents the probability of user i using facility j, and B_{ik} is the budgeted amount for user i for item k. The higher the budget, the more probable it is that a user will purchase or make use of the service. It is important to note that the model must be calibrated for the type of goods that it is modeling, in order to work (Environmental Systems Research Institute, 2007).

The last equation provides a relatively simple estimate of expected revenues for a particular location. Once the number of potential locations has been reduced to two or three, we can employ

Regression Modeling to make a much more sophisticated prediction of store revenues, based on the performance of existing stores.

Gravity Models have also been used to measure the accessibility of healthcare facilities. Rather than looking at the floor area of the store, in healthcare we are concerned with such factors as price, the quality of services, cultural appropriateness, acceptability to insurance companies, and other factors (Cromley and McLafferty, 2002).

Regression Modeling

Multiple regression is another way of predicting store sales. By creating a formula to predict sales based on the type of store operations, the quality of the site and its location, the predicted sales can be predicted. Of course, it is important to realize that there may be unconsidered factors at work when a new store is located, so a regression equation based on the sales at existing stores may not apply fully to a new equation. Furthermore, the retail sector is a dynamic environment, so a multiple regression equation that is developed today may be obsolete in a few months, as soon as a new type of product makes its way into the market.

2.4 Optimization

Optimization is the process of adjusting the parameters in a model until it produces an optimal (or at least acceptable) result. Models may be optimized manually, or automatically by using a computer program. Using a computer program has the advantage of allowing thousands or millions of possible permutations to be tested, where as manual approaches are limited to several dozen. The advantages of being able to explore more options need to be weighed against the difficulty of programming the model. If there are only a few variables to be optimized, manually optimizing them may be the best option.

If we assume that all possible results form a surface in n-dimensions, based on n parameters to be optimized, then it is possible to think of the most optimal solution as the one with the least error. In other words, the surface contains high points, which are poorly optimized, and low points, which are highly optimized. The optimal solution is then the lowest possible point on the surface.

This gives rise to the *gradient descent algorithm*, which assumes that if you adjust the variables that you are trying to optimize so that you always move "downhill," you will eventually reach an optimal solution, i.e. the lowest point on the surface, which has the least predictive error. To do this in practice involves a directed search in which variables are adjusted, and the results of the model are examined against some criteria. When working with multiple variables, one is adjusted at a time until the best possible setting for that variable is discovered. That variable value is then fixed, and the next variable is adjusted until the best possible setting for the second variable is discovered. The first and second variables are then fixed, and the third variable is adjusted in turn, until the best possible solution is achieved (Figure 23).

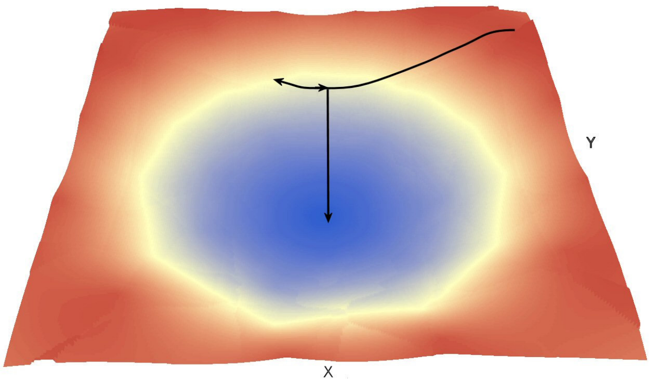


Figure 23. Optimization is relatively simple when there is a single minimum in the multidimensional surface. Here, we optimize the X value first, and then optimize the Y value.

In essence, each variable to be adjusted is one dimension in a multidimensional model. By optimizing each variable (dimension), it is possible to come up with an overall best solution. However, the gradient descent algorithm assumes that the multidimensional surface is smooth, and that there is only one minimum point. On a complex surface, there may be multiple local minima, but only one optimal solution. The problem with the gradient descent algorithm is that it may identify a local minimum. The solution to this problem is to randomize the variable settings once a local minimum has been found, to see whether randomly changing position on the multidimensional surface can place them on a slope that leads into the global minimum, in other words the optimal solution (Figure 24).

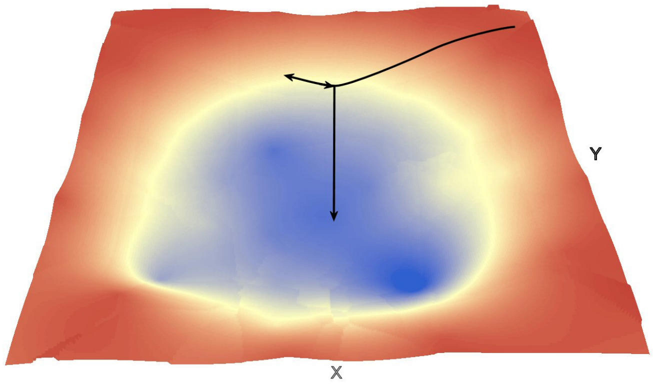


Figure 24. The gradient descent algorithm does not always return the lowest point on a multidimensional surface if there are multiple minima. This can be resolved by randomizing the starting variables and running multiple trials. Here the X variable is optimized first, followed by the Y.

If there are only a few variables to optimize (up to four), it is relatively easy to perform the gradient descent algorithm manually, keeping track of the variable values and the overall model result (Table 3). If there are more variables to be optimized or you have reason to believe that the surface is unusually complex (which normally requires more variables), then automating the gradient descent algorithm is the best option.

Table 3. An example of how to use a table to perform the gradient descent algorithm manually. Here we are working with three variables that must be optimized. We improve the result from 1.0 in the first trial to 3.0 in the 11th trial.

Trial	Variable 1	Variable 2	Variable 3	Result Being Optimized	Comment
1	1	2	3	1.0	Begin by changing Variable 1 only; 2 and 3 are kept constant. Start with large (1.0) jumps in Variable 1
2	2	2	3	2.0	Result improved. Keep going
3	3	2	3	1.0	We now know the optimal value for Variable 1 is around 2. Now start varying Variable 1 by smaller (0.2) jumps
4	2.2	2	3	2.2	Result improved. Keep going
5	2.4	2	3	2.1	
6	2.3	2	3	2.3	Variable 1 is optimized; hold it steady and begin modifying Variable 2
7	2.3	3	3	2.8	Result improved. Keep going
8	2.3	4	3	2.2	We now know that the optimal value for Variable 2 is around 3. Now we start varying Variable 2 by smaller (0.2) jumps
9	2.3	2.8	3	2.6	
10	2.3	3.2	3	2.9	
11	2.3	3.1	3	3.0	Variables 1 and 2 are now optimized; hold these values and begin modifying Variable 3

2.5 Areas of Application

2.5.1 Healthcare

Location and Allocation problems are very important in the provision of healthcare. Having a healthy population has many components, including providing physicians, having hospitals, Public Education programs, a clean Water Supply, and providing Mental Health and Social Services. GIS can help in all of these components in some way.

Using Allocation Analysis, we can determine the location of healthcare shortage areas, which can be then used as potential sites for new healthcare facilities, or which can be used to assist public health nurses, physicians, and other healthcare providers in targeting specific underserved populations.

In healthcare, it is important to note that there is a *threshold requirement*, which is the minimum demand or volume of service that is required to support a new healthcare facility. This means that, although there may be need for a facility, that need may not be sufficient to justify the construction of a new hospital or clinic, so other means, such as household visits must be found to meet the demand. Furthermore, there may be a limit on the capacity of a particular healthcare facility, such as the number of physicians on-call or the total number of beds. Healthcare facilities may also have minimum standards for service delivery, which leads to overcapacity at all times except during high-volume times and emergencies.

One problem with healthcare is that different forms of healthcare have different distributions. Take, for example, the difference between neonatal care units and oncology (cancer) clinics. If the elderly live in different areas than young mothers, then there will be a difference in the service areas for these two types of medical services, therefore it may make sense to have a number of smaller, specialized facilities then a single centralized hospital providing services to everybody (Cromley and McLafferty, 2002).

2.5.2 Policing

Policing is another area where GIS can provide many different services. Perhaps the simplest, but most effective use of GIS is to promote data sharing between different police forces and other agencies that assist in crime fighting. Organized criminals have long taken advantage of the poor communication between different police agencies. Often these groups will spread their crimes through different jurisdictions, so that the police in each jurisdiction are aware of only part of the problem, and find it difficult to prosecute the criminals as a result.

When policing and non-policing organizations work together, it is possible to solve problems that were previously ignored. For example, in Huntsville, Alabama, records of automobile accidents from the police are used to support the planning and public works departments in redesigning intersections that are prone to vehicle accidents. In Clark County, Nevada, crime records are used to help Nevada Power to determine locations where street lighting can be improved as an environmental deterrent to crime. In Lincoln, Nebraska, records of juvenile crime are used to help locate new community facilities and neighborhood groups, which help prevent youth crime by providing healthy alternatives to getting involved in crime (Leipnik & Albert. in Wang, 2005).

Perhaps the area of greatest interest in GIS by policing agencies is being what is called "Geographic Profiling." Geographic profiling is the analysis of related crimes to determine additional information about the perpetrator. Two basic facts about criminal behavior are used in Geographic Profile. The first is that criminals tend to limit the amount of effort that they exert in perpetrating a crime. The second is that criminals avoid perpetrating crimes in places that would implicate them. Taken together, this implies that a criminal will perform crimes in a ring around his current home, previous home, or place of work, and that the ring will tend to be centred on their home or place of work.

Geographic Profiling produces a *jeopardy surface*, which represents the probability of a criminal living in a particular location. The area highlighted with the jeopardy surface will typically be only a small proportion (about 5%) of the total area covered by the criminal while the crimes are being perpetrated. Although this technique will not point out the exact place of residence or work of the criminal, it will significantly narrow the possibilities, providing police with one additional clue that they would not otherwise have (Rossmo *et al*. in Wang, 2005).

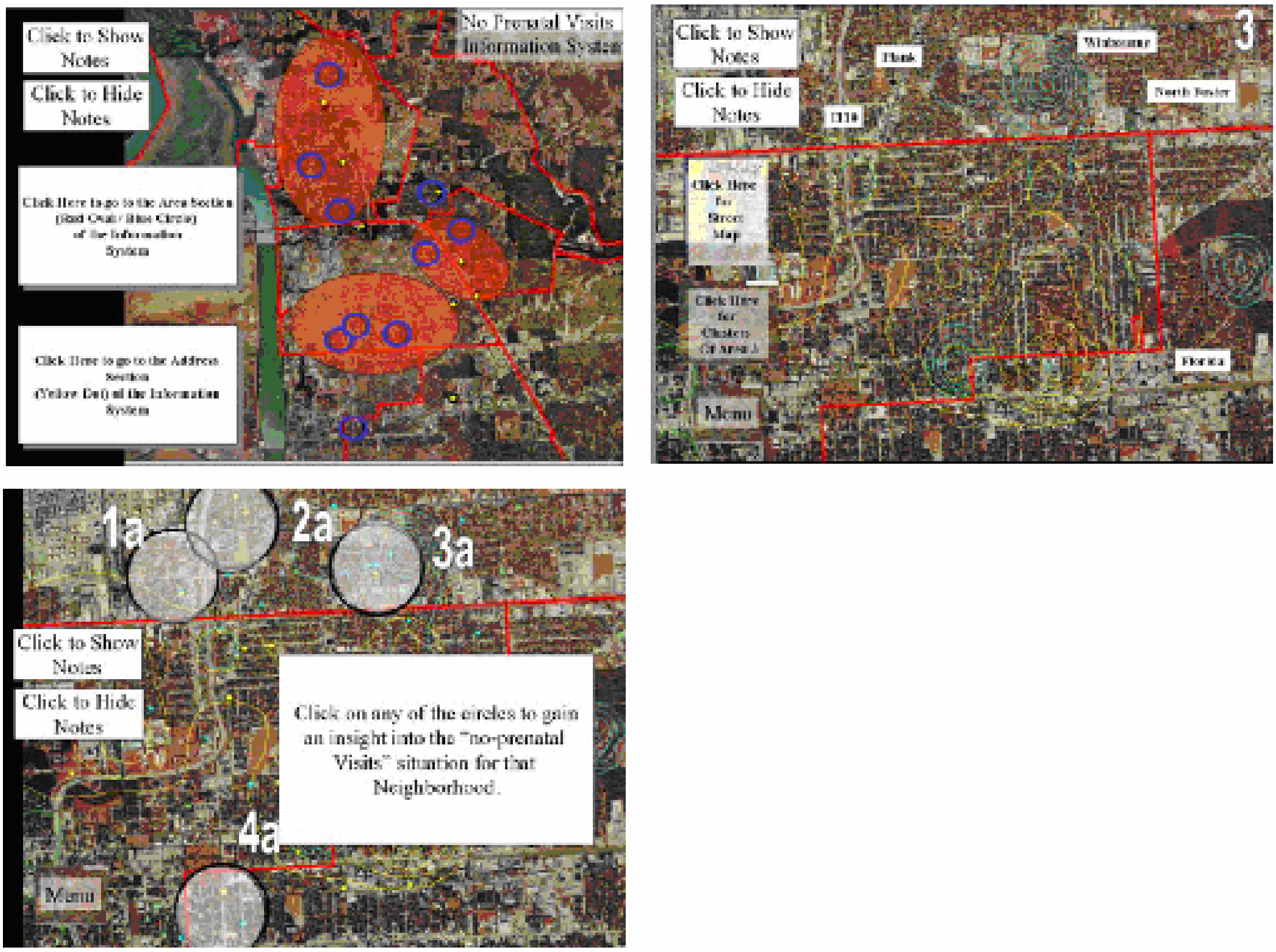
Location Analysis can also be used to determine areas for police patrols based on the distribution of crime. This helps to ensure that police officers are working where they are most likely to be needed (Curtin *et al.* in Wang, 2005).

2.6 Examples

2.6.1 Allocation Problems

In Baton Rouge, Louisiana, USA, the Baton Rouge Healthy Start GIS has been used to model access to healthcare for pregnant women in a 39 square-mile area in the central part of East Baton Rouge Parish. Certain portions of Baton Rouge have very high infant mortality rates, as a result of poverty, delays in becoming available for government-assisted health care, and difficulties in reaching health care facilities. In places where accessibility is low, health workers are able to use data from the GIS to target pregnant women to ensure that they obtain needed perinatal care. Completed maps are loaded into Powerpoint presentations that are downloaded to laptop computers used by health care workers in the field (Figure 25) (Curtis and Leitner, 2006).

Figure 25. Screens from interactive PowerPoint presentations that are used by health care workers to target



underserved pregnant women in Baton Rouge, Louisiana (Curtis and Leitner, 2006)

2.6.2 Location Problems

Federated Department Stores

A fairly classic example of using GIS to solve Location Problems is provided by Federated Department Stores in the United States, which operates more than 460 department stores in 34 states as well as in Guam and Puerto Rico. This company has been using GIS to guide it in an aggressive expansion program. New marketing strategies are being developed for existing stores, and smaller stores are being opened in areas with less "market penetration."

To do this, Federated uses many different databases, including Federated proprietary databases and databases collected from government and private industry. Through a process of index modeling, Federated combines and weights data to create maps showing an index of market penetration (Figure 26) (Environmental Systems Research Institute, 2007).

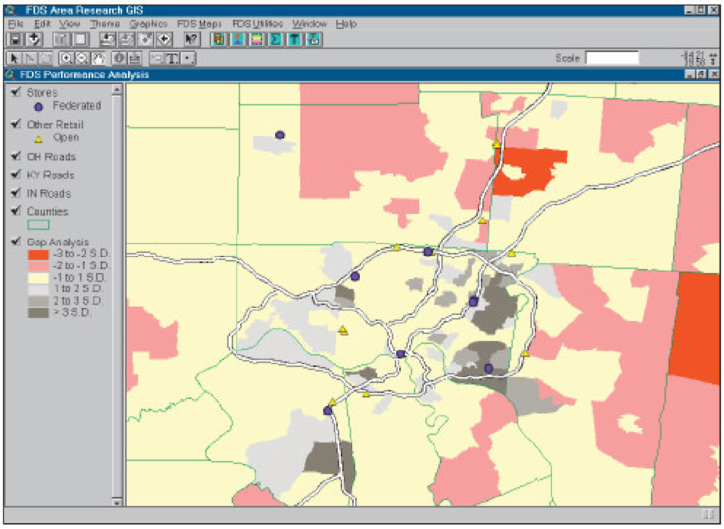


Figure 26. Map of market penetration used by Federated Department Stores: red represents under penetration and grey represents over penetration (Environmental Systems Research Institute, 2007).

Storage USA, Inc.

Storage USA, Inc. is another company that uses Location Analysis to determine locations for new facilities. Storage USA operates a network of 500 mini-storage parks, where people can rent storage lockers. Because each new facility costs US $4 million to develop, it is critical that these facilities be placed in the right location.

In general, sites need to be developed on large areas of vacant land in close proximity to subdivisions. To determine this, the company uses home values, income ranges, and population density to evaluate potential sites. Every year, Storage USA evaluates hundreds of potential sites for many-storage parks. This company uses a wide variety of GIS techniques to evaluate potential sites, including Index Modeling to determine sites with the correct demographics, Network Analysis to determine traffic volumes near potential sites, and Gravity Modeling to determine the influence of competitors at locations of interest.

Because Storage USA performs the same type of analyses for every site of interest, the application can be easily automated. Ninety percent of site analysis is now completely automated, allowing Storage USA to produce site location reports faster and at lower cost than would otherwise be possible (Environmental Systems Research Institute, 2007).

Need for Dentists

McSorley (1999) uses Binary Modeling to determine communities in New Jersey where childrens' dental health was the worst and where more publicly funded dental services are required. To do this, layers representing demographic characteristics such as the number of children below the poverty line, and the ethnic makeup of populations were overlaid with layers showing which water supplies were fluoridated, and where there were shortages of dental health professionals to obtain a map showing the areas of greatest need in New Jersey (Figure 27).

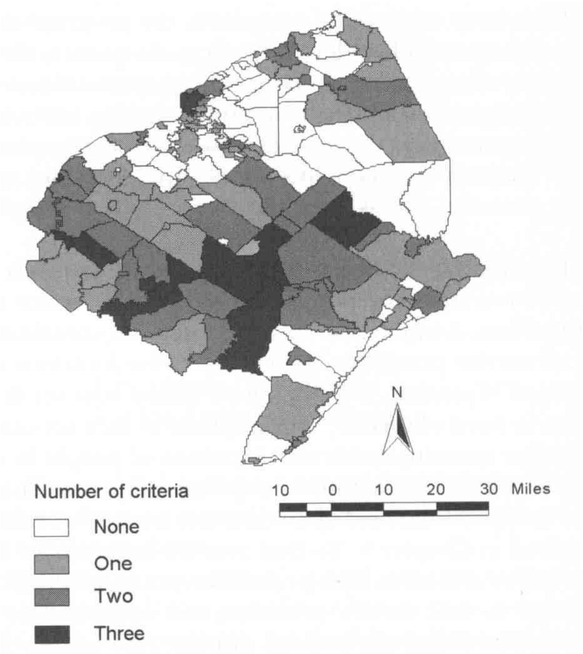


Figure 27. Communities requiring additional public dental services in New Jersey (McSorley, 1999 in Cromley and McLafferty, 2002)

2.7 Summary

In this module, we discussed a number of technologies that can be used to store and retrieve locations of parcels (Multipurpose Cadastre), to build models that show how locations are connected (network construction), how to determine the number of people who will use a particular facility (allocation), and how to determine optimal position for one or more new authorities (location). By themselves, these technologies are relatively straightforward, and offer a few advantages, and when combined, they give rise to a number of specific applications in different areas.

Analyses such as these show which areas are served well by existing facilities, and which areas are served poorly. New facilities can be positioned manually or can be optimized to reduce the differences between under-represented areas and over-represented areas. The overall goal is to ensure that minimum levels of service are established for the entire population. Behind this goal is the implicit assumption that the most efficient provision of services is the one that is most equal. Furthermore, this type of analysis highlights waste, when facilities are located to closely together so that their jurisdictions overlap.

Over the course of decades, planning the location of facilities and an optimal fashion leads to increasing levels of efficiency in the provision of public services. In most cases, this can be accomplished at very little cost, since, for various reasons, facilities must be opened and closed in any case. It made widely available, this type of analysis helps to support urban planning initiatives. For private businesses, using location and allocation tools to determine the best business location can make the difference between profitability and bankruptcy.

Module Self-Study Questions

1. Can you think of some examples of a Prescriptive Model? A Descriptive Model?

2. If you were in charge of locating a new high-end restaurant in Kaunas, describe some of the layers that would be helpful to produce an Index Model of the most appropriate locations.

Answer: Census data (income, number of children), digital elevation model (for calculation of views), road network (for determining accessibility), locations of competing restaurants (for buffering or gravity modeling), survey data (number of people on the sidewalks, questions about food preferences)

3. The rapid growth of Vilnius has encouraged the government to invest in a new hospital. Describe how location and allocation analyses can be used to optimize the location of the hospital.

Answer: Examine distribution of different types of health needs, perform hot spot analysis (are the patterns for some different types of health needs similar?), use Index Modeling to examine accessibility and availability of different sites, perform Gravity Modeling to determine influence of other healthcare facilities, attempt to match available locations with healthcare needs.

Required Readings

Environmental Systems Research Institute (2007). ArcGIS Best Practices: GIS for Retail Business (http://www.esri.com/library/bestpractices/retail-business.pdf, July 15, 2007)

Environmental Systems Research Institute (2006). *GIS Best Practices: Land Records and Cadastre*. Redlands, California: Environmental Systems Research Institute. (http://www.esri.com/library/ bestpractices/land-records-cadastre.pdf)

ESRI Virtual Campus Module

- N/A

Assignments

1. Assignment 1: Optimizing Emergency Service Facilities

References

Chang, Kang-tsung (2006). *Introduction to Geographic Information Systems, 3rd Ed*. New York: McGraw-Hill

Cromley, Ellen K., & McLafferty, Sara L. (2002) GIS *and Public Health*. New York: Guildford Press

Curtis, Andrew and Leitner, Michael (2006). Geographic Information Systems and Public Health: Eliminating Perinatal Disparity. Hershey, PA: IRM Press.

DeMers, Michael M. (2002). *GIS Modeling in Raster.* New York: Wiley

DeMers, Michael N. (2005) Fundamentals *of Geographic Information Systems, 3rd Ed.* Hoboken, N.J.: Wiley.

Environmental Systems Research Institute (2006). *GIS Best Practices: Land Records and Cadastre*. Redlands, California: Environmental Systems Research Institute. (http://www.esri.com/library/ bestpractices/land-records-cadastre.pdf)

'

Environmental Systems Research Institute (2007). *ArcGIS Best Practices: GIS for Retail Business* (http://www.esri.com/library/bestpractices/retail-business.pdf, July 15, 2007)

ESRI GIS Dictionary http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.search (July 11, 2007)

Groot, Richard & McLaughlin, John (2000). *Geospatial Data Infrastructure: Concepts, Cases, and Good Practice*. Oxford: Oxford University Press.

Wang, Fahui (2005). *Geographic Information Systems and Crime Analysis*. Hershey, PA: Idea Group

Zeiler, Michael (1999). *Modeling our World: The ESRI Guide to Geodatabase Design.* Redlands, California: ESRI Press.

Terms Used

- Accumulative Operators
- Allocation Problems
- Arc
- Arithmetic Operators
- Assignment Operators
- Binary Models
- Bitwise Operators
- Block Functions
- Boolean Operators
- Combinatorial Operators
- Conductivity
- Deductive Model
- Descriptive Model
- Deterministic Model
- Dynamic Model
- Edges
- Euclidean Distance
- Fiscal Cadastre
- Flowcharts
- Focal Functions
- Geocoding
- Geographic Profiling
- Geometric Network
- Global Functions
- Global Positioning System
- Gradient Descent Algorithm
- Gravity Model
- Impedance
- Index Models
- Inductive Model
- Judicial Cadastre
- Junctions
- Location Problems
- Logical Network
- Logical Operators
- Mandated Service Areas
- Manhattan Distance
- Map Algebra
- Maximum Covering Model
- Model Calibration
- Model Validation
- Multipurpose Cadastre
- Natural Service Areas

- Network
- Network Distance
- NoData
- Node
- Normalization
- Optimization
- Physical Network
- Planar Graph
- p-median Location Model
- Prescriptive Model
- Process Models
- Raster Calculator
- Regression Models
- Relational Operators
- Stochastic Model
- Threshold Requirement
- Topology
- Transportation Network
- Turn Table
- Turns
- Zonal Functions
- Zone of Allocation

3 Network and Unconstrained Travel

Outline:

13. Introduction
14. Construction of Networks
15. Geocoding
16. Calculation of Distance
17. Routing
18. Functional Distance
19. Utility Network Analysis
20. Applications
21. Examples

3.1 Introduction

In the last module, we looked at issues of *location*. We examined how we could determine the best location for a new facility, and we then looked at how to determine the capacity of that facility, in terms of numbers (people, buildings, census polygons) or of distance (travel time). We examined algorithms to allow us to allocate people, buildings, or groups of people to particular facilities, and we looked at ways to determine optimal locations, such as binary models, index models, gravity, and regression models. We discovered that there is a hierarchy in the determination of location for a new facility; index or binary modeling can first be used to narrow down many candidate locations to a few that are worthy of further analysis, and then gravity and regression modeling can then be used to predict the number of customers that a facility at each of the prospective sites will attract.

In this module, we will look at issues of transportation. Obviously, there is a strong connection between the previous module and this one. Both contribute to our understanding of planning, and can be combined to create some very important applications.

Once again, the development of a Geographic Information Infrastructure (GII) is the key development that makes these applications widely available. With quality data that is accessible, it is increasingly easy for computers to run the types of analyses that we will discuss in this module.

We will begin with some fundamentals. All of the applications that we discuss in this module are based on networks, which are based on a layer of features representing roads, railroads, pipelines, electrical transmission lines, rivers, or airline routes. In terms of shape, networks tend to be of three possible forms: non-intersecting lines (e.g. airline routes), branching lines (e.g. rivers), and circuits (e.g. roads) (Figure 28). In this module, we will discuss the construction of transportation and utility networks separately, because their behaviour is fundamentally different. Although networks are based on layers containing lines, which may have arc-node topology, it is important to note that the construction of a network entails the addition of information that allows a *higher order* of topology to be created (DeMers, 2005).

Figure 28. Airline routes (left), rivers (centre), and road networks (right) are examples of three possible network types using straight lines, branching lines, or circuits.

Once a network has been constructed, it permits many types of analysis. One of these is finding the best route along the network between two points. But how do we find the start and end point for routing? As we saw in the previous module, we can use various techniques to locate facilities. If the X, Y Coordinates of two facilities are known, then we can determine the best route between the two. However, we often want to determine the best path from a facility at a known location to a particular address. To do this, we need another fundamental technology, known as *address geocoding*. Address geocoding is used to provide an *estimated* X and Y Coordinate for a particular address. We can then determine the best route from a particular facility to an approximate address, or between two approximate addresses.

But how far apart are these addresses? That depends on how you consider the problem. In the simplest form, we simply draw a straight line from point to point. Unfortunately, only birds regularly travel this way! We could consider the distance along the roads, which is also known as the network distance. We could travel overland as well. There are many different ways of calculating distance depending on our mode of travel. We will discuss the different ways of calculating distance, and the advantages and disadvantages of each.

Next, we will have a close look at utility networks, and how they can be used to model the flow of commodities. Modeling of this type is extremely valuable to the companies that provide the utilities, since they can provide their services more efficiently, with fewer and shorter interruptions. Electricity, water, natural gas, oil, sewage, and rivers can all be modelled by utility networks.

Once we have discussed the fundamentals, and then the different kinds of analysis that networks and geocoding make possible, we will look at how these types of analysis benefit people, concluding this module with some concrete examples of where and by whom these techniques are being used.

3.2 Construction of Networks

Much of our society's infrastructure is based on networks, so the ability to create networks gives us a tool with which we can model much of what drives our society. A network is a series of connected lines, which can be used to simulate travel from a start point on one line to an end point on another line.

Virtually all networks are based on a vector data structure. This is because "raster GIS is not particularly efficient at handling networks because there is no way to define them explicitly except to assign specific attribute values to the grid cells" (DeMers, 2005, p. 189.) There are two types of networks, in terms of how objects or commodities move through them. In the case of commodities such as oil and water, there is no autonomous control of where they go. Gravity or pressure within pipes forces these fluids through the network, and they are controlled by sources of the commodities, places where the commodities exit the network (sinks), and valves, which direct the commodities into particular pipes. These are called directed networks. For objects that have autonomous control, such as cars on roads, or aircraft flying along flight routes, travel may occur in either direction along the network. These are called undirected networks (DeMers, 2005).

3.2.1 Network Properties

In order to represent transportation and utility networks, the properties of connectivity, directions, restrictions, and impedance must be accurately modelled in software.

Connectivity is the ability to trace the route from one point on the network to another point on the network by following a series of links and turns. Because of the nature of networks, there is often more than one route from start to finish; analysis of the network can help us optimize our route, based on our choice of fastest route, shortest route, or other restrictions such as choosing roads that are not closed to bicycles.

We cannot assume that roads always permit travel in both directions; therefore we must be able to model the direction of travel on each road segment. Furthermore, some roads may be permanently closed, or may be restricted to certain types of traffic such as buses and taxis.

Impedance values may be assigned to either roads or to turns. For example, a speed limit restricts the amount of traffic that can use a particular segment of road, and the time required to pass through an intersection or perform a left, right, or U-turn affects how many vehicles are able to pass through that intersection per unit of time. For intersections, it typically takes more time to make a left-hand turn than it does to make a right-hand turn. Straight through travel requires the least time, and U-turns, if permitted, take the most time of all.

When a network is implemented in software, it is done in two parts, the *physical network*, and the *logical network*.

3.2.2 Network Elements

The *physical network* consists of the arcs and nodes that make up the underlying layer on which the network is built. Arcs are line segments that represent portions of the network between intersections. Nodes are locations where arcs join together, and correspond to intersections or connections. The physical network contains the locational information and the shape of the road segments (Figure 29).

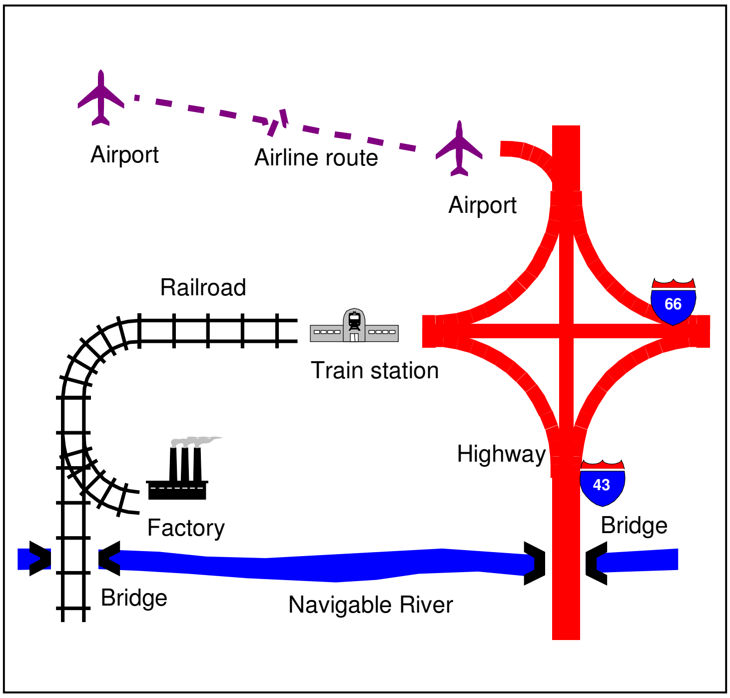


Figure 29. In the Physical Network, commodities flow from sources to destinations. A Network contains point features through which commodities for, such as railroad station. Source: Zeiler, 1999

Additional information is required to construct the *logical network*. The logical network consists of two variants, one for transportation networks, and another for utility networks. Logical networks contain additional information that is missing from the physical network, including edges, junctions, and turns (Figure 30). Edges provide the additional information for arcs, and junctions and turns provide the additional information for nodes.

In the transportation version of the logical network, edges are implemented as an attribute table for the arcs, which includes information on directionality, restrictions, and impedances. Road segments may be subdivided to represent different impedances as a result of a change in speed limit, problems with the road surface, or obstructions. Junctions are implemented as an attribute table for the nodes, which is called a "turn table." Turn tables contain information on the beginning and ending edges for each turn, the angle involved, and the turn impedance (amount of time required to complete the turn). If the turn impedance is negative (or infinite depending on the software that you are using), this indicates that the particular turn is prohibited (Figure 31).

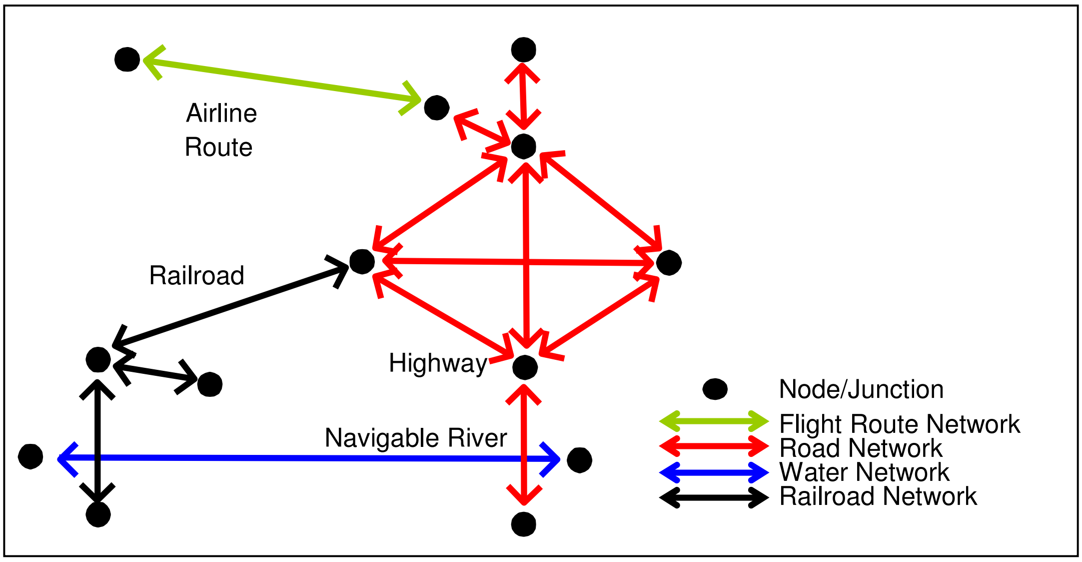
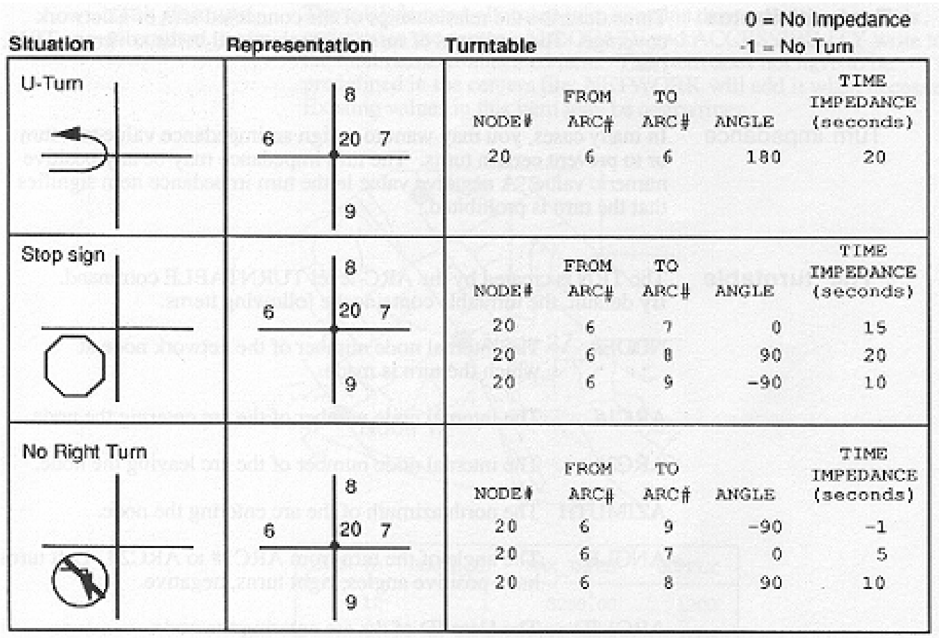


Figure 30. The Logical Network is composed of edges and junctions. Commodities flow through edges and edges connect together at junctions where flow from one edge is transferred to another. In a network, the geometry is not important, only the connectivity of edges and junctions (Source: Zeiler, 1999).



Situation	NODE#	FROM ARC#	TO ARC#	ANGLE	TIME IMPEDANCE (seconds)
U-Turn	20	6	6	180	20
Stop sign	20	6	7	0	15
	20	6	8	90	20
	20	6	9	-90	10
No Right Turn	20	6	9	-90	-1
	20	6	7	0	5
	20	6	8	90	10

Figure 31. Turn tables can be used to model the impedance for left, right, straight through, and U-turns. An impedance value of -1 means that a particular type of turn is not allowed.

One difficulty in the construction of the networks for urban areas is the number of intersections that exist. It would be a daunting task to enumerate each road intersection in a large city, and measure the time required for straight through travel, left, right, and U-turns. However, a properly structured sampling program can be used to obtain estimates of these turn impedances for intersections of a particular type. This information can then be transferred to all other similar intersections, allowing us to populate turn tables for an entire city in a reasonable amount of time. It is easier to obtain the impedances for road segments, since speed limits are usually well documented for particular sections of road.

The utility network version of the logical network is much simpler and contains an attribute table for arcs which includes the direction and weights (capacity, friction, size etc.), and an attribute table for nodes, which identifies whether junctions are "sources" or "sinks" for material.

Overpasses

Physical networks are generally identical to the underlying line layer, on which they are based, except with respect to overpasses. Normally, roads are represented as a *planar graph*; arcs in a road layer are broken when they cross to create a node. Overpasses require one arc to pass over the other without an intersection, which presents a problem.

One way of solving the overpass problem is by configuring a turn table to have zero impedance for straight through travel in either direction, and negative (or infinite) impedance on all turns to indicate that they are prohibited. A variation on the theme is to add a column to indicate relative elevation of the edges that connect at a node (Chang, 2006). In either case, even though the network forms a planar graph, an overpass is stimulated through the design of the node.

The other way to create an overpass is to deviate from a planar graph. To do this, we create a three-dimensional (non-planar) graph, which is created by allowing roads to cross without intersecting them with a node (McKinney, 2005).

Utility Networks

Utility Networks represent the channelled flow of commodities from a source to a destination. Utility Networks are less complex than Transportation Networks, since the commodities being transported are not autonomous, and move based simply on pressure gradients and the morphology of the network.

In a Utility Network, commodities flow from "sources" to "sinks," through a network of pipes or wires. In a network of water pipes, the sources would be the pumping and water treatment stations, where water is added to the system, and the sinks would be the places where the network connects to buildings, or fire hydrants. Any junction within the network may be defined as a source or as a sink. Sources and sinks help to define upstream and downstream directions in a utility network. The upstream direction is that part of the network that lies between the barrier and the source, and the downstream direction is that part which lies between the barrier and the sink. In other words, in a utility network, direction is an inherent part of the network, which is not the case for vehicles moving in a transportation network. When a Utility Network is first established, the directions of upstream and downstream have not been defined. This initial state, known as "uninitialized flow," may also occur when a utility network has been edited and has been disconnected from a source and/or sink.

Commodities flow from sources to sinks through pipes or wires, whose capacity and amount of friction are represented by weights. Weights are used to determine the "cost" of moving the commodity through the network. These can be thought of as being the amount of pressure required to move water through pressurized water lines, the gradient required to ensure that sanitary sewers do not back up, or the amount of resistance in an electrical transmission line.

Movement within the network is controlled using valves or switches. By opening a valve, the flow of commodities is permitted to a particular part of the network; and by closing one, the flow is interrupted to the part of the network beyond it. Opening and closing switches or valves changes the state of the portion of the network downstream from the switch or valve from "enabled" to "disabled." This may also be done by interactive selection of a portion of the network.

Barriers may also be established to represent places where a pipe or a wire is broken. These temporary interruptions to the network allow GIS operators to determine where service interruptions will occur if a repair is required. Barriers can be established interactively, by pointing at the map, or by selecting existing features in a layer. Barriers act like switches or valves, in that they disable the part of the network that is downstream from the barrier, except that they are not permanent features of the network.

In most cases, these sources, sinks, and valves/switches provide enough information to determine the flow for each pipe or wire that is enabled. This creates a network that has "determinate flow." However, in complex networks where multiple paths exist, there may be locations where there is insufficient information to determine what is upstream and what is downstream. This leads to a situation called "indeterminate flow," which must be resolved interactively.

3.3 Geocoding

Often, tables of data must be entered into GIS. The process of assigning X, Y coordinates to features that have their locations expressed in other systems is termed Geocoding. There are three different types of Geocoding: XY Geocoding, Dynamic Segmentation, and Address Geocoding.

3.3.1 XY Geocoding

The easiest Geocoding technique to understand is XY Geocoding, which requires a table consisting of a location name, an X-coordinate, and a Y-coordinate to be available. Using this table, we can obtain the XY coordinates of any location name, simply by looking it up in the table and finding its coordinates.

XY Geocoding is performed most often when we are interested in approximate locations. Many postal agencies, such as those in Germany and the United Kingdom, have implemented a series of postal codes to help accelerate the delivery of mail. Each postal code corresponds to a particular area, usually only a few blocks in size, although large apartment buildings may be assigned their own code. When this code is written on a letter, it is used to determine which postal carrier should deliver the letter. The postal carrier can then examine the rest of the address to ensure that the letter is delivered to the right door.

Because postal codes define a distinct geographical area, it is easy to create a table showing the geographic centre of each postal code, and the associated X, Y coordinates that are assigned to it. We can then use these data to determine the approximate location where a person lives, based on the postal code at the end of their mailing address. Although the results are not very precise, they are sufficient for many demographic studies (Figure 32).

Postal Code	Latitude	Longitude
B2G2J3	45.6011	-62.0159
B2G2J4	45.60835	-61.9882
B2G2J5	45.6011	-62.0159
B2G2J7	45.6011	-62.0159
B2G2J8	45.61746	-62.0306
B2G2J9	45.61746	-62.0306
B2G2K1	45.61746	-62.0306
B2G2K2	45.61746	-62.0306
B2G2K3	45.6011	-62.0159
B2G2K4	45.61746	-62.0306

Figure 32. A small portion of the Postal Code Conversion file for the Canadian Province of Nova Scotia (negative longitudes indicate that these locations are in the Western Hemisphere).

A related type of XY Geocoding is intersection mapping Intersection mapping is commonly used in the mapping of vehicle accident locations, since these occur most commonly at intersections. In this case, the X and Y coordinates of each intersection are recorded, and the two cross roads can be matched to this location. This method of locating addresses can be combined with address geocoding to determine address locations when they are entered in a variety of formats.

3.3.2 Linear Referencing

In addition to X and Y Coordinates, locations may also be expressed as a distance along a linear feature, or as a street address. The process of Dynamic Segmentation is used to determine the location or linearly referenced data based on a linear feature and the distance from the beginning of that feature.

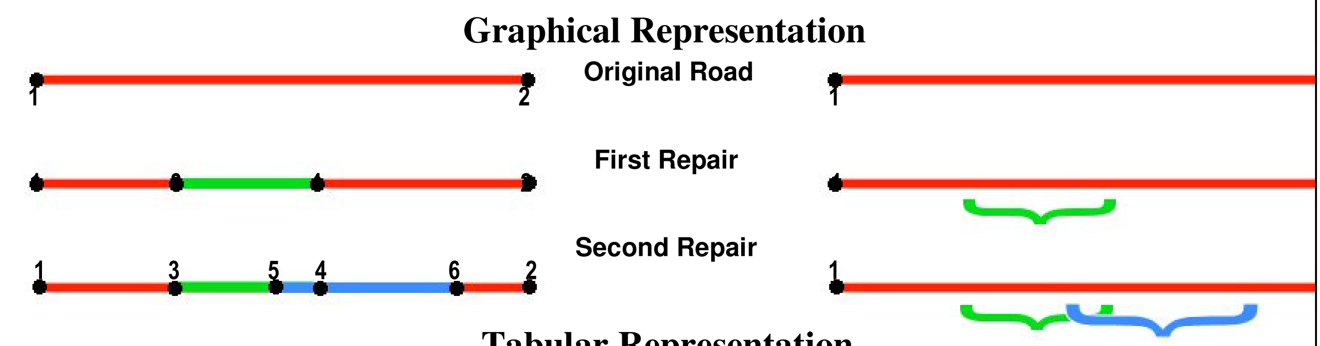
Dynamic Segmentation is different from XY Geocoding, in that it requires additional data. The X, Y Coordinates attached to the list of postal codes are complete; once you have matched the postal codes and have the X, Y Coordinates, you can simply plot these on a map. With Dynamic Segmentation, you must have a series of distances from a known starting point, and a network, along which those distances can be plotted. This is known as *linear referencing*. With linear referencing, distances from the start point may be measured in absolute terms (i.e. metres or kilometres), or in relative terms (i.e. as a percentage of the way from the start point to the end). The features to be located on the network, whether they are point or line features, are known as *events* (Chang, 2006).

Dynamic Segmentation is often used by transportation agencies, which locate objects along the road, such as signs, lights, and maintenance locations. A new sign, for example, could simply be located on Highway A1 at kilometre 143.29. This provides the location of the sign to with 10m along the length of the road, based on a known starting point. If a road has marked kilometre posts, then this system provides both a way to locate objects, and a way to record the location of areas in need of repair.

Road repair is the main reason that Dynamic Segmentation was developed. Consider what happens when a section of road is repaired. We must record that there is new pavement along a section of the road. Using traditional arc-node topology, we must divide the road up into three segments: the section before the repair, the section that has been repaired, and the section after the repair. Each segment would be assigned a different date for the "repaired" attribute. One year later, part of the repair has to be redone. Now we have to divide the road into additional segments, recording the fact that one location has been repaired twice. Very rapidly, we have a situation where the road is physically divided into many segments, and the attributes representing the status of the repairs are difficult to understand (Figure 33).

One important result of the way that Dynamic Segmentation stores its events is that multiple, overlapping attributes are possible. In Figure 34, we see how multiple variables stored with Dynamic Segmentation can be combined using a query.

Dynamic Segmentation can be used to create routes. A route is a connected series of segments that spans part of a network. Each segment has its own attributes and its own system of measurement; distance measurement on a route begins at the first segment that makes up the route. Because of the way that routes are created, multiple routes can overlap, each of which behaves independently of each other. Should, however, part of the underlying network change (for example if a road were rerouted slightly), all of the routes along the changed section of network would change as well.



Tabular Representation

Original Road

Arc ID	From-Node	To-Node	Type
1	1	2	Original

Segment	Start km	End km	Type
1	0.0	10.5	Original

First Repair

Arc ID	From-Node	To-Node	Type
2	1	3	Original
3	3	4	Repair 1
4	4	2	Original

Segment	Start km	End km	Type
1	0.0	10.5	Original
2	2.8	4.6	Repair 1

Second Repair

Arc ID	From-Node	To-Node	Type
5	1	3	Original
6	3	5	Repair 1
7	5	4	Repair 1, Repair 2

Segment	Start km	End km	Type
1	0.0	10.5	Original
2	2.8	4.6	Repair 1
3	4.2	8.5	Repair 2

Figure 33. A comparison of using Arc-Node topology to represent two overlapping repairs with the simpler structure afforded by Dynamic Segmentation.

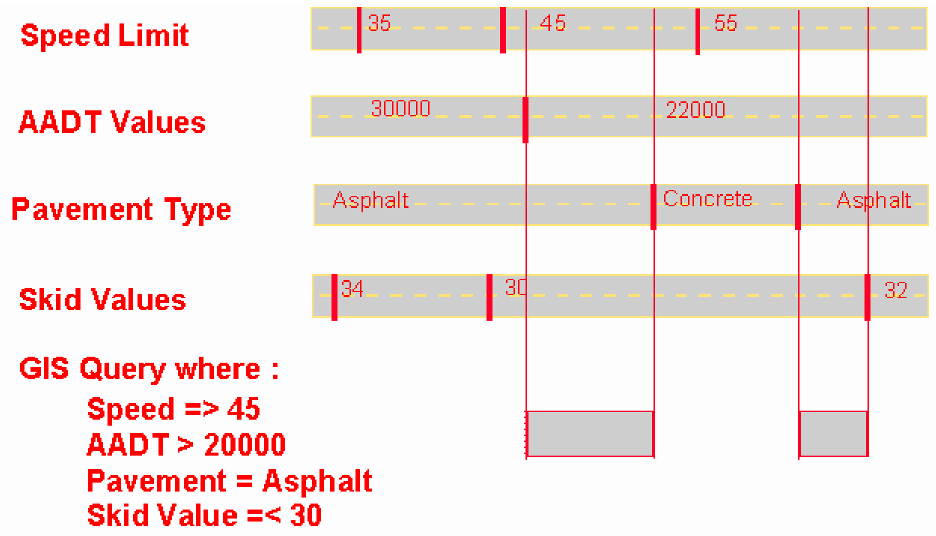


Figure 34. Multiple, overlapping attributes can be stored for a section of road using Dynamic Segmentation. Here, a query determines where a speed of 45 km/h is exceeded, the AADT (Annual Average Daily Traffic) is greater than 20,000 vehicles, the surface type as asphalt, and the skid value is less than or equal to 30.

3.3.3 Address Geocoding

Finally, Address Geocoding (also known as Address Matching) can be used to determine the location of a point on a road, based on the address of that feature. Address Geocoding requires the street name and number to be matched with an existing block on a map. Once this has been done, the street number is used to interpolate a position on the block. This is the technology used by services such as Google Maps and in-vehicle GPS Navigation Systems to guide drivers to their destinations.

Like Dynamic Segmentation, Address Geocoding can be used to locate objects based on their position along a road network. The requirements for the road layer to be used in Dynamic Segmentation are relatively simple; only distance values need to be encoded. For Address Geocoding, the informational requirements for the roads layer are much greater. At the least, each road segment must include attributes which show its name, the from block number and the to block number. More typically, the following attributes might be found:

- Street Name
- Left Starting Address
- Right Starting Address
- Left Ending Address
- Right Ending Address
- Even Side of Road
- Odd Side of Road
- Postal Code

Address Geocoding assumes that each road segment has a range of addresses defined for left and right sides (Figure 35). First, the street on which a property is located is determined; this allows the GIS to shortlist those segments on which the property may exist. Next, the numeric address of the property is used to determine whether the property in question is on the left or the right hand side of the street. When a polygon needs to be identified, an offset value can be defined so that the point that is geocoded will be offset from the centre of the road. This allows the GIS to use a point-in-polygon calculation to locate the polygon, which would not be possible if the point were exactly in the centre of the road, separating two polygons (Figure 36). Finally, the address number is used to interpolate the position on a particular street segment between the recorded start and end positions.

Attributes of roads_adr_final

FULL_NAME	CITYL	CITYR	METERS	FROM_EVEN	TO_EVEN	FROM_ODD	TO_ODD	EVEN_SIDE	ODD_SIDE
A. V. Kojelavièiaus g.	Vilniaus m.	Vilniaus m.	50.14669	108	114	83	83	R	L
A. V. Kojelavièiaus g.	Vilniaus m.	Vilniaus m.	129.68456	92	100	45	69	R	L
A. V. Kojelavièiaus g.	Vilniaus m.	Vilniaus m.	176.01752	68	88	39	47	R	L
A. V. Kojelavièiaus g.	Vilniaus m.	Vilniaus m.	68.62041	276	280	0	0	L	R
A. V. Kojelavièiaus g.	Vilniaus m.	Vilniaus m.	66.88754	14	18	1	1	R	L
A. V. Kojelavièiaus g.	Vilniaus m.	Vilniaus m.	136.27267	222	230	0	0	R	L
A. V. Kojelavièiaus g.	Vilniaus m.	Vilniaus m.	21.8107	100	102	67	71	R	L
A. V. Kojelavièiaus g.	Vilniaus m.	Vilniaus m.	242.84067	20	28	3	3	R	L
A. V. Kojelavièiaus g.	Vilniaus m.	Vilniaus m.	185.2287	236	248	0	0	L	R

Record: 0 Show: All Selected Records (0 out of 19125 Selected) Options

Figure 35. Even and Odd Address Ranges for A. V. Kojelavièiaus g. in Vilnius. These allow the position of a building with a particular address to be estimated along a street segment.

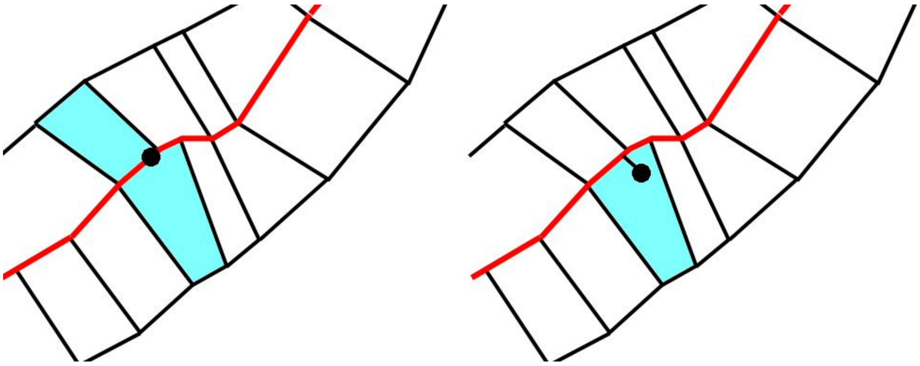


Figure 36. Geocoding based on just the street address can lead to an ambiguous situation in which two polygons may be chosen (left). Using a small offset value leads to the added point being placed unambiguously in a single polygon (right).

In order to match an address to a particular road, we must set up an address locator to determine the position on the road based on the address format. Because different countries have different address formats, and many countries have multiple address formats, we must make use of one or more address locators for addresses in each country (Figure 37). If a particular format is not available, then custom Address Locators can be created. In an operational geocoding system, a number of different address locators will be attempted. Those addresses that are not matched by the first address locator are passed onto the second, and so on. The reason for this is that there are often inconsistencies and errors in both road networks and addresses. If necessary, the user can adjust the closeness of the matching, so that less than exact matches can be used.

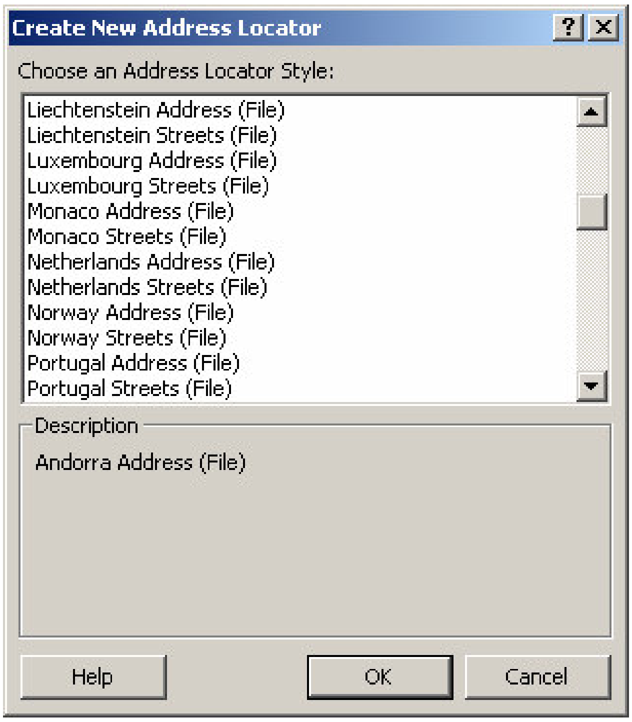


Figure 37. Address Locator set up for different countries

Problems with addresses may include half street addresses (e.g. 18½ Main Street), different ways of spelling the same street type (Avenue vs. Ave. vs. Ave), road names that have been misspelled, road types that are not recognized by the address locator, incorrect road numbers or street names, incorrect street direction (e.g. Main Street N.) and missing information. Problems with the road layer may include new roads that have not been added, roads that have been closed, incorrect block range values, misspelled road names, road names that have been changed, and missing information.

Adopting forced entry fields in the data entry software can prevent some of the data entry errors. For example, we could present a list of properly spelled street names, so that the street name cannot be misspelled (Chang, 2006).

Once an Address Locator has been defined, it simply needs to be provided a table containing addresses in order to geocode it (Figure 38). The table will be used to generate a new point feature class containing the features that have been located (Figure 39), which can then be place on a map (Figure 40). If some addresses cannot be geocoded, then there is an option to manually locate the road segments on which the features occur.

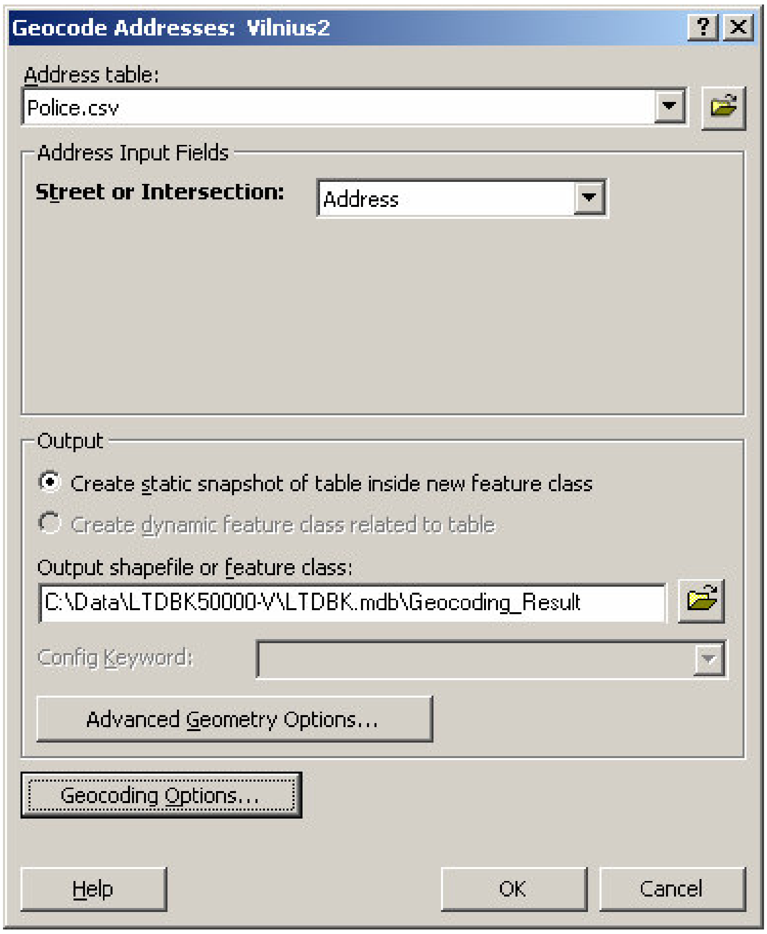


Figure 38. Making use of an Address Locator (Vilnius2) to geocode the addresses in a database table.

Attributes of Geocoding Result: Geocoding_Result_American

FID	Shape *	Status	Score	Side	X	Y	Match_addr	ARC_Street	Station	Address	Telephone	Fax
0	Point	M	100	L	582124.81493	6060988.45358	14 MINDAUGO G	14 Mindaugo g.	Vilnius Police Commissariat (head office)	14 Mindaugo g.	27 17 158	27 16 080
1	Point	M	88	R	582630.459804	6063051.129377	59 KALVARIJØ G	59 Kalvariju g.	Vilnius Police Commissariat No. 1	59 Kalvariju g.	27 17 600	27 17 701
2	Point	M	100	L	581982.849597	6060803.004027	20 ALGIRDO G	20 Algirdo g.	Vilnius Police Commissariat No. 2	20 Algirdo g.	27 16 7 06	27 16 716
3	Point	M	100	L	582812.146239	6060605.407101	52 PYLIMO G	52 Pylimo g.	Vilnius Police Commissariat No. 3	52 Pylimo g.	27 16 078	27 16 2 92
4	Point	M	85	R	591723.878296	6062502.088441	29 GEROVĖS G	29 Geroves g.	Vilnius Police Commissariat No. 4	29 Geroves g.	27 15 014	26 71 638
5	Point	T	100	L	594727.95495	6079832.77595	8 T KOSCIUĐKOS G	8 T. Kosciuđkos g.	Vilnius Police Commissariat No. 5	8 T. Kosciuđkos g.	26 25 752	27 18 121
6	Point	T	87	L	580177.147824	6064989.237185	30 ĐEĐKINĖS G	30 Đeđkines g.	Vilnius Police Commissariat No. 6	30 Đeđkines g.	24 16 766	24 16 766
7	Point	M	90	L	578418.901478	6063821.2446	14 JUSTINIĐKIØ G	14A Justiniđkiu g.	Vilnius Police Commissariat No. 7	14A Justiniđkiu g.	27 16 4 81	27 16 454

Record: 1 Show: All Selected Records (0 out of 8 Selected) Options

Figure 39. Geocoded table of addresses for Police Stations, with X and Y coordinates added by the Address Locator.

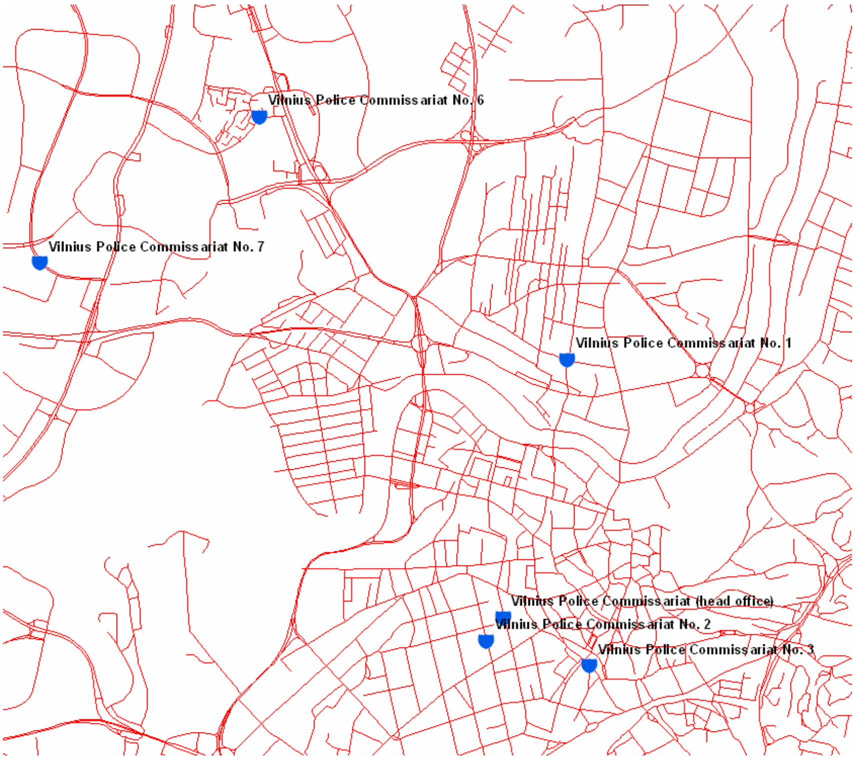


Figure 40. Some of the geocoded Police Stations in central Vilnius

Address Geocoding presents a fast and effective way to create new point data in a GIS based on database tables that have address information within them. It is important, however, to understand the limitations of this technique. There is an implicit assumption that all properties on a particular block are the same size. On blocks that have different size properties, the actual location may be slightly different than what geocoding predicts. McCarthy and Ratcliffe (2005) found that geocoded addresses were, on average, positioned 31 metres from the actual locations.

When an address needs to be located approximately, then address geocoding is an appropriate tool. For example, an emergency call operator can type in the address of the person making an emergency call, and an ambulance can be dispatched to the geocoded address. Although the ambulance will not arrive at the exact address, the few extra seconds that they take to find the correct address will be more than compensated for by the reduced travel time from the ambulance station allowed by the calculation of the shortest route.

When addresses need to be located exactly, a database of address locations must be created beforehand. In such a case, the locations should be determined using more accurate techniques, such as differential GPS or surveying. Needless to say, this is a much more expensive proposition than simply using Address Geocoding.

The accuracy of address geocoding is dependent in part on the accuracy of the base map. Streets must be mapped accurately, and the address ranges must be assigned correctly. Unfortunately, assigning address ranges is a large and tedious job, and is prone to error. Furthermore, although the vast majority of addresses can be located this way, there are always exceptions. In Toronto, Canada, an emergency crew was recently dispatched to the wrong location because there were two roads with the same name, as a result of the amalgamation of a number of suburbs into the city of Toronto, Canada (CityNews, 2006).

3.4 Calculation of Distance

One of the issues with printing maps on sheets of paper is that it encourages us to measure distances in straight lines, and on flat surfaces. Unfortunately, travel in the real world is almost never straight, in either the X-Y plane, or in the Z plane. The reality is that virtually any journey across the surface of the Earth involves a path that curves in three dimensions.

Highly powered vehicles such as automobiles tend to mask the effort involved in travelling over irregular terrain. Just compare a trip over rolling hills in a car with a trip on a bicycle! The driver of the vehicle may not even be aware of the changes in elevation, but the bicycle rider most certainly will. Nevertheless, energy is still expended, whether it is in the form of gasoline or muscle power.

The simplest form of distance measurement is the straight-line, or Euclidean distance. On large-scale maps, it is easy to draw a line connecting two points. While this may represent the shortest distance to travel, it does not necessarily represent the shortest travel time. A more sinuous route may take less time to travel, particularly if the straight-line path forces us to cross mountains! Over short distances, the map sheet is a good approximation of the Earth's surface, since the amount of the Earth's curvature over such distances is negligible. This is why jet aircraft tend to travel straight lines. Of course, over longer distances, we need to consider the curvature of the Earth. If it were possible, the shortest path between London and Vilnius would not be the great circle route that a jet aircraft would fly, which is the shortest path between two points on the surface of a sphere, but would be a straight line that burrowed through the Earth's crust!

For most travel, then, Euclidean distance isn't a good approximation. If we ignore the curvature of the Earth for short trips, then we are left with two possibilities: Manhattan Distance or Network distance. Manhattan distance is simply the distance travelled on a regular grid of roads (such as those found in Manhattan, New York, USA). As we saw in the previous module, the formula for

$$d_{ij} = \left|X_i - X_j\right| + \left|Y_i - Y_j\right|$$

Manhattan Distance is simply

In other words, you have to move the difference between coordinates in one direction and then the difference in the other. The greatest advantage of Manhattan distance is its simplicity; the only thing that we need to know is that a regular grid of roads is involved, and we can calculate the distance. However, this system only works when roads are laid out using the British System, where roads and avenues meet at 90-degree angles.

In most parts of the world, Manhattan distance does not work, because roads are sinuous (for defensive purposes, because they follow old routes, or because they follow a river or coastline), or radiate outward from important locations. In cases such as this, Network Distance should be used, because it allows the exact distance along the road to be measured. Unfortunately, determining Network Distance requires a transportation network to be built, and determining the shortest path between two points can be computationally intensive.

The Euclidean Distance formula can be generalized to create the Manhattan Distance formula, and it can also be generalized to create other distance metrics using the General Distance formula (McGrew and Monroe, 2000).

$$d_{ij} = ((X_i - X_j)^k + (Y_i - Y_j)^k)^{1/k}$$

In this formula, the k exponent allows us to model the difficulty in traveling. The k value must always be lower than 2, since if *k* is 2, we end up with the Euclidean Distance (and your distance cannot be shorter than a straight line). When *k* is 1, we end up with the Manhattan Distance formula. So the lower the value of *k*, the most difficult it is to travel between two points. A very low *k* value of 0.6 might be used to model the movement of a boat traveling along a meandering river. We could determine the average distance between two points, and use it to create a value of *k* for any city. The advantage of using this method is that we don't need to have a constructed road network; given the X and Y Coordinates of any two points, we can estimate the distance along a road between the two, since our *k* value provides an estimate of how indirect the route is likely to be.

Network Distance is more difficult to compute, but easier to understand. If we have two points and a network, we can plot a route from one point to the other by following the network. Although many different paths can be followed in a busy urban area, there is generally a shortest path between the two points, which can be determined through a series of software iterations. We will discuss the technique for determining the shortest path distance in the next section.

3.5 Routing

Vector data structures can be used to efficiently represent road networks. The main reason for this is that the discrete nature of lines in a vector data structure ensures that travel off of roads is not allowed. Using Graph Theory, a network of vector lines can be topologically connected, and can be assigned attributes that allow the simulation of vehicle travel along road networks to be made.

The Dijkstra (1959) algorithm is the most commonly used method for determining the shortest path between points. The algorithm builds up a table of all impedances by working outward from the two nodes, and identifying those arcs with the least impedance first. The nodes that are connected by these arcs are then evaluated to determine the arcs with the least impedance that flow from them. If a circuitous route is found that has less total impedance than a direct route, then the direct route is replaced in the table. This process continues until all possible routes between the start and end points have been evaluated. When this process is complete, the route with the least impedance between the two points has been identified, and the total impedance cost has been calculated.

Of course, not all routing problems are as simple as travelling between two points. There are two common variants of routing problems. The first is that we have a single point, and we need to determine which of *n* possible destinations is closest. The other form of the problem is to solve the *traveling salesman problem*, in which, given a certain number of destination points, the shortest path that visits each stop only once must be determined (Chang, 2006).

The solution to the first problem is really quite easy, once Dijkstra's algorithm for determining the shortest path has been implemented. We simply calculate the shortest path from the start to each of the end nodes, and determine which has the shortest route (Chang, 2006).

The second problem is much more complex. Consider that for n possible stops, there are 2^n possible routes. Determining the optimal route from each node to each other must then be calculated using the Dijkstra algorithm. Lin (1965) created an algorithm that produces locally optimal solutions by swapping the stops until the shortest path is determined, but this problem has yet to be solved for large numbers of stops (Chang, 2006).

3.6 Functional Distance

When travel is not limited to a network of roads, or the network has not yet been physically constructed, a Raster data structure provides better modeling than a network. Raster modeling does not limit travel to a network, instead allowing for travel in any of eight directions from one cell to the next. This allows for more flexible paths to be created, although the modeling of such travel is more computationally intensive than the modeling of vehicle travel in a Vector data structure.

Suppose that we are traveling on foot. Obviously, the dynamics of traveling on foot are different from traveling in a vehicle. For one, it is much more important to travel in a straight line, since your speed is limited. When you combine this with the fact that a pedestrian is not necessarily limited to travel along the roads, you end up with a different type of travel being modelled. In such a system, we need to reflect the fact that it is probably easier to walk along a road, but that significant savings in travel distance may be worth taking shortcuts. These differences mean that we need to be able to model both on and off-road travel.

There are a number of ways of modeling functional distance on a raster, some of which are quite sophisticated. At their simplest, these methods require a source raster (a raster representing the location from which travel begins), and a cost raster. The cost raster may simply result from calculating the cost to travel through certain types of land cover, or it may be a sophisticated model in its own right, combining many different layers to create an expression of travel cost. Also, the cost raster may represent actual travel costs, or merely standardized relative travel costs between cells (Chang, 2006).

Another situation where functional distance calculations are used is when we are building a new road between two points. Since the road has not yet been constructed, it could follow almost any path between the two points. In virtually all cases (except for specially-designed Scenic Drives), we are trying to minimize the cost of construction between the source and destination. Functional distance commands can help us to determine the least cost route between these two points.

An interesting side effect of this type of analysis is that the road which costs the least to construct, because it minimizes changes in topography, it is likely also to be the least cost route to travel once the road is completed. In other words, the savings that result from Functional Distance calculations not only reduce the cost of road construction, but they also provide ongoing fuel savings for vehicles once it is completed.

As we discussed in section 3.4, the simplest way to conceptualize distance is to use Euclidean Distance. This can be done easily on a raster. On a raster, there are eight possible directions: North, Northeast, East, Southeast, South, Southwest, West, and Northwest. We can follow a path from the centre of one pixel to the centre of any adjacent pixel. Because of this, the distances in the cardinal directions (North, South, East, and West) are 1/2 pixels, whereas the distances in the diagonal directions (Northeast, Southeast, Southwest, and Northwest) are $\sqrt{2}/2$ or 0.7071 pixels (Figure 41).

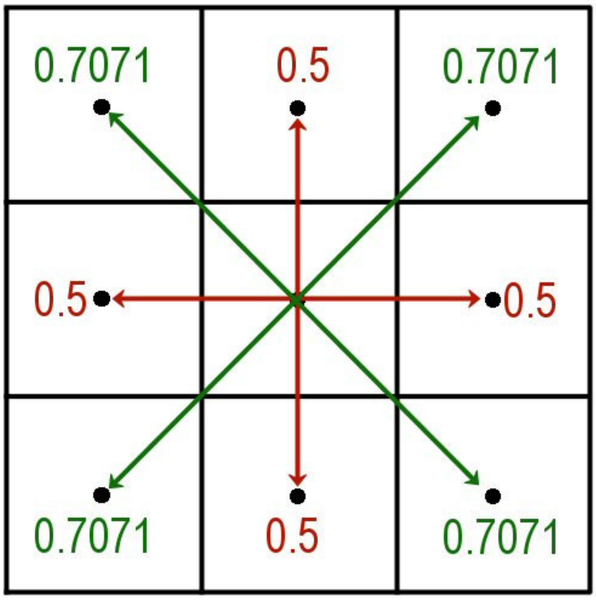


Figure 41. Eight pixels surround any particular pixel. Distances between centres are 0.5 in the cardinal directions (red) and 0.7071 in the diagonal directions (green).

Based on these pixel distances, it is relatively easy to determine the distance of any pixel from any other. The pixels can then be colour coded by their distance from the central pixel (Figure 42).

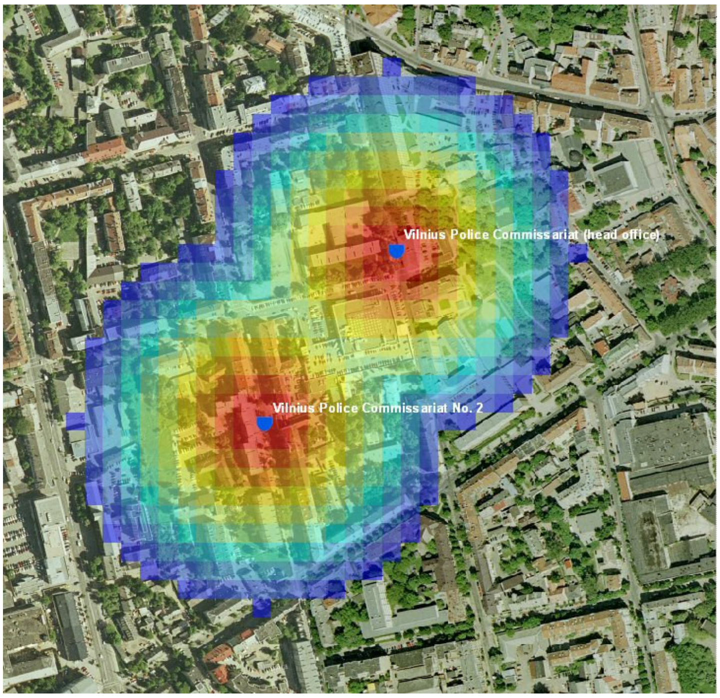


Figure 42. Euclidean Distance calculations from two Police Stations. The shade is dependent on the distance value stored in each pixel. Maximum distance is 200 metres.

A more complex way of looking at distance is by considering the difficulty of travelling over the surface. Like the Euclidean Distance calculations, Cost Distance functions assume that there is no elevation change or wind blowing. They, do, however, incorporate a cost surface, which represents the amount of "friction" involved in crossing the surface. Like the impedance value that is used to determine the cost of traversing a road segment in a network, the cost surface represents the difficulty of traversing each cell. Cost distance calculations are also similar to shortest path calculations on a network, in that they both assume that the difficulty of climbing hills is negligible. Put another way, they both work well to model the travel of powered vehicles between two points.

The process of providing reliable estimates for a cost surface is difficult, just as it is difficult to provide estimates for link and turn impedance values in a network. In both cases, we often make use of estimates and averages, because it is difficult to obtain accurate information about off-road travel times. The most common process to determine a cost surface is to reclassify a land cover raster to reflect the difficulty of travelling through different types of land cover. For extensive studies of overland travel, it might be sensible to perform a study to determine travel speeds over different types of terrain beforehand. This is difficult to do when we are dealing with off-road vehicles, and even more difficult when we are trying to model the movement of animals across terrain -- radio collar tracking might be the only way of accurately determining animal travel speeds over different types of terrain (DeMers, 2005).

The algorithm for the Cost Distance calculations combines the cost raster with the distance between cells. We can examine every pixel in the cost raster, and compute the cost of travel to all surrounding pixels. If we assume that the cost for the central pixel is C_i, and the cost for the adjoining pixel is C_j, then the cost of travelling from C_i to C_j is:

$0.5\ (C_i + C_j)$

for travel in the cardinal directions, and

$0.7071\ (C_i + C_j)$

for travel in the diagonal directions.

Based on this information, we can determine the cost of travelling from a start point to all adjoining pixels, based on the cost raster. We can stop the calculations after a particular cost has meant exceeded, or we can allow the costs to accumulate until the edge of the raster is reached (Chang, 2006).

The third, and most sophisticated functional distance calculation is the Path Distance calculation, which incorporates the effects of topography and winds on travel (Figure 43). To accomplish this, two additional raster inputs are required: a Digital Elevation Model (a raster with elevation values recorded for each pixel), and a raster representing the direction of winds.

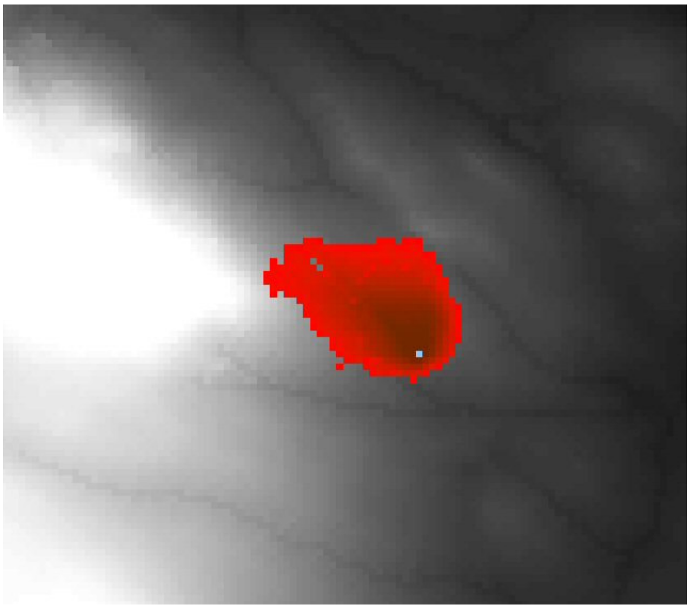


Figure 43. Model created using Path Distance command showing predicted travel of a forest fire from a source (light blue pixel), based on topography (underlying shaded raster) and a wind from the Southeast.

The path distance calculation combines the distance between cells with multipliers based on the difficulty of travelling up (or down) hill, the strength and direction of the wind, and the cost raster. The effects of winds can be set to zero in this command to model travel by off-road vehicles, by foot, or by wildlife.

To calculate the cost for travelling to an adjacent cell using Path Distance, we must first determine horizontal and vertical factors, which are the weights applied to travel from one cell to the next based on the wind direction and velocity and the difference in elevation. To come up with the horizontal and vertical factors, a graph is applied to the wind and elevation information.

For example, we might use a linear graph to convert the difference in elevation into a vertical factor. The linear graph implies that as the angular increase in elevation becomes greater, the difficulty of moving from one cell to the next increases as well. At a certain point, the difference becomes so great that movement from one cell to the next is impossible. In this case, the vertical factor becomes infinite. If travel is downhill from one cell to the next, then travel becomes increasingly easy, until at some point, it becomes too steep to travel downhill, so once again, the vertical factor becomes infinite (Figure 44).

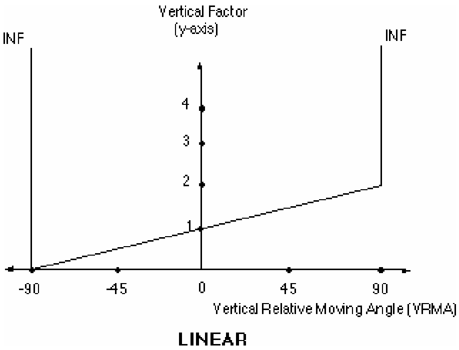


Figure 44. Linear graph for the conversion an angle difference into a vertical factor. Travel at all angles except for straight up and straight down is permitted (Environmental Systems Research Institute, 2007).

The horizontal factor is also calculated using a graph. Again, the linear graph shows an increase in the amount of "push" as the direction of travel aligns with the wind direction. At some point, the strength of the wind prevents travel into the wind. In the example shown (Figure 45), the wind isn't very strong, so only travel directly into the wind is impossible; if the wind were stronger, we could move the point at which the Horizontal Factor becomes infinite to a lower angle (Environmental Systems Research Institute, 2007).

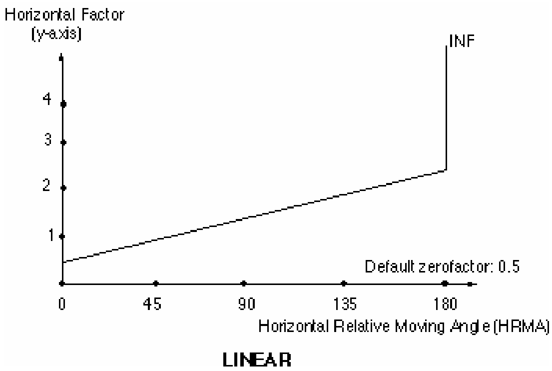


Figure 45. Linear graph for the conversion of horizontal angles (relative to the wind direction) to a horizontal factor. Only at 180 degrees (facing directly into the wind) is travel impossible, which is indicated by the infinite horizontal factor (Environmental Systems Research Institute, 2007).

The formula for incorporating the horizontal and vertical factors, as well as the cost surface is:

$D * V * (C_i * H_i + C_j * H_j) /2)$

Where:
D is the planimetric distance
V is the vertical factor (the difficulty of moving up or downhill)
C_i is the cost of the central pixel
H_i is the horizontal factor of the central pixel
C_j is the cost of the neighbouring pixel
H_j is the horizontal factor of the neighbouring pixel

By combining functional distance calculations for a source and a destination, it is possible to determine the cost of travelling from source to destination across all possible paths between the two points. Given a maximum cost (amount of fuel, construction cost, or energy that can be expended by an animal), we can highlight a corridor of travel between the source and destination

(Figure 46). Within this corridor, the energy or fuel budget is not exceeded, and outside the corridor, travel is not possible because too much energy or fuel is used. Of course, in the centre of this corridor is a single line representing the line of least resistance between source and destination (Environmental Systems Research Institute, 2007).

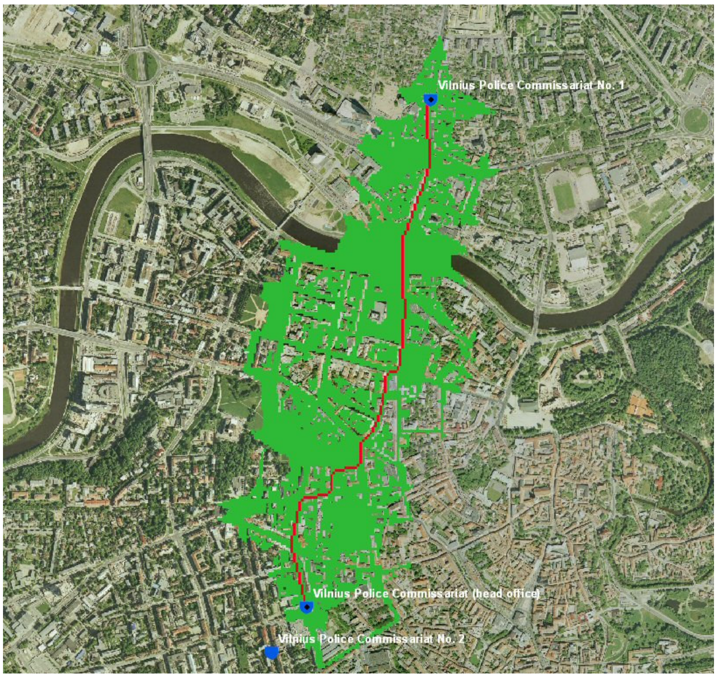


Figure 46. Corridor Analysis showing all paths that cost less than 15,000 (arbitrary) units, and the path of least cost through the centre between Vilnius Police Commissariat No. 1 and the Police Headquarters.

Now that we have discussed methods for analyzing travel on a network and off the network, we can examine the transportation of commodities using networks, which is how many goods are delivered in our society.

3.7 Utility Network Analysis

With a fully functional utility network, where there is no indeterminate flow, we can analyze network flow by *tracing* the network. Tracing can be used to determine the areas that are upstream or downstream from a particular point (Figure 47). Some other tasks that can be performed by tracing the network include:

- Finding Common Ancestors (i.e. finding all segments that lie upstream and are common to a series of locations)
- Finding all parts of the network that are disconnected from a particular segment
- Finding all parts of the network that are connected to a particular segment
- Finding loops (cycles) in the network
- Finding the path between two points
- Finding the amount of accumulation upstream (useful for river networks) (Zeiler, 2005)

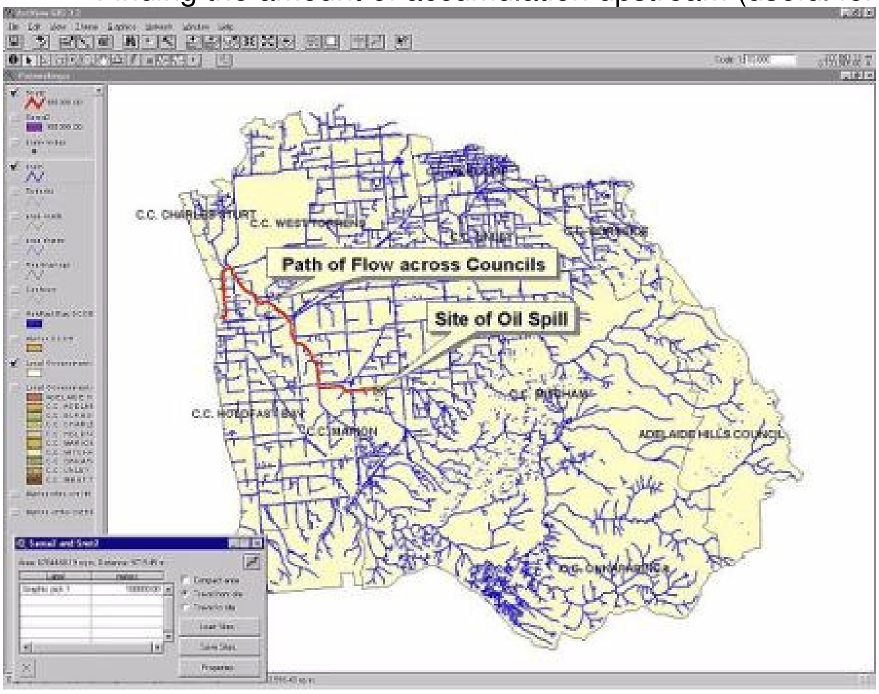


Figure 47. Tracing the path of oil contamination downstream from an oil spill in Adelaide, Australia (Overton, 2007)

3.8 Applications

3.8.1 Dynamic Segmentation

In natural resources applications, researchers often want to record the location along a river where a particular observation was made, or where a particular tributary begins. This can be done by recording the linear distance up the river network from the point of launch. Furthermore, the ability of Dynamic Segmentation to support multiple overlapping data sets allows multiple attributes such as pH and salinity to be stored for river networks.

The ability of Dynamic Segmentation to support routes makes it an ideal tool for transportation planning. For example, multiple train routes can be mapped on the same network, and many bus routes can share the same street (Figure 48). Even transportation routes that don't specifically follow exactly the same path every time, such air airline and marine travel routes, can be modelled using this technique, although it is necessary to assume that planes and ships will follow the path laid out by the network.

For those agencies that are charged with the maintenance of infrastructure, Dynamic Segmentation provides a method of locating repairs that need to be made, and keeping track of those repairs that have been completed. This allows a complete record of infrastructure maintenance to be built up over time.

Indeed, any feature that is linear in nature can be modelled using Dynamic Segmentation. For example, the amount of pollution and types of ecosystem that are found along a shoreline can be mapped.

Figure 48. Multiple rail routes can share the same segments of track. Dynamic Segmentation can be used to connect the same segments into different routes.

3.8.2 Geocoding

Address Geocoding can be used for much more than just locating addresses for the dispatch of emergency services. In a Multipurpose Cadastre, Geocoding can be very helpful in locating parcels based on their address -- people always know their address, but they can rarely give the latitude and longitude of the location where they live!

The mapping of locations where crimes occur is one of the most important applications for Address Geocoding. Once again, when a crime is reported, the caller tends to give the street address where the crime occurred. That address can then be converted into a point on a map using Address Geocoding. Over time, the accumulation of points begins to give policing agencies information on the distribution of crimes. Clusters of similar crimes can be mapped, and these can be used to create a geographic profile of a criminal.

On the less serious side of things, geocoding can be used on social networking and dating websites to identify people of similar interests who live near you. For fundraising agencies, such profiling tools can be very lucrative as well. For example, if the addresses of people who donate to a charity can be located, clusters of people who are sympathetic to a particular charity's cause can be identified. The charity can then target its fundraising activities more productively, by aiming at those people who are most likely to donate (Chang, 2006).

An example of using Address Geocoding in health care is provided by Cromley and McLafferty (2002). The entire health care system can be mapped to identify locations of over and undersupply of services. If the locations of facilities, services provided, and the hours of operation are provided to referral operators in a GIS application, then the operators can refer callers to the most appropriate facility at different times of the day, based on the address of the caller. This application locates the caller based on their address, calculates distance to the nearest open facility using the Shortest Path Algorithm, and can even provide a list of directions for the operator to relay to the caller.

Of course, once this level of sophistication is developed, it is not difficult to automate the service and make it available on a web page. Users simply type in their address, and the system provides the nearest clinic or hospital that is open when the request for data is made.

3.8.3 Routing

Shortest-path calculations are the basis for In-Vehicle Navigation Systems such as the Hertz Rent-a-Car NeverLost system. These systems combine a GPS receiver with a specialized GIS application that uses a road network and special software that allows shortest-path calculations to be performed. Combined with a utility to create a list of directions and a voice synthesizer, this tool allows you to enter your destination. The GPS receiver determines your current location, and the shortest route is then calculated from your location to your destination. A list of direction is generated, and then the voice synthesizer reads off these directions step by step to get you to your destination.

GIS applications that determine the closest facility are very important for telephone referral services. When a caller phones to inquire about the closest school, the operator can enter the address of the caller, and this can be used to determine the closest facility. If call display services are available, the system can go one step further, by matching the phone number to the caller's address, then using address geocoding to determine the X and Y coordinate of their address, then determining the closest facility.

For companies that perform regular deliveries, such as dairies or newspaper publishers, solving the travelling salesman problem can be the key to reducing delivery costs, and thus, increasing competitiveness. For disabled persons and the elderly, the travelling salesman algorithm can be used to assist with the provision of "meals on wheels" services. The addresses of all the people who require the service can be entered into a GIS, and once they have been Address Geocoded, a route can be generated to allow each location to be serviced effectively (Cromley & McLafferty, 2002).

3.8.4 Functional Distance

Although the Euclidean Distance calculation produces rasters that are poor models of real-travel, the Cost Distance and Path Distance commands produce much more realistic results. Since the Cost Distance command ignores the effects of terrain and wind on travel, it is a good model in situations where it is safe to ignore the energy expenditure required for travel. Thus, for travel by off-road vehicle, the Cost Distance model works reasonably well. Of course, for certain types of analysis, such as determining fuel costs for off-road travel, the Cost Distance model is inappropriate, and the Path Distance model produces more valuable results. In situations where the effects of terrain and wind need to be modelled accurately, such as the distribution of plant seeds, the movement of forest fires, the modeling of gas or air pollution dispersion or travel by bicycle, the Path Distance command is most appropriate (Chang, 2006).

Combining the results of two Cost Distance or Path Distance analyses allow us to model travel corridors between the source and destination points. This is useful in mapping the routes for overland travel, wildlife travel, or for the construction of new infrastructure, such as roads, pipelines, or electrical transmission lines.

3.8.5 Utility Networks

Utility Networks can be analyzed to help manage the flow of commodities to customers effectively. For example, being able to trace the network gives us the ability to use reports of service outages to isolate the location of breaks in a network, and thus assist repair crews in locating the break in the network (Figure 49). The ability to determine the portions of the network that are downstream or upstream of sampling locations can help us to isolate the source of contamination in the network. We can also determine the capacity of the network to transport commodities, and can identify locations that require repairs or future upgrades to improve capacity as a result (Zeiler, 2005).

Figure 49. Tracing a utility network can help to make sense of issues of distribution and flow, particularly when it is very complex.

3.9 Examples

In Prince William County, Virginia, the county government is using Address Geocoding to map the locations of incidents that affect public safety, such as burglaries and traffic accidents. This gives the government a handle on areas that pose a significant risk to public well-being and allows environmental crime deterrence, policing, and changes to roads and intersections to be made to reduce the risk in these areas. In addition, areas where social and community services are required the most can be identified, which gives the county the ability to provide services where they are needed the most. Dynamic Segmentation is used in the county to assist with transportation planning and school bus routing (Korte, 2001).

The Baton Rouge Healthy Start GIS in Baton Rouge, Louisiana uses Address Geocoding to map the homes of women who visit perinatal clinics. This allows areas that are poorly served by perinatal clinics to be identified, and for health care workers to be able to help women who live in underserved areas. After their children have been born, the women who visited the clinics are contacted to ensure that their children are doing well; where children have not survived, the system helps to generate maps showing the density of infant mortality as a tool to help reduce future infant mortality rates (Curtis & Leitner, 2006).

The Vermont [USA] Agency of Transportation (VTrans) has replaced its entire system of paper Route Logs with a Dynamic Segmentation system, to simplify the management of roads in the state. In the past, all roads were mapped using linear Route Logs on paper; these paper records contained information on pavement conditions and history, traffic volumes, roadway widths, speed zones, as well as the locations of infrastructure including bridges, curves, crashes, intersections, traffic signals, and railway crossings. Unfortunately, due to staff cutbacks, it became difficult to maintain the paper Route Logs. In response, VTrans implemented a dynamic segmentation application over an 18-month period to store all of this information. The new system enables Route Logs to be viewed over the Internet using a simple application; more sophisticated analysis can be performed on a dedicated workstation. The next phase of the project will add many secondary capabilities, including the ability to view video logs of movement along the roads, orthophotograph overlays, links to CAD data and scanned documents, and DEMs to allow the production of route profiles from the road data (Kiel, 2004).

In Texas, USA, CenterPoint Energy is a large, integrated energy production, transmission, and distribution company. It provides natural gas and electricity to 5 million customers in seven states. The company uses network analysis extensively to maintain its distribution networks. For example, the company can use network analysis to trace its electrical distribution networks and identify outage locations, and to assist with the connection of new customers. A special gas outage analysis application allows the utility to analyze its distribution system for weaknesses and to identify locations of breaks in pipelines. By solving the travelling salesman problem, it is able to define efficient routes for its meter readers as they collect information about each customer's natural gas and electrical usage (ArcUser, 2003).

3.10 Conclusion

The analysis of travel, both on and off-road, is a key capability of GIS. To be able to perform this analysis, we must have appropriately structured raster or vector data. For vector data, the construction of a network is a key step, which allows many different types of analyses to be performed. Another key step is the location of objects that are stored using a non-Cartesian coordinate system using Geocoding. Geocoding allows data that are stored as street addresses, postal codes, or distances along a linear route to be converted into an X, Y coordinate system, which can be placed on a map.

We can use raster data to analyze cross-country travel over any surface. The simplest way of performing this analysis is to calculate Euclidean distance, but more sophisticated techniques are available which consider the amount of "friction" encountered while travelling, as well as elevation differences and the effects of wind. By combining multiple distance analyses, it is possible to generate travel corridors and optimal paths between points.

A network permits the efficient analysis of travel between points. Of course, the key assumption is that travel occurs only along the network, and no travel is permitted outside the network. The advantage of making this assumption (which is correct for many different types of travel) is that analysis on a network is highly efficient, and permits the determination of the shortest route, optimal routes for visiting a certain number of stops, and the automated creation of a list of directions.

There are two types of network analysis: transportation network analysis, and utility network analysis. In a transportation network, vehicles are autonomous and can move in different directions along the network, assuming this is permitted. In a utility network, travel is always limited to a single direction, and commodities move from sources to sinks either through the flow of gravity or because of pressurized or electrified networks.

Together, the ability to locate people or objects through geocoding, and the ability to determine optimal routes from source to destination are a powerful combination, and can be used to make the provision of government services more effective and less costly. Whether this is routing an ambulance from its current location to the location of a car accident, or assisting a delivery truck driver in making as many deliveries as possible in the shortest amount of time using the least fuel, the analysis of travel can help almost everyone in one way or another.

Module Self-Study Questions

1. Virtually all trips across the surface of the Earth involve a curving path in three dimensions. Just to understand the issue better, can you think of any situations where land vehicles travel in virtually straight lines?

Answer: Some that I can think of are aircraft landing on a runway, vehicles attempting to set speed records on dry lake beds, and trains travelling in exceptionally flat areas such as Ukraine.

2. Explain how a river having many dams, such as the Dnieper River in Ukraine, can be modelled as a utility network. What form of network does a river take?

Answer: Each tributary can be defined as a source for water; each dam acts as a valve, and the outlet of the River can be treated as a sink. If pumped storage is used (pumping water back into a reservoir when energy prices are low, so that generators can produce additional electricity when energy prices are high), then we have loops in the network, which form a circuit, but otherwise the river acts as a branched network.

3. Can you think of any unusual addresses that would give an Address Locator particular problems?

Answer: There are many examples. Frontage roads may have the same name as a main road, but run parallel to the main road and allow for access to businesses. Half addresses are an issue. Roads that have the same name as cities (i.e. Vilnaius gatve) can sometimes cause problems.

Required Readings

- Chang, K-Ts. (2006). Introduction to Geographic Information Systems, 3rd Ed. New York: McGraw-Hill

ESRI Virtual Campus Module

- Introduction to Urban and Regional Planning Using ArcGIS 9 Module 2: Data for Urban and Regional Planning

Assignments

- Assignment 2: Emergency Service Levels

References

ArcUser Online, 2003. "Real Benefits from Interoperability" (http://www.esri.com/news/arcuser/0403/centerpoint.html, Aug 8, 2007)

Chang, Kang-tsung (2006). Introduction to Geographic Information Systems, 3rd Ed. New York: McGraw-Hill

CityNews, 2006 "EMS Crews Go to Wrong Address in Fatal Fire" http://www.citynews.ca/news/news_3482.aspx (July 13, 2006)

Cromley, Ellen K., & McLafferty, Sara L. (2002) GIS *and Public Health.* New York: Guildford Press

Curtis, Andrew and Leitner, Michael (2006). *Geographic Information Systems and Public Health: Eliminating Perinatal Disparity.* Hershey, PA: IRM Press.

DeMers, Michael N. (2005) Fundamentals of Geographic Information Systems, 3rd Ed. Hoboken, N.J.: Wiley.

Environmental Systems Research Institute (2007). How Path Distance Works. (http://webhelp.esri.com/arcgisdesktop/9.1/body.cfm? tocVisable=1&ID=-1&TopicName=How%20PathDistance%20works, Aug. 7, 2007)

Groot, Richard & McLaughlin, John (2000). Geospatial Data Infrastructure: Concepts, Cases, and Good Practice. Oxford: Oxford University Press.

Kiel, Don (2004) "Vermont Develops Spatially Enabled Transportation Tools" *ArcUser Online* (http://www.esri.com/news/arcuser/0704/vtrans.html, Aug 8, 2007)

Korte, George (2001). *The GIS Book, 5th Ed.* Albany, New York: OnWord Press

McKinney, Jim (2005). ArcGIS Network Analyst and Network Dataset. Presented at Business GeoInfo Summit, April 18-19, 2005, Chicago. http://gis2.esri.com/library/userconf/geoinfo05/docs/routing-logistics.pdf (July 11, 2007).

Overton, Ian. "Integrated Stormwater Mapping Across Local Government Areas in Adelaide, South Australia" (http://gis.esri.com/library/userconf/proc00/professional/papers/PAP253/p253.htm, Aug. 12, 2007)

Zeiler, Michael (2005) Modeling our World. Redlands, California, ESRI Press.

Terms Used

- Address Geocoding (Address Matching)
- Address Locator
- Arc
- Barrier
- Cost Distance
- Cost Surface (Cost Raster)
- Determinant Flow
- Directed Networks
- Disabled
- Downstream
- Dynamic Segmentation
- Edge
- Enabled
- Euclidean Distance
- Functional Distance
- Geocoding
- Geographic Information Infrastructure (GII)
- Great Circle Route
- Horizontal Factor
- Impedance
- Indeterminate Flow
- Junction
- Linear Referencing
- Logical Network
- Manhattan Distance
- Network
- Network Distance
- Node
- Planar Graph
- Physical Network
- Raster Data Structure
- Sink
- Source
- Topology
- Transportation Networks
- Turn
- Turn Table
- Undirected Networks
- Uninitialized Flow
- Upstream
- Utility Networks
- Vector Data Structure
- Vertical Factor
- XY Geocoding

4 Land Use Analysis

Understanding the type and amount of land cover and land use is an important process from the standpoint of understanding the Earth as a system.

This module introduces the basics of image classification for the purpose of land cover and land use analysis. The module is based on fundamentals of remote sensing and actively uses definitions, terms, and other material from previous modules both related to remote sensing and GIS.

The module highlights differences between land use and land cover and elaborates on principles of classification systems available worldwide to classify objects on the Earth both at local and regional scale, as well as globally.

The module elaborates on the main approaches to image classification – automated (supervised, unsupervised), manual (or visual), and hybrid, which combines advantages of all methods. Different types of classifiers (mathematical algorithms), widely used for automated image classification, are discussed in the module. This part of the module culminates with an examination of issues of accuracy assessment for image classification.

Change detection in land cover and land use is an inherent part of remote sensing. The module introduces several approaches and techniques used to detect and map such changes – from simple on-screen vectorization to complicated image transformations using spectral rationing and principal component analysis.

The module outlines the significance of remote sensing and GIS for effective use of geospatial data in a variety of corresponding applications.

Theoretical studies and practical assessments in this module are based on knowledge and skills acquired by students through previous units.

Module Outline

Topic 1: Land Cover and Land Use
Topic 2 Image Classification
Topic 3: Remote Sensing for Change Detection and Mapping

4.1 Land Cover and Land Use

4.1.1 Land Cover Analysis

Land cover type is a combination of the physical, climatic, biotic, and human-influenced factors present at a specific location. We need to study land cover since it influences a large number of processes related to a wide range of problems - from environmental impact assessment at the local level to global climate change. Study of global land cover is important to model environmental processes, such as exchange of mass and energy between land and atmosphere that require deriving essential surface parameters for environmental change models.

In many cases the terms "land cover" and "land use" are used interchangeably, however, there is a difference between land cover and land use. Simply put, land cover is what covers the surface of the earth and land use describes how the land is used. Examples of land cover classes include: water, snow, grassland, deciduous forest, and bare soil. Land use examples include: forest management area, agricultural land, urban development, and recreational areas. It is important that each class on the map be clearly defined and distinct from other classes.

Remote sensing has become a powerful tool for land cover and land use identification and for the classification of land surface features produced on any image taken by a satellite. Digital processing of remote sensing data has gained momentum in the last decade, providing a number of advanced methods for automated image interpretation, especially for land use analysis. However, at the same time, one needs to keep in mind the limitations of the remote sensing interpretation process. For example, the CORINE Land Cover (CLC) Programme applied a method for land cover data collection based on a hardcopy inventory from low-resolution imagery, like Landsat 5 satellite printouts. This proved to be the most feasible approach in the mid-eighties, when the program started. This method made only limited use of image processing and GIS software. Due to low-resolution imagery and the limitations of applied remote-sensing and GIS interpretation techniques, one could guess that a lot of Lithuanian forest was lost in five years (1995 - 2000). But in reality, this was not the case.

Land use analysis is widely used in many applications: environmental protection, risk management (deforestation, insect infestation), agriculture (preservation of agricultural lands), urban and regional planning, and many other areas of interest.

Global land cover mapping requires global training data to implement image classification. To implement global land cover mapping the whole surface of the Earth should be covered with satellite images that must be mosaiced (“stitched”) together and radiometrically calibrated. Such process is a time consuming and typically requires powerful computers and parallel processing algorithms. Many questions need to be answered, such as, what sensors would be appropriate (or inappropriate) for doing this research and which classification scheme shall I use?

4.1.2 Land Cover Classification Schemes

A classification scheme effectively defines the legend that will be used for the final map. For example: will the map show forest and non-forest or will it have several or even dozens of different categories? Should the final map categories represent land cover or something else, such as land use, habitat, or conservation areas? The intended use of the map, along with some practical realities, will dictate the content of the classification scheme.

One place to start is to look at some of the common classification schemes. There are a number of classification schemes used for land use and land cover (LULC) maps throughout the world. Some of the common schemes are the CORINE classification system (Europe) and the USGS classification system (USA). Today, ISO 19144-2 "Geographic information - Classification Systems - Part 2: Land Cover Classification System LCCS" standard is under development by the TC 211 Technical Committee of International Organization for Standardization. When choosing an appropriate classification scheme it is important to consider if compatibility with existing schemes is necessary or desirable. Some advantages of using an existing system is that the classes are already defined and the map you produce can be easily compared with other maps using the same system.

The three aims of the CORINE (Coordination of Information on the Environment) program of the Commission European are:

- To compile information on the state of the environment with regard to certain topics which have priority for all the Member States of the Community
- To coordinate the compilation of data and the organization of information within the Member States or at the international level
- To ensure that information is consistent and that data are compatible.

The land cover project is part of the CORINE program and is intended to provide consistent localized geographical information on the land cover of the 12 Member States of the European Community. The accepted common methodology for CORINE Land Cover consists of computer-assisted photo-interpretation of satellite images, with the simultaneous consultation of ancillary data, is classifying data into classes of the CORINE Land Cover nomenclature. During more than 10 years of inventory of the CORINE Land Cover classes much experience has been built up with respect to the visual interpretation of Landsat and SPOT satellite images. CORINE CLC nomenclature is very rich and provides a conceptual framework for understanding the different types of land cover, including procedures for locating, delineating, and identifying land cover units during the satellite image interpretation phase. For each item of the nomenclature, the CORINE guide includes a satellite image, delineation of a unit on the image, an example of a document (ancillary or additional documentation) which will help delineate and identify the unit, and a short commentary on the three illustrations. The illustrations used in the nomenclature have been provided by the various national teams participating in the project. The CORINE land cover central team would like all the teams to submit additional examples of land cover in the format used in the nomenclature (http://www.ec-gis.org/docs/F10418/CLCTECHNICAL_GUIDE.PDF).

4.1.3 Creating Land Cover Maps Using Remote Sensing Data

Mapping of species, communities, or ecosystems is a significant application area of remote sensing and GIS. Land use and land cover maps (LULC) and some of the physical parameters can be efficiently derived from remotely sensed data using different techniques of image interpretation and analysis.

There are several options for creating land cover maps. The following is a list of a few data and interpretation alternatives that can be considered for mapping land cover:

- Visual satellite image interpretation
- Digital satellite image interpretation
- Aerial photo interpretation using stereo photos
- Aerial photo interpretation using single uncorrected photos
- Aerial photo interpretation using orthophotos
- Field surveys using simple angle and distance measurements
- Interpretation of videography
- Interpretation of small format photography

Land use analysis involves the following key processes:

- Importing and geo-registering remote sensing images (use of metadata)
- Development of ground truth database: for training algorithms, operators, accuracy assessment
- Automated image classification : supervised and unsupervised
- Accuracy assessment: calculation of error matrices
- Vectorization of classified data
- Detection of land use changes

How can land use and corresponding changes be detected? By using and analyzing historical and current data (obtaining current data if necessary) and determining the corresponding difference in land cover and use.

The following specific aspects need to be mentioned when using remote sensing for LULC mapping:

- Automated classification utilizes a multiband data set
- To get finer ecosystem details it is necessary to employ sensors with higher resolutions
- In most cases the analysis assesses the utility of automated classification for distinguishing land cover
- When classifying the emphasis is given on empirical classification rather than physical models
- To identify and isolate particular terrestrial features one needs to proceed in a smooth and systematic manner and the data need to be grouped into a suitable framework
- Classification procedures might be adopted for identifying particular land features
- Statistical error matrices are used to determine the effectiveness of automated classification, but ground truthing is essential for overall accuracy assessment.

4.2 *Image Classification*

4.2.1 Why and How Do We Classify Images?

Why do we classify images? To make sense of a landscape by placing it into categories (classes) such as Forest, Agriculture, Water, etc. To classify we use a classification scheme that represents the structure of classes. As discussed above, there are a number of classification schemes. The choice of a scheme (or creation of a custom classification scheme) very much depends on the needs of users.

How does one classify remotely sensed imagery? Information extraction can be made by human methods (visual analysis) or computer (digital analysis). Visual and digital analyses of remote sensing imagery are not mutually exclusive and both have their merits. In most cases, a mix of both methods is usually employed when analyzing imagery. In fact, the ultimate decision of the utility and relevance of the information, extracted at the end of the analysis process, must still be made by humans.

Automated classification is based on the use of computer programs performing statistical analysis of multi-channel imagery. Automated classification is based on statistical analysis of the brightness of image elements (pixels) while manual classification is based on the expertise and empirical skills of the human interpreter who usually uses natural or pseudo color imagery. In many cases manual interpretation utilizes feature identification keys (see Figure 1 for example, or the full set online http://cbc.rs-gis.amnh.org/remote_sensing/guides/image_interp/landsat_key).

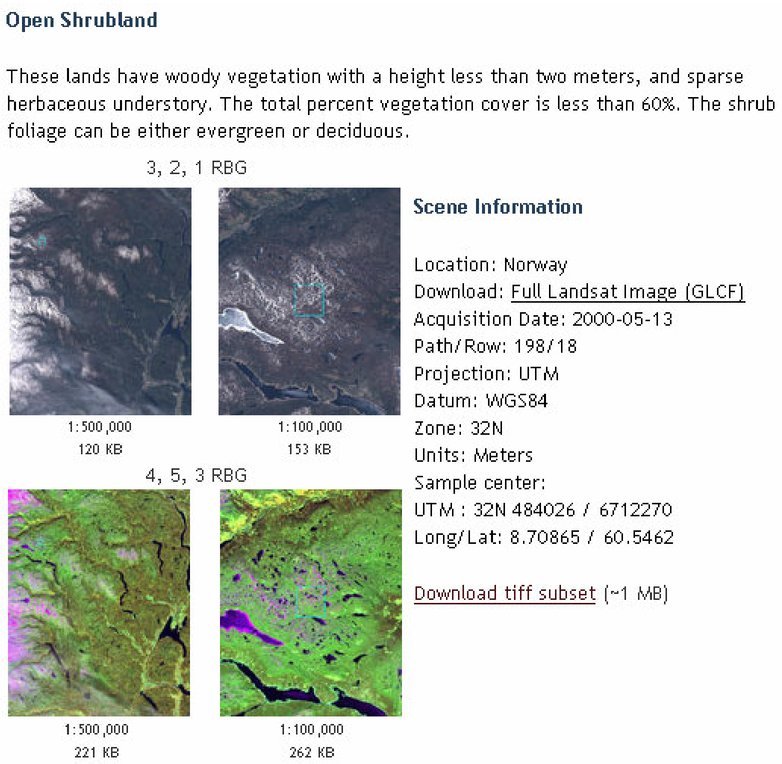


Figure 1. Example of a feature identification key

How accurate are the results of the classification? Accuracy of classification can be assessed statistically (to satisfy statistical criteria) or conceptually (to satisfy an expert's opinion). In both cases it is important to implement ground truthing (or field verification) – i.e. identification of features in image and on the ground by actually visiting the place. Ground truthing can be done prior to or after classification.

Field verification can be considered a form of collateral material. Field verification is typically conducted to assist in the analysis of the data to be analyzed. Essentially, this is familiarizing the interpreter with the area or type of feature or object to be interpreted. This type of verification is done prior to the interpretation. After an interpretation, field verification can be accomplished to verify the accuracy of the interpretation conducted. The choice depends on the goal of image interpretation, techniques employed, and the likelihood of implementing fieldwork.

4.2.2 Automated classification using remotely sensed data and GIS

Figure 2 illustrates the general workflow pattern of an automated classification of remotely sensed data that involves image processing and GIS techniques. In many cases, steps shown in the figure are omitted or reordered, depending on the purpose of the analysis.

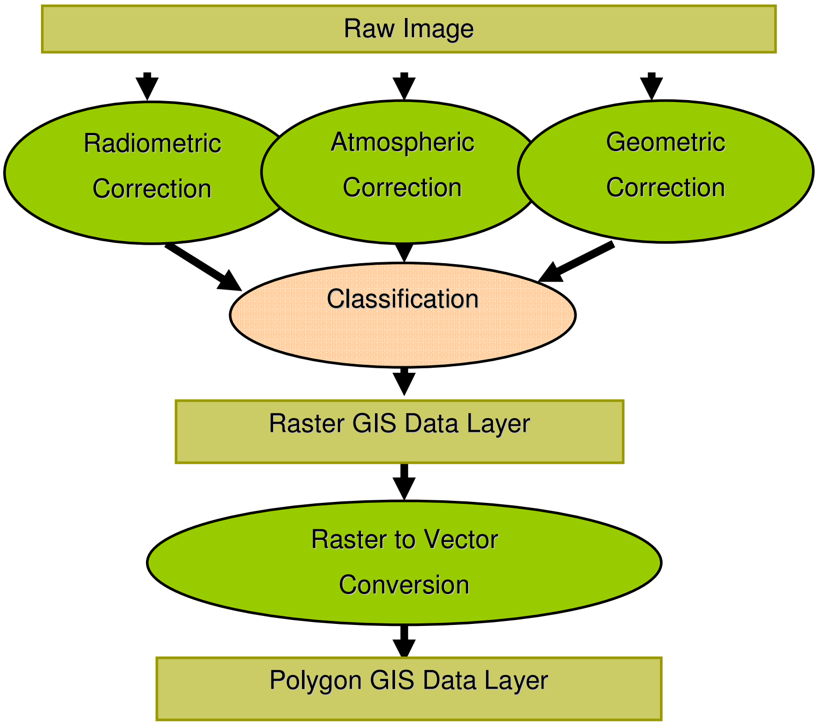


Figure 2. Automated image classification

A raw image, acquired by a sensor, is burdened by many errors that needed to be corrected. Apart from calibration of a sensor, image **radiometric correction** deals with such casual problems as striping and (partially) missing lines. Differences in illumination caused by sun angle, viewing angle (sensor's position relative to the object), and terrain effects need to be corrected at this stage. Nowadays, the vendors and suppliers are usually shipping remote sensing images to customers after corrections of some systematic radiometric distortions have been made.

Image **atmospheric correction** is applied to remove the impact of the atmosphere on the signal received by a sensor. Atmospheric correction is needed to remove the atmospheric component in the whole chain of physical relation of radiance to surface property. Since scattering increases inversely with wavelength, it impacts on the processing of multispectral data used for visual analysis. There are a number of approaches to atmospheric correction. Image-based methods utilize the histogram minimum method and the regression method. Other methods are based on radiative transfer models or implement the empirical line method.

Geometric correction. All remote sensing imagery is inherently subject to geometric distortions. These distortions may be due to several factors, including perspective of sensor optics, motion of scanning system, motion of platform, platform altitude, attitude, and velocity, surface terrain, curvature, and rotation of the Earth. These variations might be systematic or random. Correction of these distortions is done by geometric transformations of the imagery.

Image classification. Automated classification is the process of taking the “brightness values” associated with each pixel and using them to assign a class to the corresponding output pixels (See figure 3). The resulting output pixels are not continuous. Instead, they are discrete values that represent a class or category (e.g. land cover classes).

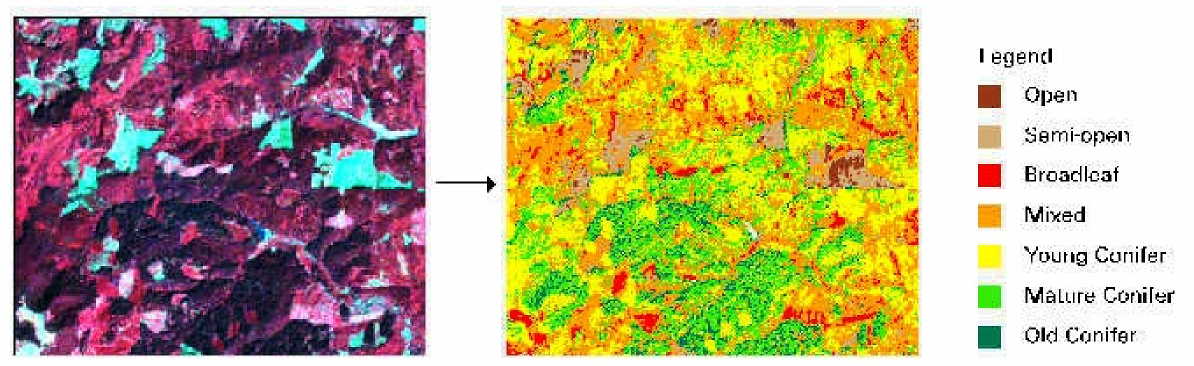


Figure 3. Automated classification of remotely sensed imagery

Vectorization of classified rasters. Result of automated image classification is a classification map in raster format. Variation of pixel values in raster classification maps are very limited and defined by the number of classes (usually 5 to 40). Vectorization is, therefore, a relatively straightforward process. At the same time, there are some drawbacks of vectorization of classified rasters. Single “orphan” pixels need to be detected and deleted or aggregated because automatic conversion to polygons is not feasible, especially if there are many tiny polygons. To aggregate pixels into more homogenous groupings, one uses different filtering techniques following the automatic conversion of the aggregated pixel groups into corresponding polygons.

4.2.3 Spectral Classification: Basic Strategy

Different objects on the Earth have different spectral signatures. EM energy, reflected by objects, is registered by a remote sensor in different spectral bands. Therefore, the same object is depicted differently in different spectral bands with corresponding values of pixels in different bands of a digital imagery (Figure 4). Automated image classification is based on the use of the spectral signature of an object, registered by a multi-spectral sensor.

Example: A forest has its unique spectral signature; all “Forest” pixels would have similar digital pixel values within the same spectral band of a sensor. If we decide that pixels with the value (20, 30, 190) (in band 1 thru 3, respectively) indicate that the output pixel should be assigned a value of “10” – corresponding to class “10”=forest, then we could just say that any pixel in an image with that signature was forest. We could do the same for soil and end up with a map of classes.

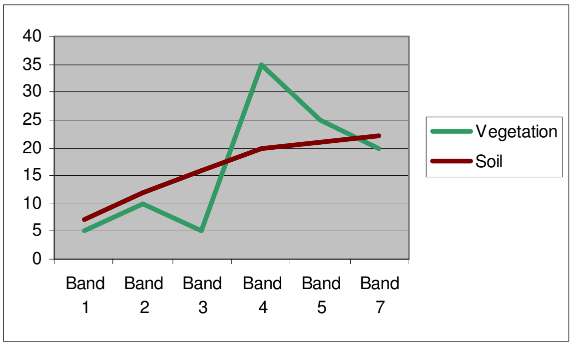


Figure 4. Spectral signatures for two objects (simplified, for illustration only)

4.2.4 Methods of Image Classification

The automated classification process involves translating the pixel values in a satellite image into meaningful categories. In the case of land cover classification, these categories comprise different types of land cover defined by the classification scheme that is being implemented. There are dozens, if not hundreds, of classification methods that can be used to group image pixels into meaningful categories. Unfortunately, there is not a single "best" approach to image classification. The choice you make depends a lot on the algorithms that are available to you with the image processing software you use and your familiarity and experience with the different methods.

There are three fundamental approaches to classification

- Supervised
- Unsupervised (no extraneous data is used: classes are determined purely on difference in spectral values. The computer selects classes based on clustering of brightness values)
- Hybrid (use unsupervised and supervised classification together)

Supervised classification requires “training pixels”, pixels where both the spectral values and the class is known. User specifies the classes to be used and provides “signatures” for each class. In supervised classification, you help the computer to select “signatures” that represent each land cover class. Signatures are statistical descriptions of the brightness values of a given land cover type (e.g., the mean band 1 value, the mean band 2 value, etc.). You select signatures using a tool that provides “seed” values.

Typical steps in supervised classification:

1. Decide on classes
2. Choose training pixels that represent these classes
3. Use the training data with a classifier algorithm to determine the spectral signature for each class
4. Using the classifier, each pixel can be labeled in one of the pre-determined classes (or, potentially, in the “other” class)

In **unsupervised classification,** no extraneous data are used in order to run a classification algorithm: classes are determined purely on differences in spectral values. The computer selects classes based on clustering of brightness values. Unsupervised classification refers to a variety of different techniques that share some features in common. They use statistical “clustering” techniques to decide which pixels should be grouped together. With luck, these clusters of pixels will correspond to land cover classes. Unfortunately, this correspondence may not be one-to-one - one land cover class may be represented by more than one cluster or one cluster may represent more than one land cover type (not easily fixed – may need to specify more clusters).

A **hybrid approach** combines the advantages of the automated and manual methods to produce a land cover map that is better than if just a single method was used. One hybrid approach is to use one of the automated classification methods to do an initial classification and then use manual methods to refine the classification and correct obvious errors. With this approach, you can get a reasonably good classification quickly with the automated approach and then use manual methods to refine the classes that did not get labeled correctly.

Signature selection. The key to a good supervised classification is proper selection of “signatures”. What makes for a good signature? A spectral signature should uniquely characterize a land cover type of interest and it needs to be distinguishable from other signatures.

There are two main methods for obtaining signatures:

- Area-based – where you use a box on the screen to select the area to be statistically characterized as a signature, and
- Growing an area – where you select a point and other “similar” adjacent points are added to the sample. Please note that choice of minimum similarity can have a big effect on the results of this method.

4.2.5 Classification and Classifier Algorithms

Classification is a repeated process to add new classes. Due to color variations within a class, often multiple signatures will be needed to capture a single cover class.
The classes can be recoded and lumped (using basic editing functions) so that they correspond to the desired classes.

There are a number of classifier algorithms used both in supervised and unsupervised classifications. Among the most frequently used are:

- Table look up
- Parallelepiped
- Minimum distance
- Maximum likelihood

Many software packages for image analysis offer a variety of other algorithms (e.g., binary decision trees, artificial networks, etc.). Quite often classifiers are not limited to raw spectral digital numbers (DNs) (values of pixels in the source image), but use the other inputs such as:

- Transformed spectral data: vegetation indices, end member fractions, depth of absorption features
- Spatial data: distance to certain landscape features
- Texture data: difference between spectral values of a pixel and its neighbors
- Temporal data: change detection results

Table Look Up. For each class, a table of band DNs are produced with their corresponding classes. For each image pixel, the image DNs are matched against the table to generate the class. If the combination of band DNs is not found, the class cannot be determined. Modification is possible. In this case, the table can be a range of values. Table Look Up is conceptually easy and computationally fast, but every potential combination of band DNs and their class is known (or the range), since matching pixels against spectral libraries is a type of table look up.

Parallelepiped uses maximum and minimum DNs on each individual band (as derived from each signature) to decide which pixels fall within a given class (see Figure 5). This approach has certain benefits (simple to train and use, computationally fast) and drawbacks (pixels in the gaps between the parallelepipeds cannot be classified; pixels in the region of overlapping parallelepipeds cannot be classified; uses only a fraction of the information contained in the signature data).

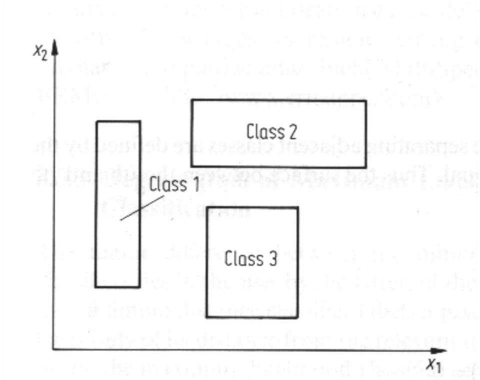


Figure 5. Parallelepiped classifier

Minimum Distance. A “centroid” for each class is determined from the data by calculating the mean value by band for each class (see Figure 6). For each image pixel, the distance in ***n*** -dimensional distance to each of these centroids is calculated and the closest centroid determines the class, where ***n*** is the number of image bands or channels used for classification. Benefits: mathematically simple and computationally efficient. Drawback: insensitive to different degrees of variance in spectral response data.

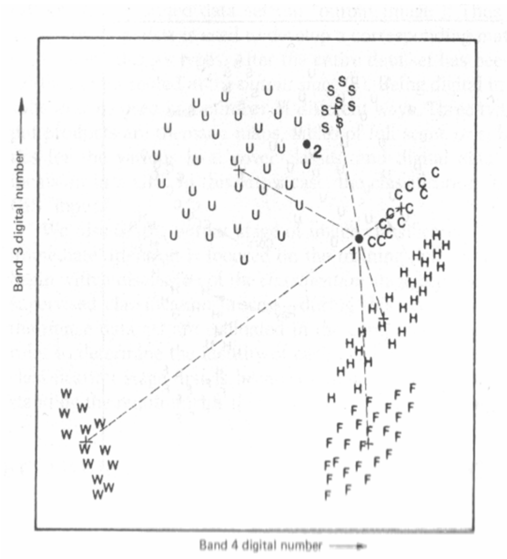


Figure 6. Minimum distance classifier

Maximum Likelihood uses statistical techniques to decide which class a pixel falls into. This classifier uses full signature information (mean, variation and inter-band covariation) to tease apart similar classes; thus it is more computer intensive.

Gaussian Maximum Likelihood uses the variance and covariance between training class spectra to classify the data. It assumes that the spectral responses for a given class fit a normal (Gaussian) distribution (Figure 7).

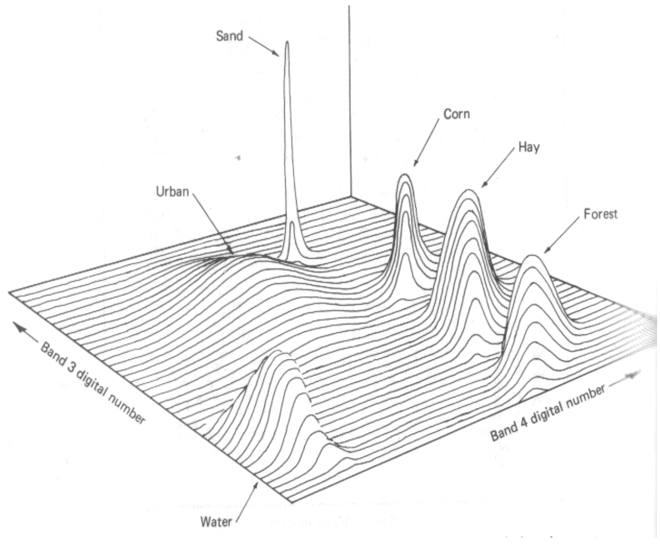


Figure 7. Gaussian Maximum Likelihood classifier

In Gaussian Maximum Likelihood, we assign any given point in spectral space (i.e. a set of DN values of a pixel) to a specific class. The assignment ('classification') is made to that class whose probability of occurrence at that point in spectral space is highest ('most likely') - or to "Other" if the probability is not over some threshold.

Benefits: uses the class covariance matrix representing not only the first but also the second order statistics that often contain much of the information in RS data. Drawbacks: computationally inefficient, multimodal or non-normally distributed classes must be split into normally distributed subclasses.

Both Gaussian Maximum Likelihood and Maximum Likelihood classify according to what is most likely – according to which class probability density curve is higher at a particular DN value. Maximum Likelihood assumes nothing about the shape of that probability density curve. It can be parametric or non-parametric, Gaussian, normal, or whatever - it is whatever you provide. On the other hand, Gaussian Maximum Likelihood assumes that the probability density curves of the information classes fit normal (Gaussian) distributions.

Binary Decision Trees are a common machine-learning tool that has taken hold in the remote sensing arena. Decision trees are a set of binary rules that define how specific land cover classes should be assigned to individual pixels. At each node a true/false decision is made thereby creating a branch in the tree (Figure 8). This approach makes it easy to integrate ancillary data into the classification process.

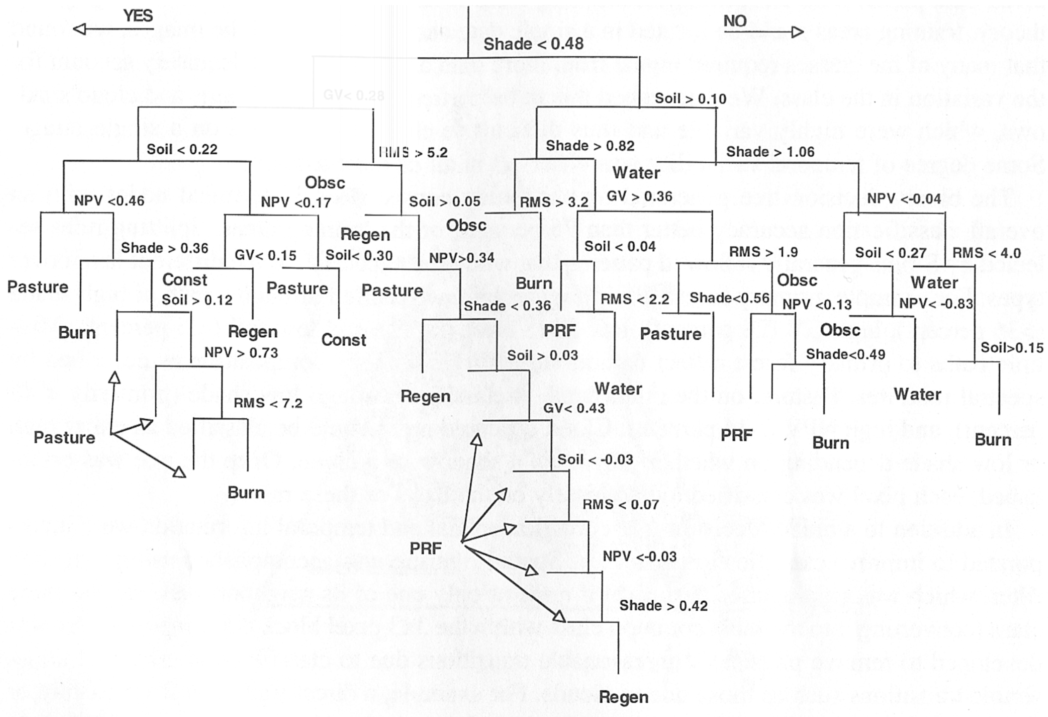


Figure 8. Binary Decision Trees

Artificial Neural Network classification algorithms attempt to mimic the human learning process to associate the correct land cover class label to image pixels. Neural networks have their roots in the field of artificial intelligence and although the algorithm still pales in comparison with the human brain its use in image classification has been quite successful. One advantage to using neural networks is that it is easy to incorporate ancillary data in the classification process. This, in itself, can greatly improve classification accuracy. The training process for a neural network classification can be time consuming and is not as simple as the supervised statistical approach.

4.2.6 Accuracy Assessment of Image Classification

"Without an accuracy assessment, a classified map is just a pretty picture."

A classification process is not complete without assessment of accuracy of the final map but some thought should be given to what is the required accuracy and how it can be assessed. The accuracy can refer to either spatial accuracy or thematic accuracy. Spatial accuracy is directly related to the base information that is used. Thematic accuracy specifies how well individual mapped classes match what is on the ground.

Thus, the goals or accuracy assessment are:

- Assess how well a classification worked
- Understand how to interpret the usefulness of someone else's classification

To implement accuracy assessment we often need reference data, which can be obtained from aerial photo interpretation, ground truth with GPS, GIS layers, and other sources.

The general approach can be outlined as follows:

- Collect reference data: "ground truth"
- Determine class types at specific locations
- Compare reference to classified map
- Does class type on classified map = class type determined from reference data?

In many cases, the accuracy of a land cover map is assessed using statistical sampling procedures. The basic idea is to select a number of sample sites from each of the cover types in the final image and then go into the field to see what type of cover is actually on the ground. This information is then compiled into a contingency table so that the accuracy of each class can be determined (see Figure 9).

	Class types determined from **reference source**				
	# Plots	Conifer	Hardwood	Water	**Totals**
Class types determined from **classified map**	Conifer	50	5	2	57
	Hardwood	14	13	0	27
	Water	3	5	8	16
	Totals	67	23	10	100

Figure 9. Example of error matrix for image classification

Classification: Final thoughts. Classification of remotely sensed imagery is a fusion of art, science, and skill. Common practice suggests that classifications are never complete - they end when time and money run out. Classification is iterative - its tough to get it right for the first few iterations. It is always wise to consider a hybrid classification - part supervised and part unsupervised, including manual classification. Classification often requires editing and should not be considered cheating!

4.3 Remote Sensing for Change Detection and Mapping

Remote sensing data and methods are widely used to identify and map changes in land cover and land use. Change detection methods and techniques very much depend on data availability. There are two major approaches to mapping land use change: to compare source data (remotely sensed imagery) and to compare derived vector data (maps).

4.3.1 Mapping Land Use and Land Cover Changes

Mapping land cover and land use changes is a very common task in environmental studies. To implement LULC change detection and mapping we need to apply some criteria to identify and document: past land use, current land use, and predicted land use or plan and, if possible, future land use. LULC change detection and mapping is based on temporal resolution of sensors and employs time series analysis.

To detect and map changes we can use:

- Raw (source) images to classify two images and then find difference
- Derived image products to derive image components or indexes and then find difference
- Derived vector products to derive vector features and then find difference

The particular approach depends on the type of data: raster data requires image arithmetic as the main technique while vectors employ different operators of map algebra.

There are several practical techniques used for change detection mapping:

- Vector analysis based approach
  - "Heads-up" on-screen digitizing and differencing
- Raster analysis based approach
  - Multi-date RGB visualization
  - Image algebra using differencing/rationing
  - Vegetation indexes' spectral differencing
  - Post-classification comparison
  - Multi-date composite image using PCA

"Heads-up" On-screen Digitizing and Differencing is a totally manual technique requiring great amount of manual work. Two temporally separated images are geo-referenced and then digitized using a GIS package by employing its manual editing tools. Both vector outputs (GIS layers) are then compared either manually or automatically (using map algebra) and the differences are interpreted as spatio-temporal changes.

Multi-date RGB Visualization. Visual examination of 2 or 3 time periods by "coloring" up land cover change areas**.** Georeferenced images are colored in three primary colors (red, green, and blue). Being superimposed, they result in a pseudo-color image when color is interpreted as change (theoretically, if no changes occur, the result is a grey image). This approach enhances the visual interpreter's ability to discern change, but at the same time is non-quantitative, that is, it does not implicitly indicate type of changes, such as "from class A to class B".

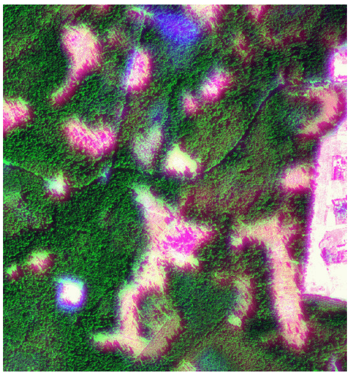


Figure 10. False color Landsat image is pan-sharpened with panchromatic ortho-photo image – blue area on the top clearly is cut area. This approach was tested a decade ago for forest inventory in Lithuania with outdated ortho-photos

Image Algebra Using Differencing/Ratioing. Two color images are differenced or rationed to generate a binary image indicating occurrence of changes. This is an efficient method of identifying "change" pixels but requires a careful selection of 'change/no-change' threshold. Similar to the previous one, this method does not indicate "from-to" change classes. To determine 'change/no-change' threshold the user must examine a histogram of a difference image where peak of histogram corresponds to no change and tails of the histogram indicate change. Typical shortcomings of this method are: there is no single 'change/no-change' threshold. A too liberal threshold leads to commission errors; i.e., calling non-change pixels changed. A too conservative threshold introduces omission errors; i.e., calling changed pixels no change.

Image division or spectral ratioing is one of the most common transforms applied to image data:

Ratio = band (A) / band (B)

Image ratioing serves to highlight subtle variations in the spectral responses of various surface covers. By ratioing the data from two different spectral bands, the resultant image enhances variations in the *slopes of the spectral reflectance curves* between the two different spectral ranges that may otherwise be masked by the pixel brightness variations in each of the bands. Results of image ratioing is called indices (e.g., vegetation indexes).

4.3.2 Vegetation Indices

Vegetation indices are simple band combinations and their ratios that are commonly used for either vegetation or mineral delineation. Indices are used extensively in mineral exploration and vegetation analyses to bring out small differences between various rock types and vegetation classes.

In many cases, judiciously chosen indices can highlight and enhance differences that cannot be observed in the display of the original color bands.

Commonly used ratios for Landsat images are:

- Clay Minerals = TM 5/7
- Ferrous Minerals = TM 5/4
- Ferric Minerals (iron oxide) = TM 3/1
- Mineral Composite = TM 5/7, 5/4, 3/1

- Hydrothermal Composite = TM 5/7, 3/1, 4/3

Normalized Difference Vegetation Index (NDVI) is widely used for vegetation studies using remotely sensed data. It is well known that healthy vegetation strongly reflects in the near-infrared zone (NIR - 0.8 to 1.1 mm), and, at the same time, strongly absorbs in the visible red (0.6 to 0.7 mm), while other surface types (soil and water) equally reflect in both the near-infrared and red zones.

NDVI is calculated as NDVI = (NIR-R)/(NIR+R)

The formulae gives RatioValues significantly exceeding 1.0 for vegetation and RatioValues approximately equal to 1.0 for soil and water. Thus, the discrimination of vegetation from other surface cover types is significantly enhanced. Using NDVI we can better identify areas of unhealthy or stressed vegetation, which show low near-infrared reflectance, as the ratios would be lower than for healthy green vegetation.

NDVI model using Landsat TM imagery is derived using band numbers

$$NDVI = \frac{TM4 - TM3}{TM4 + TM3}$$

Other algorithms available for Landsat images:

- Vegetation Index = TM4 - TM3
- IR/R = TM4/TM3
- Sqrt(IR/R) = Sqrt(TM4/TM3)

Post-classification Comparison: Two raster maps. This approach requires two preliminary made classifications to be compared to each other. The output provides ‘from-to’ change information, but it is very dependent on accuracy of individual date classifications. This method requires two separate classifications (for each data set).

4.3.3 Image Arithmetic: Cross Tabulation

Raster map comparison is easily facilitated using the *Tabulate Area* function in ArcGIS (Spatial Analyst Tools) by determining the cross tabulation between two grid themes on a cell-by-cell basis. Once the tabulations are made, the data is displayed in a simple matrix where Map One is the X axis and Map Two is the Y axis.

Assume we have a 9-cell land cover map from 1990 with three categories: A, B, and C. We also have another map from 2000. Then, the cross-tabulation calculation will look like the following expression:

1990

A	B	B
B	B	C
B	A	C

Cross Tabulate

2000

A	A	B
B	C	C
A	A	B

=

Cross Tabulated Grid

AA	BA	BB
BB	BC	CC
BA	AA	CB

You can see that the resulting cross tabulation provides a pixel-by-pixel comparison of the interpreted land cover types with the reference land cover. Therefore, for the upper right-hand cell, the 1990 data was B, and the 2000 data was also B. Therefore, this is a match between the two data sets. However, in the lower right cell you can see that the 1990 data indicated a value of C and the 2000 data set had a value of B.

We can now quantify the results into a matrix as shown below.

		1990 Data		
		A	B	C
2000 Data	A	2	3	0
	B	0	2	1
	C	0	1	1

This matrix shows the comparison of two hypothetical data sets. The **1990** data set represents the land use in 1990, while the **2000** data set represents the land use in 2000.

4.3.4 Principal Components Analysis (PCA)

Different bands of multispectral data are often highly correlated and thus contain similar information. For example, Landsat MSS Bands 4 and 5 (green and red, respectively) typically have similar visual appearances since reflectances for the same surface cover types are almost equal.

Image transformation techniques based on complex processing of the statistical characteristics of multi-band data sets can be used to reduce this data redundancy and correlation between bands. One such transform is called principal components analysis (PCA). The objective of this transformation is to reduce the dimensionality (i.e. the number of bands) in the data and compress as much of the information in the original bands into fewer bands. The "new" bands that result from this statistical procedure are called components.

This process attempts to maximize (statistically) the amount of information (or variance) from the original data into the least number of new components. As an example of the use of Principal Components Analysis, a seven-band Thematic Mapper (TM) data set may be transformed such that the first three principal components contain over 90 percent of the information in the original seven bands. Interpretation and analysis of these three bands of data, combining them either visually or digitally, is simpler and more efficient than trying to use all of the original seven bands. Principal components analysis, and other complex transforms, can be used either as an enhancement technique to improve visual interpretation or to reduce the number of bands to be used as input to digital classification procedures, discussed in the next lecture.

General Steps in image classification using PCA:

1. Geometric correction of images
2. PCA transformation
3. Interpretation of PCA images (each image will contain a single axis).

Figure 11 illustrates results of Principal Component Analysis for a Landsat image.

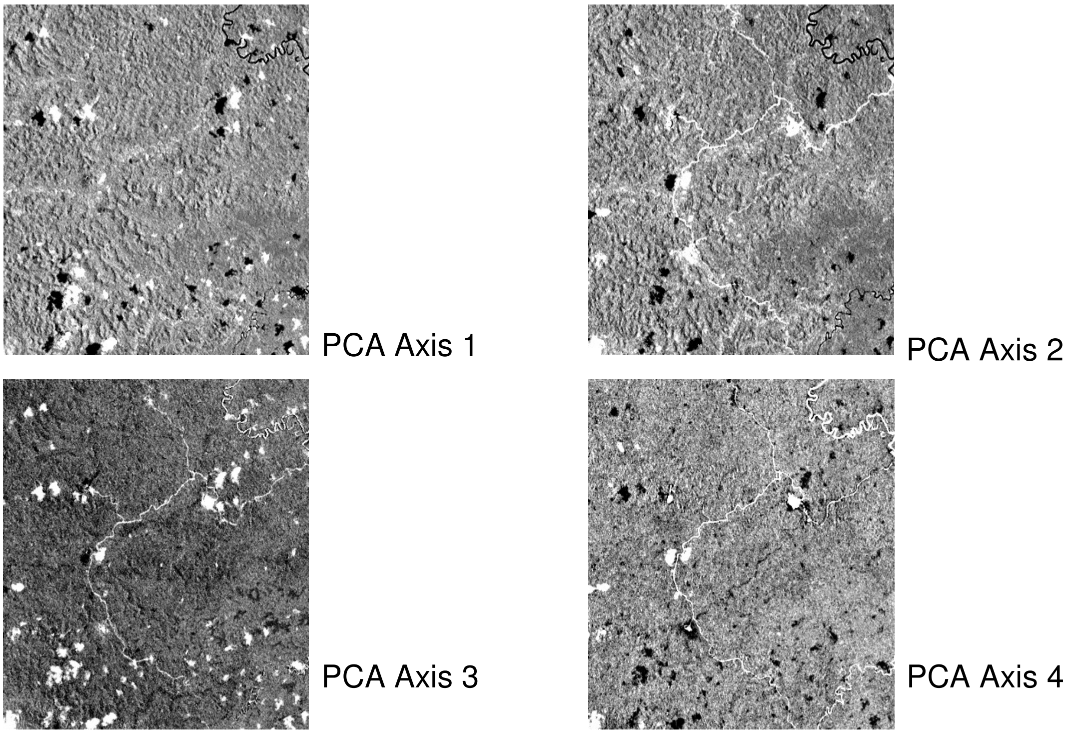


Figure 11. Result images of Principal Component Analysis for a Landsat image

Multi-Date Compositing using Principal Component Analysis (PCA). Principle Components Analysis can be used for change detection. PCA takes multidimensional or/and multi-temporal data and reduces it to axes of uncorrelated data. The first PCA axis (PC1) will describe the most variance in the image data; the second axis will have the next most variance, etc… PC1 and PC2 have been found to represent unchanging land cover, and PC3 and higher tend to represent changing land cover. This is an efficient method of identifying "change" pixels. At the same time, this method provide very little "from-to" change classes, which are difficult to label. This method is relatively easy to implement, it can reduce a large dataset into a much smaller dataset, but interpretation can be difficult.

4.3.5 Applications of Image Classification and Change Detection

The techniques discussed above are widely used in many applications, such as:

- Environmental protection (acid rain, water pollution):
  - For example, on the following image, the dark areas off the coast represent the areas where oil is present and areas of lighter tone directly to the south are areas where dispersant was sprayed on the oil to encourage emulsification:

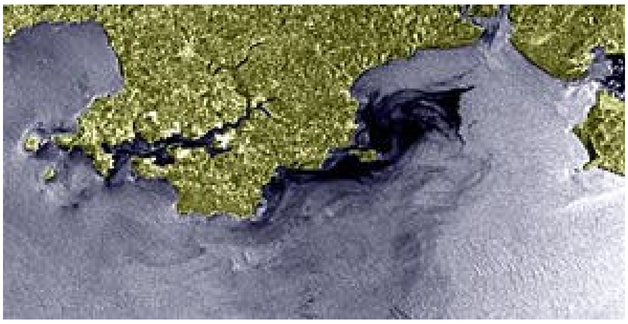


Figure 12. RADARSAT Image (http://ccrs.nrcan.gc.ca/)

- Management of natural hazards (deforestation, insect infestation, effects of global warming)
    - Higher order PCs may contain change information:

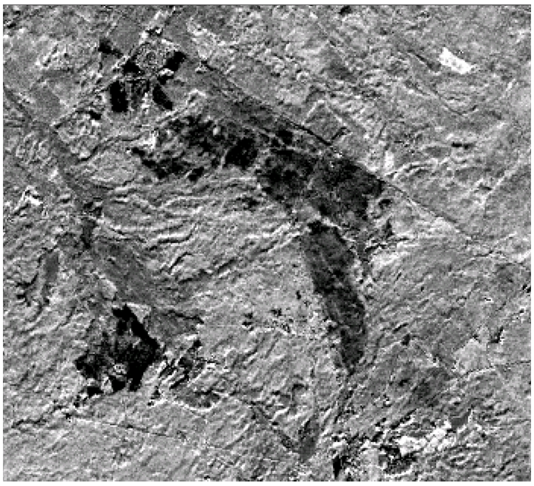


Figure 13. PC4: Blow-up of fire scar

- Agriculture (preservation of agricultural lands, promoting efficiency of farms)
    - A severe drought monitoring and planning program to seek help from outside wheat markets:

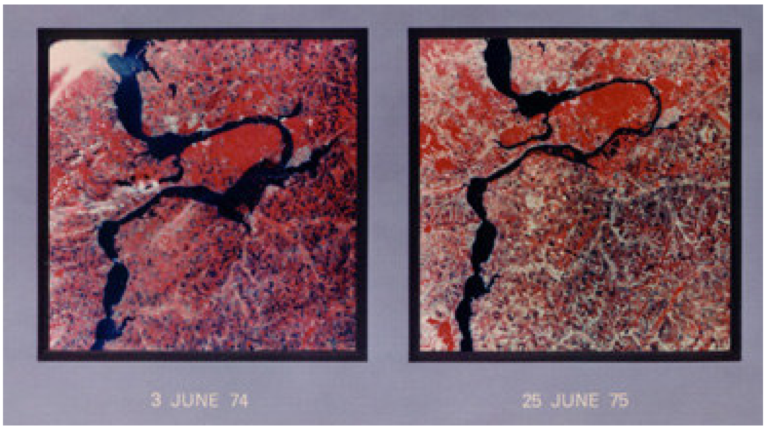


Figure 14. Landsat images proved the veracity of the Soviet Union plea for help in 1975. 3 June 1974 sub-scene shows a large bend in the Volga River and the fields are in normal crop stages. A year later, mature

crops should have increased the scene redness as the year before, instead much of the farmland is fallow (darker grays and tans), confirming the drought claims (http://www.fas.org/irp/imint/docs/rst/Front/tofc.html).

- Urban and Regional Planning:
    - For example, thermal images can be calibrated in order to show temperature at the surface. Municipal workers can analyze heat lost in urban areas or study house occupancy.

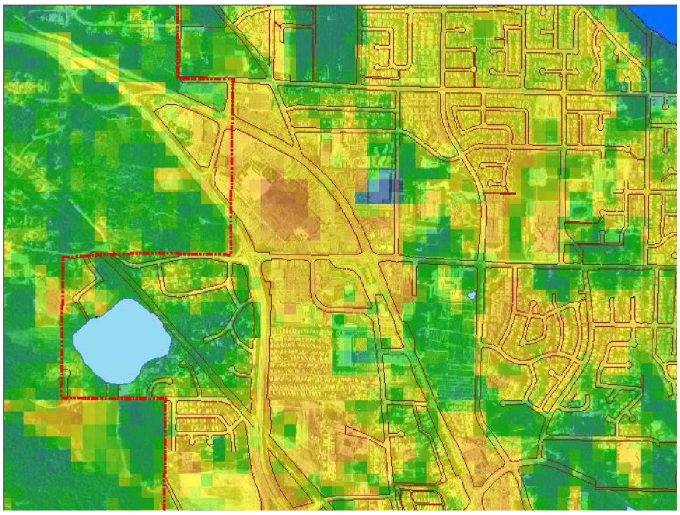


Figure 15. Calibrated thermal band of Landsat 7 imagery, red pixels represent hotter spots, green pixels represent cooler spots

4.4 Summary

- Understanding the type and amount of land cover in an area is an important characteristic from the standpoint of understanding of Earth as a system
- Land cover is not necessarily equivalent to land use
- Remote sensing has become a powerful tool for land cover identification and classification of various features of land surfaces in images taken from satellites
- Classification is one of the most widely used analysis techniques in remote sensing (it is easy to collect class data relative to many continuous data)
- Classification uses spectral (radiometric) differences to distinguish objects
- Supervised classification, unsupervised classification, and hybrid classification are the key methods for image analysis for land cover and land use mapping

Module self-study questions:

1. What is land cover and how does it differ from land use?
2. Describe the main approaches to image classification. Outline a typical image classification workflow.
3. Elaborate on the various types of satellite image corrections.
4. What is the fundamental concept of automated image classification?
5. Explain the principles of main classifier algorithms (parallelepiped and minimum distance).
6. Describe the goals and outline the general approach of accuracy assessment in image classification.
7. List and describe the methods used for mapping land cover changes using remotely sensed imagery.

Required Readings:

[1] Section 1: Image Processing and Interpretation - Morro Bay, California, The "Short" Tutorial: Short, N., NASA, http://www.fas.org/irp/imint/docs/rst/Front/tofc.html

[2] Section 3: Vegetation Applications - Agriculture, Forestry, and Ecology; Other Ecology Examples, The "Short" Tutorial: Short, N., NASA, http://www.fas.org/irp/imint/docs/rst/Front/tofc.html

[3] Section 4: Urban and Land Use Applications - From Los Angeles to Beijing, The "Short" Tutorial: Short, N., NASA, http://www.fas.org/irp/imint/docs/rst/Front/tofc.html

[4] Part One –Methology, CORINE Land Cover. Technical Guide, http://www.ec-gis.org/docs/F10418/CLCTECHNICAL_GUIDE.PDF

References

- Lillesand, T.M., Kiefer, R.W. (2000), *Remote Sensing and Image Interpretation*, 4th Edition. John Wiley and Sons, Inc.
- CCRS Tutorial: Image Analysis http://ccrs.nrcan.gc.ca/resource/tutor/fundam/index_e.php
- The "Short" Tutorial: Short, N., NASA, http://www.fas.org/irp/imint/docs/rst/Front/tofc.html
- Land Cover Classification System. Classification concepts and user manual. Software version (2). Antonio Di Gregorio and Louisa J.M. Jansen, Food and Agriculture Organization of the United Nations, Rome, 2005
- A Land Use And Land Cover Classification System For Use With Remote Sensor Data, James R. Anderson, Ernest E. Hardy, John T. Roach, and Richard E. Witmer, USGS, 1976
- CORINE land cover technical guide, M. Bossard, J. Feranec, J. Otahel, European Environment Agency, 2000

Terms used

- Land use and land cover
- Temporal image resolution
- Radiometric correction
- Atmospheric correction
- Geometric correction
- Image classification
- Unsupervised classification
- Supervised classification
- Hybrid classification
- Classification accuracy assessment
- Change detection
- Spectral ratio
- Normalized difference vegetation index (NDVI)
- Principal component analysis (PCA)

5 Working with Digital Terrain Models

Outline:

22. Introduction
23. Representation of Elevation
24. Digital Terrain Models
25. Construction of Digital Terrain Models
26. Digital Terrain Model Analysis and Applications
27. Examples
28. Conclusion

5.1 Introduction

The Earth is not flat. We take this fact for granted in modern times, but yet, when working with medium-scale maps, it is quite reasonable to model the Earth's surface as if it were flat. For thousands of years, we have created maps on flat surfaces, using media such as clay tablets, papyrus, vellum, or paper.

The reason that we can do this is that, at such scales, the Earth is *close to* flat. Consider a map of the Himalayas, the highest mountain range on Earth. At a scale of 1:2,500,000, the 2400 km long range would be just under one metre in width. If a three-dimensional model were to be produced at the same scale, the highest peak (Mt. Everest) would be only 3.5 millimetres high. However, if we model the curvature of the Earth, it would add an additional 9 millimetres of height at the centre of the map. So, for small-scale mapping, topography is insignificant, but the Earth's curvature is important.

Now, let us consider a map of Mt. Everest at a scale of 1:20,000. If the map were still 1 metre wide, it would cover an area 20 kilometres across (Figure 50). In this case, Mt. Everest would be over 44 centimetres high, and the Earth's curvature would add only 5 centimetres at the centre of the three-dimensional model. Clearly, for medium and large scales, topography is important!

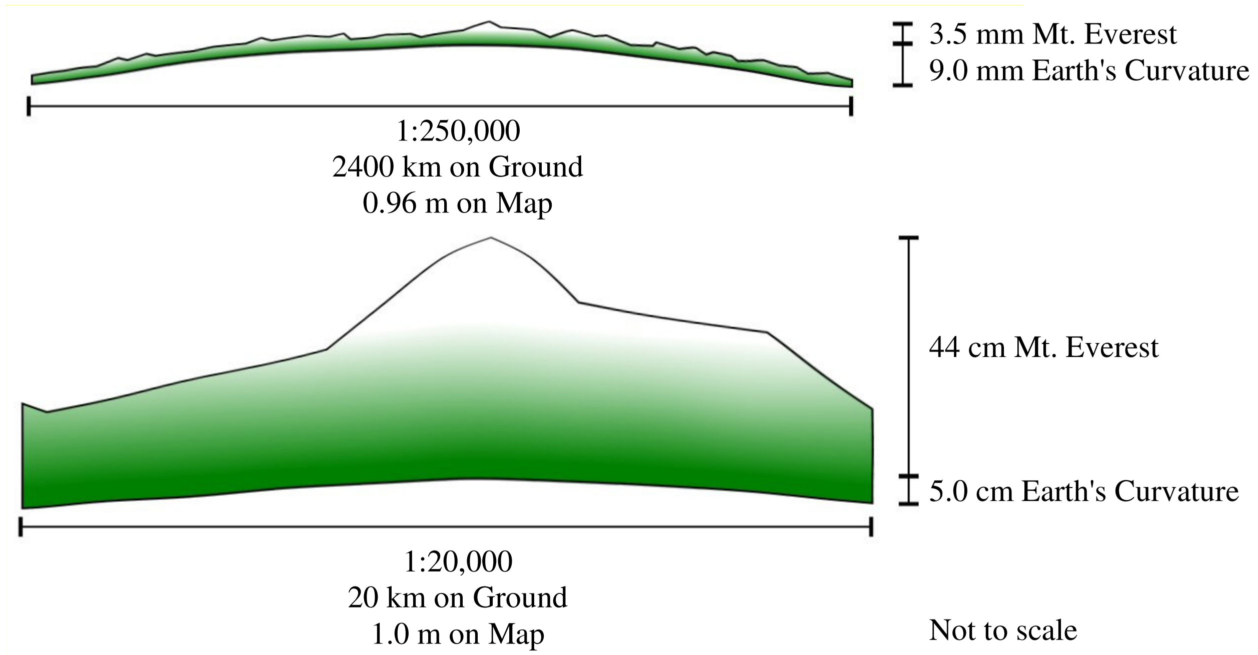


Figure 50. The effects of topography become more noticeable at larger scales, while the effects of the Earth's curvature become less noticeable.

As we move from small map scales to larger scales, we see the effect of topography becoming more pronounced, and the effect of the Earth's curvature becoming less important. As more and more large-scale spatial data become available, the relative importance of topography increases. Fortunately, GIS offers us a set of tools for modeling topography. In this module, we will examine how to represent topography using digital terrain models (DTMs), and how to use DTMs to study terrain.

5.2 Representation of Elevation

In GIS, we are concerned mainly with the surface expression of terrain. This is admittedly a generalization, and there are a number of areas where GIS might be used to model subsurface conditions, most notably for geological maps. For studies involving terrain and its effects on human activity, however, we generally do not concern ourselves with what is going on beneath the ground.

GIS therefore concerns itself with what are called "2.5D" representations of the Earth's surface. These are similar to a slightly crumpled sheet of paper. We can see the shape of the surface, but the surface is infinitely thin, and we have no information about what lies below. The reason that full three-dimensional representations are not used has to do with the volume of data that must be represented. 2.5D representations are reasonably compact, and can be drawn quickly on a computer screen. Until recently, very few computers had the graphics-processing power to display fully three-dimensional objects. This situation has now changed because of the popularity of three-dimensional video games.

For full 3-D representations, another technology called voxel representation can be used to represent subsurface structures. These are often used in engineering and geological applications where it is necessary to have a full understanding of subsurface structures. As you might have guessed from the name, voxels (short for "volume elements") are like three-dimensional pixels (i.e. cubes instead of squares), which can be subdivided or aggregated as necessary to represent finer or less detailed features (Figure 51).

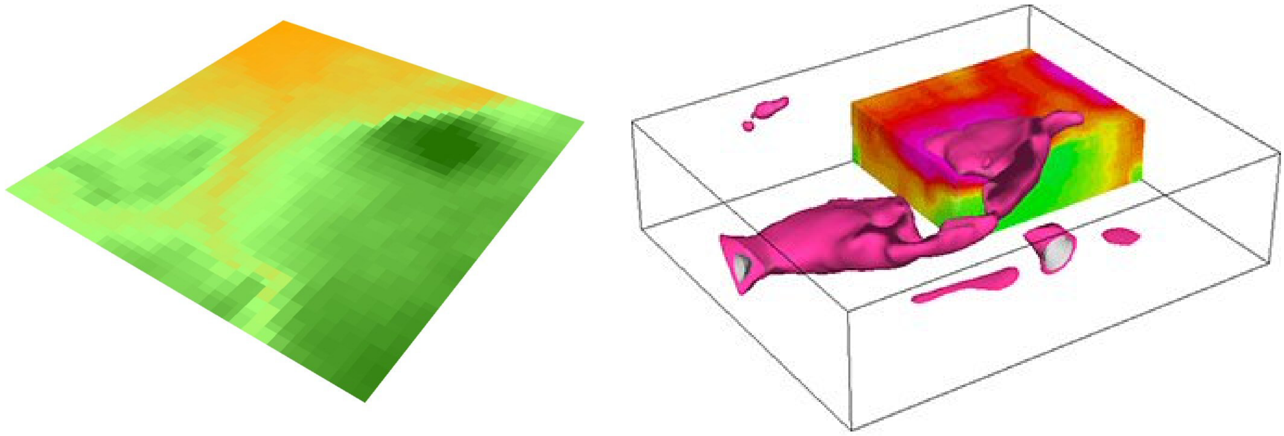


Figure 51. DEMs (left) have no thickness and are composed of square pixels, whereas voxels (right) represent full three-dimensional objects and are composed of variously size cubes (right image courtesy of Geosoft (http://www.geosoft.com/)

Although it seems obvious that elevation should be displayed on the Z dimension of a three-dimensional display, it is common in scientific visualization applications to display a numeric variable on the Z dimension. For example in a meteorological application, we might display a map of temperature variation, with the temperature shown on the Z-axis. The display capabilities of GIS are generally limited to three dimensions, unless we employ colour to represent the fourth dimension, and/or make use of animation.

5.3 Digital Terrain Models

Geographical Information Systems use three different types of data structures to represent spatial data, as you may recall. The vector data structure represents features as points, lines, or polygons. The simplest of these features is the point, which is a precisely represented location in space. Lines are built from at least two points (one on each end) and may contain a number of vertices (internal points) to allow the line to change direction. Polygons are made out of points and lines, and these form an enclosed area, which is assumed to be uniform. Thus, the vector data structure is used to represent *discrete* data. The second data structure used in GIS is the raster data structure, in which features are represented using a grid of values called pixels. Rasters are used to represent *continuous* spatial data; this data does not start and stop, but varies continuously over the entire surface. The Object-oriented data structure represents features as individual objects, which have particular characteristics depending on what they represent.

Not surprisingly, digital terrain models (DTMs) can be represented using both the vector and raster data structures. The raster variant of the DTM is a grid that contains elevation values and is known as a digital elevation model (DEM)[1]. DEMs are the most commonly used type of DTM, because they are easy to understand and manipulate. However, as with all rasters, DEMs are inherently limited by the pixel size that is used. No matter how fine is the resolution of a DEM, it will always be an approximation of the underlying terrain, and furthermore, to perform analysis of a DEM, the slope must be interpolated from the cell values.

The vector variant of the DTM is the triangulated irregular network (TIN), in which precisely located points (in three dimensions) are connected by lines to form triangles. Each triangle in a TIN represents an area of uniform slope and aspect. Unlike DEMs, TINs do not have a fixed resolution. In regions of high terrain complexity, points can be clustered very closely together to represent local high and low points. In areas of uniform terrain, points can be spaced far apart, which creates large triangles having uniform slope and aspect values (Figure 52). Because the triangles in a TIN are already a mathematical representation of the underlying surface (a piecewise linear interpolation), extracting slope and aspect information from the surface is very easy.

[1] In this document, we will use the convention that a DTM is a terrain model of any form, in that a DEM is the raster variant of a DTM. Therefore DTM = (DEM,TIN). Other authors may use the terms DTM and DEM interchangeably to represent the raster representation of elevation.

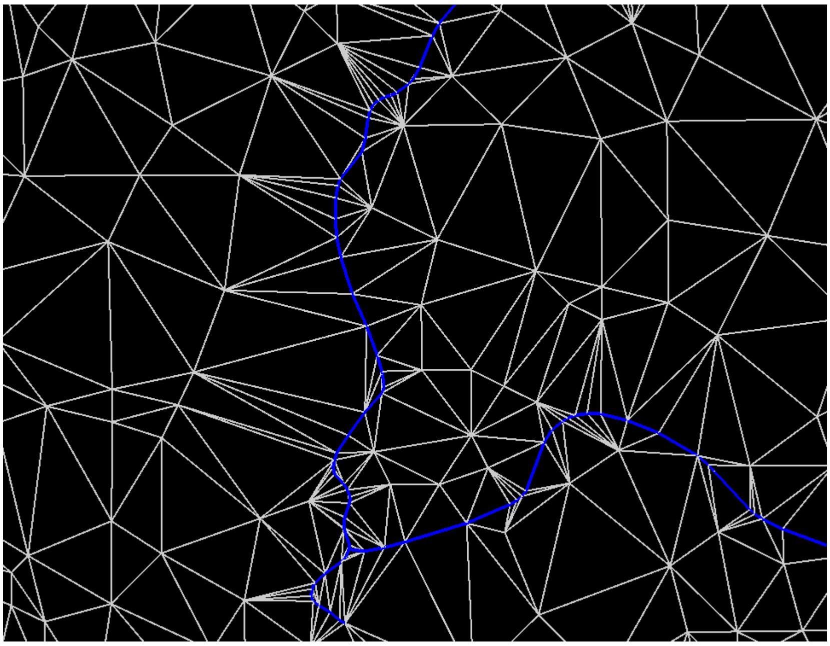


Figure 52. Close-up of a TIN showing the increase in the numbers of triangles in areas of complex terrain, such as along this river.

Depending on the type of work that you wish to do, you may choose DEMs or TINs to represent the land surface. For general analysis, involving many overlays, DEMs are a clear choice, since they are easy to manipulate like other raster layers. For precise analysis, such as in hydrology and engineering applications, TINs have a clear advantage since they are able to represent slopes much more accurately than DEMs can.

While DEMs are a very obvious application of raster technology, TINs were not developed until the 1970s, because they are conceptually more difficult to understand. In an example of parallel independent development, three groups around the same time independently invented TINs. Tom Poiker, at Simon Fraser University, came up with the concept in 1973 in a project funded by the US Office of Naval Research. Chris Gold at the University of Alberta came up with a similar idea in 1975, and Grayman and Males at Engineering-Science, a consulting firm in Virginia came up with the idea around the same time.

5.4 Construction of Digital Terrain Models

Constructing a DTM requires the selection of appropriate data, the verification of elevation values, and the construction of a DEM or a TIN. Although it is possible to convert between DEMs and TINs, doing so can lead to some interesting results. Once we have a satisfactory DTM, we may wish to export this for use by somebody else. Alternately, we may wish to simply use a DTM produced by a reputable mapping agency, in which most of the difficult work has already been done for us.

5.4.1 Sources of Data

If you are a frequent user of topographic maps, you are probably very comfortable with the interpretation of contours. Intuitively, it might seem appropriate to represent elevation data in a GIS using digitized contour. The problem with this approach is that contours are designed for the *human* interpretation of elevation. In order to perform terrain analysis, a computer must create a model of the land surface from which properties such as slope and aspect can be determined. Although it is possible to interpolate between contours to obtain this information, the representation of contours leads to flat areas along ridgelines. During interpolation, values will be interpolated from each contour to its nearest neighbour. Along ridgelines, a contour will be closest to itself, since the contour makes a sharp bend as it crosses the ridgeline. The result is that the interpolation will be flat for the portion where each contour is closer to itself than to any other contour (Figure 53). For this reason, contours are not generally used for the input or storage of elevation data in a GIS, unless no other source of elevation data is available. Contours are generally created as an output product from DTM data structures.

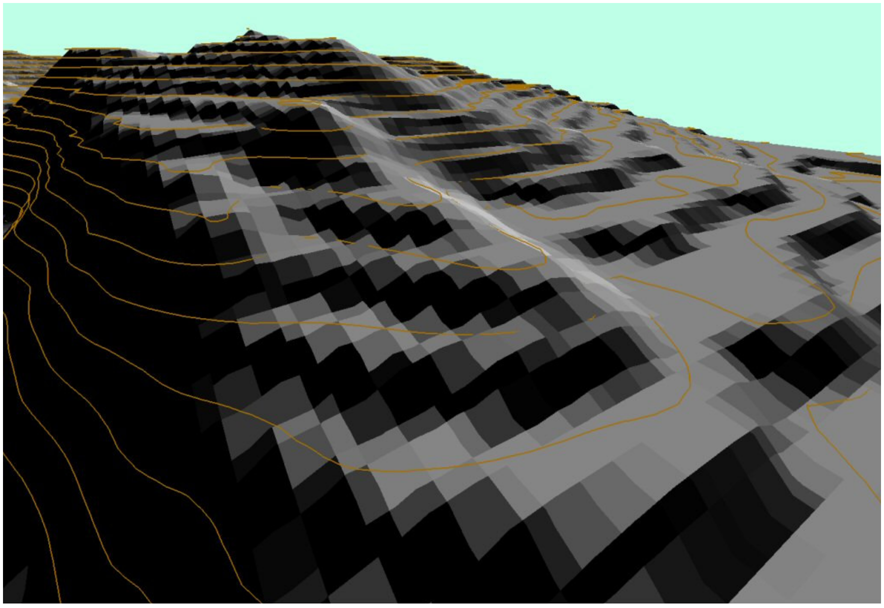


Figure 53. DEM built from contours. Notice the "stair step" appearance along the ridges.

Rather than using contours, georeferenced elevation data are used to create most DTMs. These data may simply be in the form of a table of X,Y,Z triplets, or they may have already been processed to some degree and the available in a file.

Three-dimensional data can be produced using a variety of means. The time-tested method makes use of photogrammetry to determine Z values from stereo pairs of aerial photographs. Using a stereoplotter, an operator positions a "floating dot" at significant points on the surface of a stereo pair of aerial photographs. When the enter key is pressed, a single X,Y,Z coordinate is registered. This technique, followed by manual editing, was used to collect elevation data for the forested areas of Lithuania.

For field researchers, an accurate GPS receiver is an option for recording X,Y,Z triplets over small areas. This can be an extremely difficult task, however, in areas of thick brush, tall trees, and complex terrain, so it is limited in its application, and is generally used only to augment the accuracy of digital terrain models in confined areas.

More modern techniques involve the use of radio waves or LASER beams to collect the elevation data over large areas. RADAR systems may be based in aircraft or on satellites, and can rapidly obtain elevation data over large areas at moderate resolution (10 metres vertical accuracy). These systems, which are based on interferometry, may have difficulty in areas of complex terrain, such as mountains, however (Natural Resources Canada, 2007).

The new "gold standard" in the collection of three-dimensional data is LIDAR. Using a LASER beam, the distance from an aircraft or satellite to the ground and back can be measured very accurately. Thousands of measurements can be taken every second, and airborne LIDAR is able to measure elevation to an accuracy of about 10 centimetres. LIDAR also has the ability to record multiple elevations in areas of heavy forest cover, since the top of the forest canopy will reflect some measurements, some will be reflected by the understory, and the ground itself will reflect the rest (Figure 54). This can be a very important tool for foresters, since it allows accurate measurements of tree height, and estimates of timber volume as a result. LIDAR was used to collect elevation data for urban areas in Lithuania.

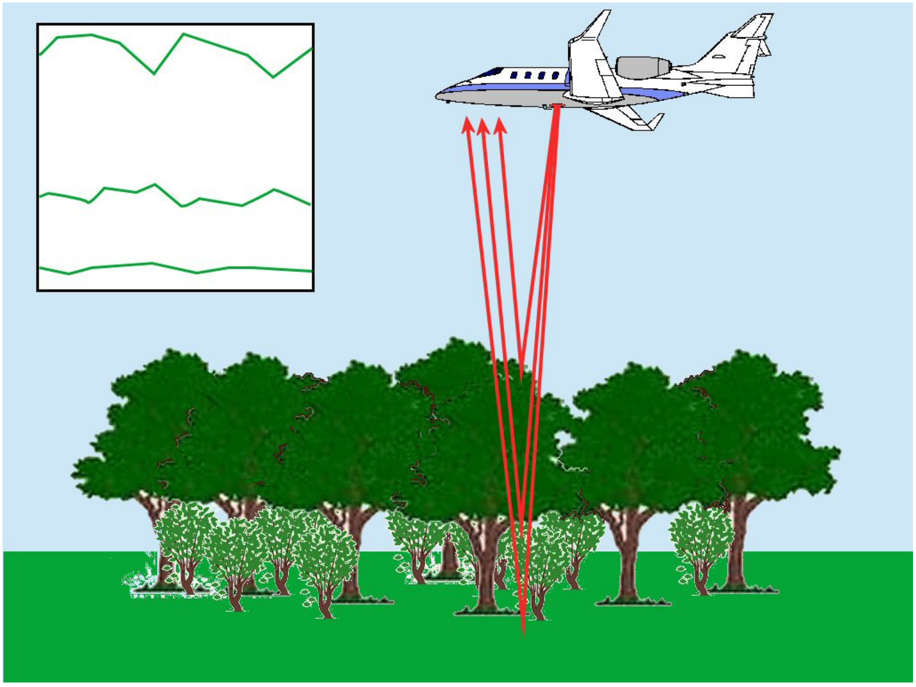


Figure 54. LASER beams from a LIDAR instrument are reflected off the canopy, understory, and ground. These three types of reflections can be separated to create three separate DEMs from the data.

5.4.2 Data Preparation

Raw X,Y,Z data may be of various levels of quality, depending on the method by which it was collected. In studies of terrain, it is important to ensure that erroneous Z values are removed from the data set, because they can affect the accuracy of a fairly large area once the DTM is created. Perhaps the most common error is for a point to be given an elevation of 0 m. In many locations, it is easy to search for these values since they are outside the possible range of elevation values. They can be identified by performing an attribute selection on the elevation column (elevation < minimum_value or elevation > maximum_value), or by building a DTM and looking for anomalous pits or peaks in the resulting surface.

Of course, the more subtle the error, the more difficult it is to detect. The best way to identify such errors is to build a preliminary DTM, and make a thorough inspection of it. The human eye is a highly sophisticated sensor, and has the ability to detect fairly subtle errors. If something does not "look right" on a preliminary DTM, checking the underlying Z values can often help to identify incorrect points.

Another type of error occurs as a result of the process of collecting X,Y,Z values using photogrammetry. Generally, a photogrammetrist will scan across a stereo model, collecting points in one direction, and when he or she reaches the edge of the stereo model, will move over a few meters and then collect data in the other direction. Some stereoplotters will actually automate the movement in the X-Y plane, allowing the operator to concentrate on collecting the Z values. The problem occurs when a steep hill occurs in the centre of the stereo model. The stereoplotter drives the motion in the X-Y plane at a constant speed, forcing the operator to play "catch up," so the Z values may be slightly low when headed uphill, and slightly high when headed downhill. On alternate scanlines, this is reversed, which results in a "staircase" appearance to hills (Figure 55). If this occurs only in a small area, then it might be possible to manually correct the data, however if it occurs over large areas, it might be better to request that the data be reprocessed from the original stereo models.

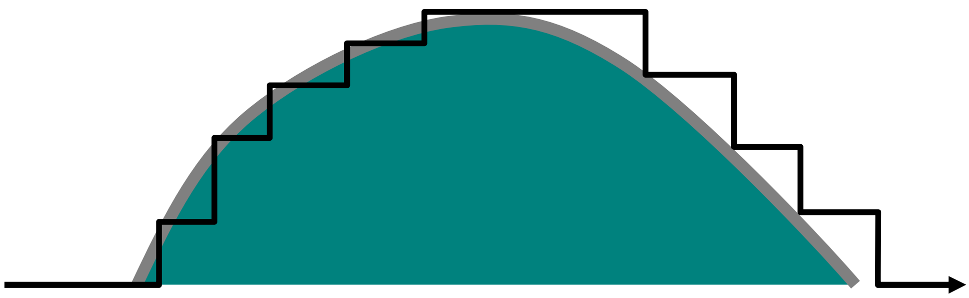


Figure 55. When encountering a hill, the stereoplotter operator must play "catch up," causing a staircase appearance to hills. Alternating lines run in different directions, so the effect is reversed every few metres along the side of the hill.

Even the most accurately collected data will not result in a perfect DTM. Consider what happens near the edges of water bodies. It is extremely unlikely that a point will be collected exactly along the shoreline. What happens is that points are collected on land, or they are collected on water. When the DTM is created from this source data, elevation pixels are interpolated from the land elevation to the water elevation, leading to a smooth transition between land and water, not a definite shoreline. This is most noticeable on narrow lakes, where the lake takes on the shape of a trough, instead of being flat. To resolve this problem, we generally add "breaklines," which are lines composed of three-dimensional points that represent significant breaks of slope, such as shorelines, the tops of ridges, streambeds, and the edges of road embankments. Photogrammetrists collect breaklines after they have collected the mass points that are used to generate the bulk of the model.

When DTMs are going to be used for hydrological analysis, it is even more critical that it be accurate. If we are simulating the flow of water across a DTM from source to outlet, then it is critical that the water is able to flow downhill along the streamline without interruption. In other words, there must be a constant decrease in elevation along the river. Flat areas are only permissible for lakes and the ocean. Although it is possible to collect the elevations of the points that make up a DTM with great accuracy, even minor variations in elevation (a few centimetres) can lead to the creation of "lakes" that do not really exist. The solution to this problem is a post-processing step to smooth streambeds to ensure that they always permit downhill flow (i.e. they are monotonic).

Once the DTM data has been examined and corrected, it will be possible to create a reasonably accurate model of the Earth's surface. As be discussed earlier, this can be done using two methods, Digital elevation models, the raster method, and triangulated irregular networks, the vector method (Figure 56). The construction of each of these types of DTM is quite different, so we will discuss the details separately.

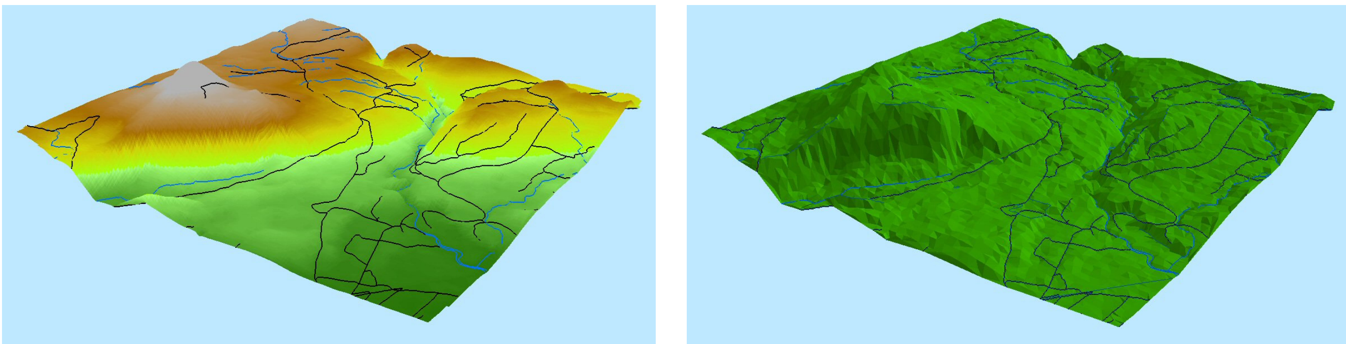


Figure 56. A DEM (left) and a TIN (right) representation of the same location

5.4.3 Digital Elevation Models

As we saw in Section 5.3, digital elevation models (DEMs) are rasters of elevation values, which force a generalization of the underlying surface. Depending on the resolution of the DEM, there may be no measurements, one measurement, or many measurements underlying a particular pixel. The process of converting from input points to a DEM requires interpolation.

Interpolation is the process of predicting a value at a point where no data is available. By using the surrounding data values, we can estimate what the value will be. Many different techniques exist for creating a DEM from underlying points. We can group these techniques along three axes: local versus global, exact versus inexact, and deterministic versus stochastic.

Properties of Interpolators

Local interpolators make use of surrounding data points within the neighbourhood when predicting a value. The operator may be able to define the number of points used in making the calculation, or it may be completely automatic, depending on the technique used. Global interpolators, on the other hand, make use of all the points in the data set. A local polynomial interpolator takes a number of points in a local area and uses it to create a small surface representing that area. This process is repeated over the entire data set, until a "patchwork quilt" has been created from many local interpolations. The global polynomial interpolator uses all of the points in the data set to create a single surface for the entire data set. Since all points are represented, this type of interpolator tends to smooth out local variations in terrain.

Exact interpolators ensure that the surface created that is used to make a prediction passes exactly through all of the data values in the neighbourhood (local) or in the entire data set (global). Inexact interpolators do not require the surface to pass exactly through all of the data points. This enables the interpolator to produce a smoother surface than would ordinarily be possible due to local variation in the data (Figure 57).

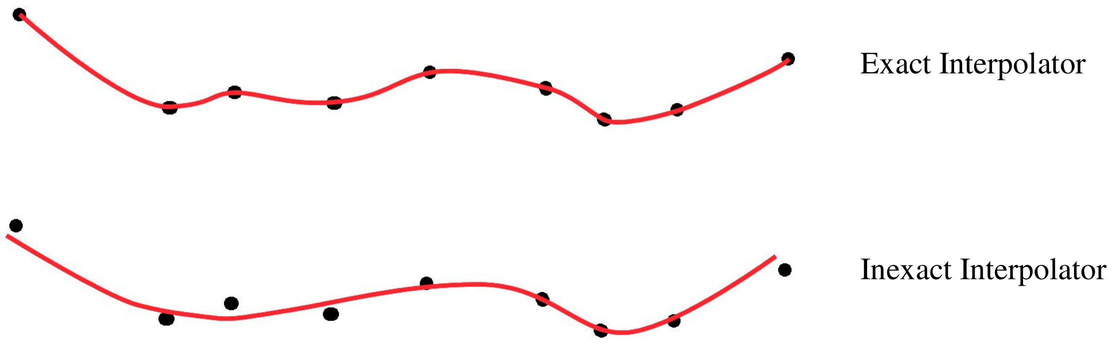


Figure 57. Profiles of DEMs produced using Exact Interpolators (top) and Inexact Interpolators (bottom). Notice that inexact interpolators do not pass through all points, but provide a better overall representation of the trends in the terrain.

Deterministic interpolators use a mathematical formula, such as a weighted average of neighbourhood values, to make the prediction. Stochastic interpolators make use of probability model to make the prediction. The main difference between these two techniques is that stochastic interpolators provide an assessment of the accuracy level for each interpolated value; this is not available for deterministic interpolators.

In Appendix A, there is a detailed discussion of interpolators. Section A.1 discusses local interpolators and Section A.2 discusses global interpolators. Table 4 summarizes the Interpolation Techniques by their properties.

Technique	Local/Global	Exact/Inexact	Deterministic/Stochastic
Natural Neighbours	Local	Inexact	Deterministic
Inverse Distance Weighting	Local	Inexact	Deterministic
Kriging	Local	Inexact	Stochastic
Splines	Local	Exact	Deterministic
Local Polynomial	Local	Inexact	Deterministic
Global Polynomial	Global	Inexact	Deterministic
Topo to Raster[2]	Global	Inexact	Deterministic

Table 4. Properties of Interpolation Techniques

[2] This is not really a specific interpolation technique, but an implementation of the ANUDEM algorithm which allows multiple elevation data sets to be combined to create a single, hydrologically correct raster

Reduced Resolution Datasets

Reduced resolution datasets (RRDs), also known as "pyramids," are a technique for displaying rasters more rapidly. The key idea behind RRDs, is that there is no point in processing high-resolution data that cannot be displayed on a computer monitor. Many rasters, when displayed as a small-scale, have millions more pixels and can be displayed on the average computer display. It is much faster to display a generalized version of the raster containing fewer pixels when you are zoomed out. Only when you are looking at a small portion of the raster does it make sense to display the raster at full resolution. For this reason, when a new raster is imported, the GIS may spend some time creating a number of reduced resolution raster files for optimal display at different viewing scales (in ArcGIS, these files are stored with a ".rrd" extension).

5.4.4 Triangulated Irregular Networks

Triangulated irregular networks, commonly known as TINs, are a vector technique for representing elevation data in a GIS. Unlike DEMs, which are raster-based, and use many of the same techniques that are used for analysis of rasters that do not contain elevation information, TINs were designed specifically for the display of elevation data, and so are more specialized and less well known than DEMs, except among certain disciplines (hydrology and engineering) where they are used extensively. Despite this, TINs play a very important role in GIS, since they are the most accurate way of depicting terrain that has been devised.

The reason that TINs are the preferred way for accurately displaying terrain is a result of the way that they are constructed. In a TIN, there is no interpolation of the input data. A selection of the input data in its raw form is used to generate the surface.

Not all of the input points are used in the construction of the TIN. The reason for this is that TINs exhibit undesirable behaviour when they are composed of long, thin triangles. Thin triangles tend to occur when there are clusters of points in areas of low point density. The problem with long, thin triangles is that one triangle may have strikingly different properties from its neighbour, which may result in inconsistent results being produced in the area. For example, it is quite easy for one triangle to have a slope of 30°, and the neighbour in triangle to have a slope of only 15°, when both triangles are quite small. Such a situation may occur when three points lie in close proximity, with one point having an elevation which is different than the other two (on the order of 1-2 m). In general, this is an undesirable situation, because the variation shown is not real, but is a product of "noise" in the data set. Removing some of the points enlarges the triangles, reduces the variation in the triangle properties, and results in a better overall TIN.

There is one case where long, thin triangles are appropriate, which is along river banks. Because rivers are often highly incised, there will be a breakline along the line where the river incision begins, and the river bed will be represented by another breakline. Because these breaklines are dense sources of elevation points that are used in TIN construction, it is common for long, thin triangles to be created along the river banks. This permits the river to meander in its channel, and be represented with accuracy in the TIN.

There are two commonly used algorithms for the construction of TINs. These algorithms choose points to produce TINs with well-formed triangles. The first of these techniques is called Delaunay triangulation. Delaunay triangulation selects points from the input data set one at a time and adds them to the triangulation. Once three points have been added, the first triangle will be formed. The Delaunay triangulation algorithm then calculates an imaginary circle on which all three points lie (a "circumcircle"), and then rejects any further points that fall within that circle. As each new point is added into the triangulation, it forms a new triangle, using two of the points from existing triangles.

A new circumcircle is created, and points may not be added within its boundaries. This ensures that long, thin triangles cannot be created (Figure 58).

One limitation of the Delaunay triangulation algorithm is that it only selects points based on their X and Y coordinates; no consideration of the Z Coordinate is made when accepting or rejecting a point for use in the triangulation. What if the Z Coordinate is significant? Even if the Z Coordinate represents the top of a hill, or the lowest point in a surface, or even if it lies along a break of slope, it may be eliminated by the Delaunay triangulation algorithm. This would result in a TIN that has significant features "smoothed out" of the surface.

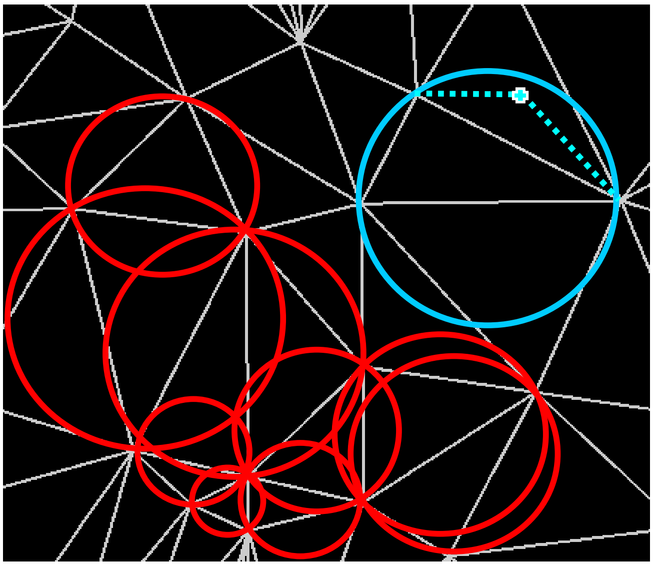


Figure 58. The blue point will be eliminated in this Delaunay triangulation, because it falls within the blue circle. This prevents the formation of a long, thin triangle (blue dashed lines).

To resolve this problem, a second algorithm, called the very important points (VIP) algorithm has been developed. The VIP algorithm uses this same certain circle method as does Delaunay triangulation, but before a point is eliminated, it's value is compared with a elevation value created by interpolating the existing points. Based on a threshold value, if the elevation is different than the expected elevation, then it is preserved in the output triangulation (Figure 59). Otherwise, the point is eliminated as it is in the standard Delaunay triangulation. The VIP algorithm ensures that the triangles that make up a TIN represent areas of uniform slope, at least as far as can be determined using the input data.

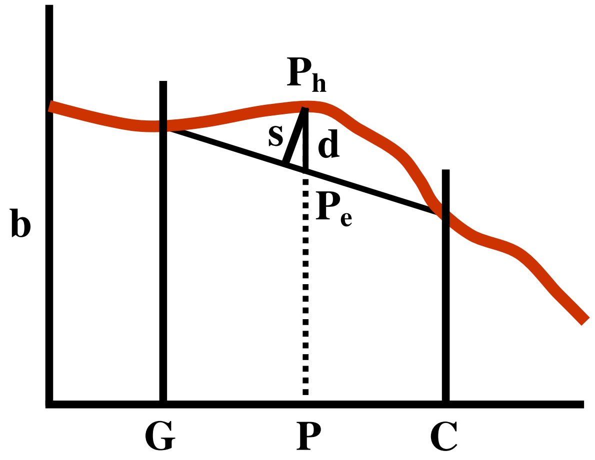


Figure 59. The very important points (VIP) algorithm overrides Delaunay triangulation when the actual elevation of a new point (Ph) exceeds the expected height of a new point (Pe) by more than a threshold value (d).

When TINs are analyzed, we must account for a "drop-off" effect at the edge of the model (Figure 60). Frequently when data is collected for TIN construction, extra data is included in a small buffer outside the normal map boundary in order to compensate for this effect. The result can then be clipped to the proper map area, removing the erroneous data from the edge of the model.

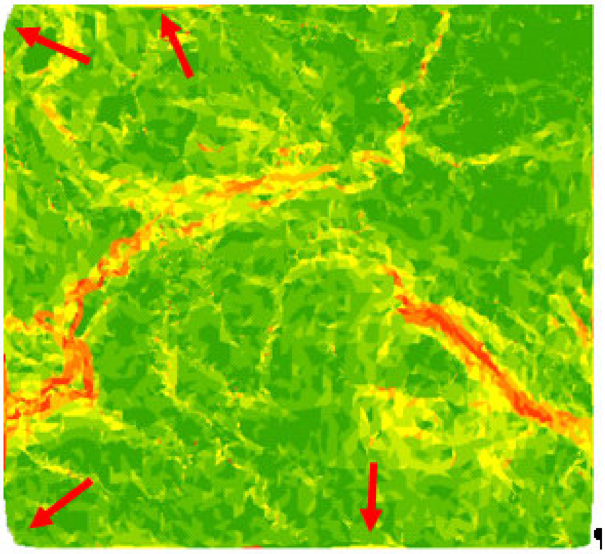


Figure 60. "Drop off" at the edge of a TIN (arrows) needs to be accounted for when analyzing the results.

5.4.5 Terrains

In Section 0, we discussed the use of Reduced Resolution Data sets to increase the speed of the DTM display, depending on the scale at which it is being viewed. We also discussed how certain interpolation techniques are able to produce DEMs that are very generalized, whereas others produce highly detailed models of the Earth's surface. Not surprisingly, some of these ideas also apply to TINs. We discussed how TINs are created on a point-by-point basis using either Delaunay or the very important points algorithm. In fact, it is computationally very easy to add and delete points from a TIN.

The idea behind a terrain is that a TIN can be created that varies its resolution depending on the scale at which it is viewed (Figure 61). Points can be added or removed as necessary so that the TIN increases its resolution as you zoom in. This process continues, until all possible points in the input data set have been used to create a TIN (subject to the limits imposed by the VIP algorithm). This allows TINs to be drawn rapidly at all scales, and appropriate levels of detail to be displayed depending on the scale at which analysis is performed.

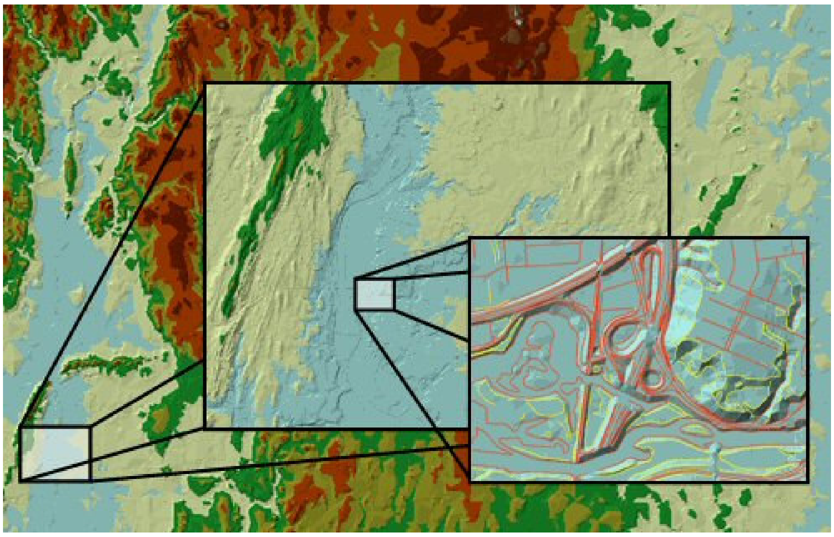


Figure 61. By adding or removing points from the TIN at different scales, terrains allow for variable resolution. The number of points in each window is approximately the same (higher resolution, but smaller area) (Source: ESRI).

Terrains also allow the creation of models with irregular boundaries, and with holes in the middle. This is useful, because areas with no data are often covered with large triangles in a traditional TIN model. In areas where there is a concave edge, triangles would normally extend from one part of the edge across empty space to another part of the edge, filling the concavity with triangles, despite the fact that we have no data for it (Figure 62).

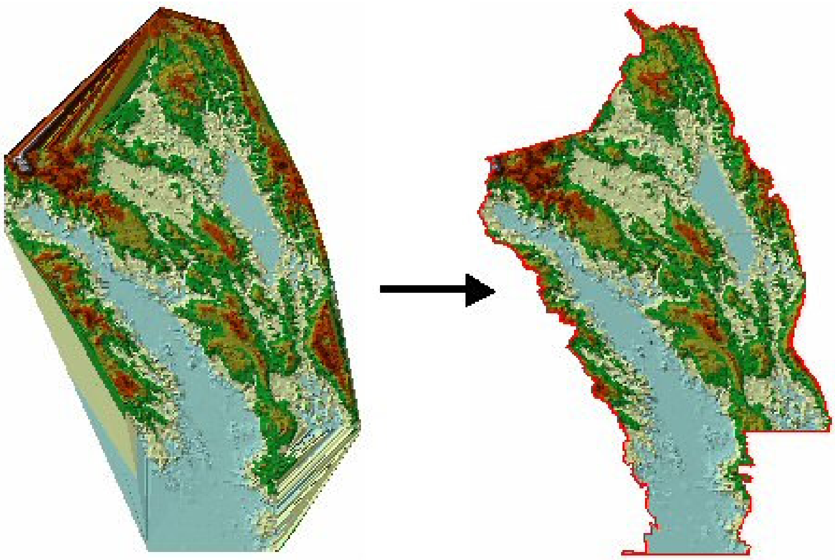


Figure 62. Clipping polygons can be used to remove long, thin triangles from the concave areas in the edges of terrains (Source: ESRI).

Another interesting feature is that "replace polygons" can be used to set all triangles to uniform elevations along lakes and other bodies of water. The replace polygons override all elevation values within them to create a uniform, flat surface.

5.4.6 Converting Between Surface Representations

No method for representing surface is perfect, and the limitations of each technique can sometimes be seen when terrain data is converted from one type of surface representation to another. When data is converted from one terrain representation to another, the limitations of the first terrain representation are reflected in the second, so there is propagation and compounding of error. In an ideal world, rather than convert between a DEM and a TIN or vice versa, it would be better to create the TIN from scratch using the original data that was used to create DEM. However, the source data is not always available, so this may force operators to convert from one surface representation to another.

5.4.7 Data Exchange Formats

There are data exchange formats available to transfer both raster and vector data (points, lines, or completed TINS). A majority of these formats are "de facto" standards, although there are a number of government standards that are widely used as well.

Unfortunately, the current state of data exchange formats is a bit messy. There are hundreds of data exchange formats available, with the vast majority of functions being duplicated, but each standard having some specialty niche where it is indispensable. This can be attributed to the following factors:

- Technology is rapidly changing, and new methods for representing surfaces are under constant development
- GIS vendors are going bankrupt, or are being bought out, leaving behind orphaned standards that are no longer supported.
- Because of the wide variety of techniques being used, it is difficult to create a "one size fits all" data exchange standard. Governmental efforts to do so result in unwieldy standards that only add to the number of standards that must be maintained.

In Appendix B, there is a list of data exchange formats that are applicable to the transfer of completed DEMs or TINs, as well as the points and lines that might be used to build them.

5.5 Digital Terrain Model Analysis and Applications

Digital terrain models give us the ability to model real-world physical processes that are a product of the three-dimensional nature of the Earth's surface. If, for example, we were trying to model areas that are susceptible to landslides, we could use a DTM to determine where the steepest slopes occur, which areas might be most susceptible to freeze/thaw cycles in the winter based on solar illumination, and the amount of water that is found in the soil at a particular point. All three of these physical characteristics are strongly affected by the shape and orientation of surfaces. Similarly, if we were interested in predicting which areas are most susceptible to flooding, we could model the surface to determine elevation, and the amount of protection afforded by dykes.

In the previous module, we discussed how the construction of a network is a necessary prerequisite for analytical activities, such as finding the shortest path between two points, or tracing the flow of a commodity to a utility network. Similarly, the construction of a DEM or TIN sets up an environment in which analysis can be performed. The analysis of surface characteristics is essentially the same for TINs, and DEMs. There are minor differences in how the data are organized and processed, but this is done internally by the analysis software, and is hidden from your view. Therefore, we will discuss these techniques as if there were no difference between the two types of DTM, but we will note the exceptions and expected differences in results when they occur.

By default, DEMs appear as a colour shaded raster with the colours representing the elevation values and TINs appear as a hillshaded series of triangles (refer back to Figure 56). Software can be used to analyze these default forms to create types of data that are useful for mapping and further analysis.

5.5.1 Visualization

Perhaps the most obvious application of digital terrain models is to visualize unfamiliar terrain. This can be immensely helpful tool when trying to communicate information about a place to people who have never been there. In its simplest form, a three-dimensional model can help people understand terrain, but because this is the computer-based model, we can add statistical information, overlay satellite pour aerial photographs on top of the model, or we can alter the very terrain itself to show the effects of open-pit mining or forestry.

Graphics quality varies greatly between packages. Although most GISs have rudimentary terrain modeling and display capabilities, specialized packages take the degree of realism much further. Developed for the video game industry, products such as World Construction Set (http://3dnature.com/) have allowed new degrees of realism in terrain modeling (Figure 63). GIS information can be imported into this package and can be used to produce images of the Earth's surface as it is, as it could be, or as it never was!

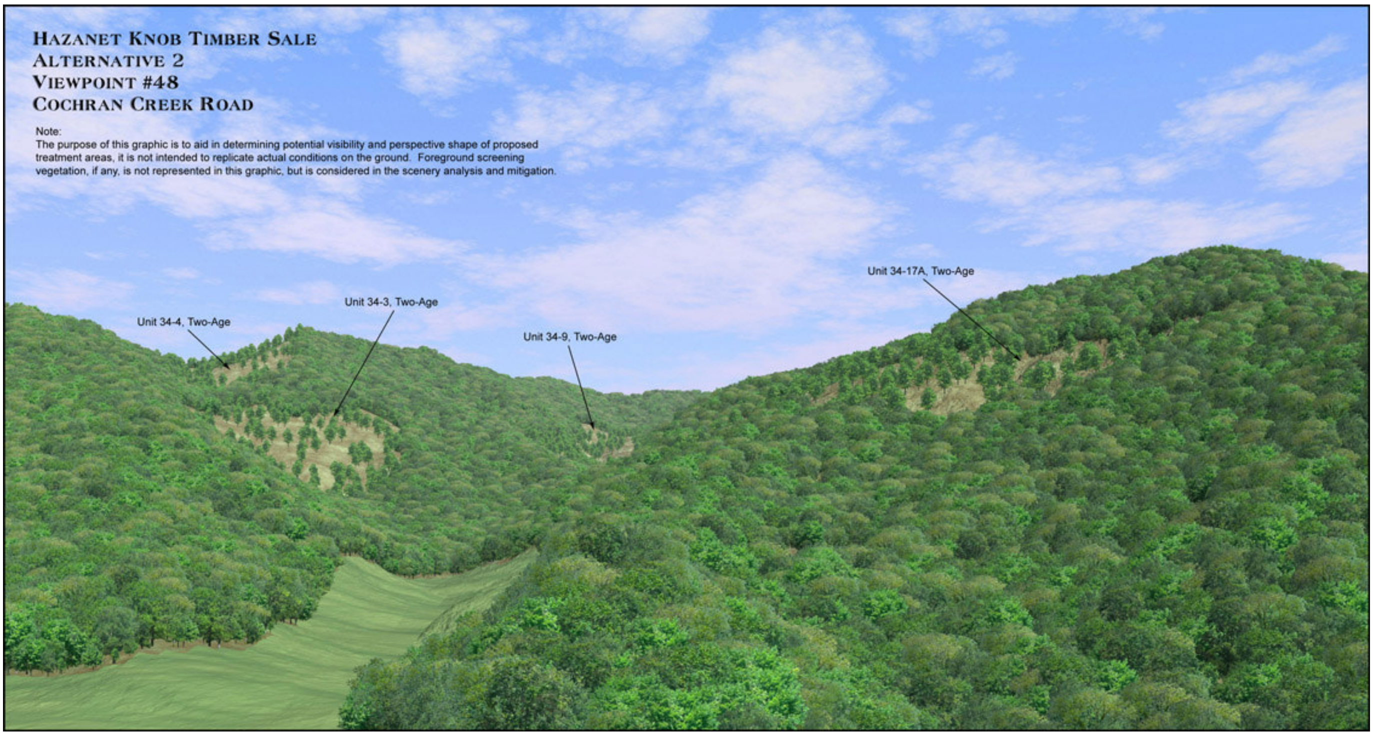


Figure 63. World Construction Set and other photorealistic rendering software can produce very realistic images from GIS data (Source: 3D Nature, LLC http://www.3dnature.com).

As with any technology, three-dimensional visualization techniques can be used for both beneficial and harmful purposes. Consider the siting of an industrial facility. If a number of sites that are in close proximity to population centres have been shortlisted, using visualization tools to identify which site is least visible is clearly an ethical use of the technology. However, it is not difficult to move into an ethically ambiguous position, by choosing sites first based on their visibility, and then determining where the facility should be located so that it is least visible. If visibility is the main siting criteria, this is probably an acceptable action, but if other criteria are more important, such as safety or environmental appropriateness, then this approach is somewhat unethical. Now consider a third option, which is to use the visualization tool for propaganda purposes. Suppose that a number of sites have been chosen, and all are at least partially visible from the urban area. To solve this problem, the GIS operator edits the elevation of a number of points and rebuilds the DTM so that a hill becomes a few meters higher, obscuring the view of the industrial facility. Here, the visualization tool is being used to misinform people in a way that is very convincing, which is clearly unethical.

A number of tools have recently become available to allow DTMs to be viewed on a computer as three-dimensional models. Perhaps the most well known of these is Google Earth, which presents satellite images draped on DTMs to show a virtual model of the entire planet. Desktop versions of software such as this also exist, such as the NASA World Wind project, and ESRI's ArcScene (Figure 64) and ArcGlobe products. Some of these products allow you to plot a flight path over the DTM and produce an animation of the resulting flyover.

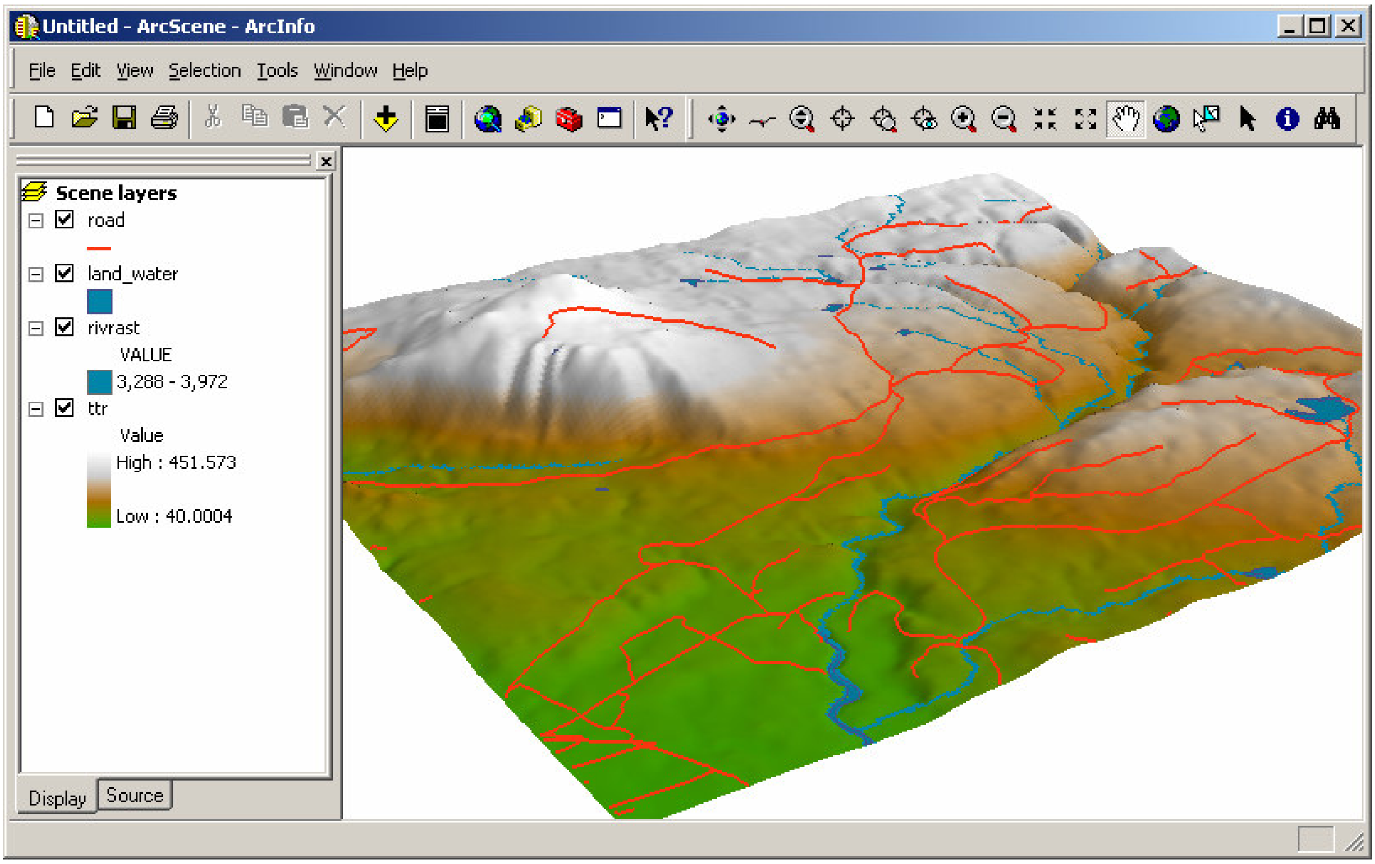


Figure 64. ArcScene is a software tool for viewing DTMs in three dimensions

5.5.2 Slope

From a human point of view, possibly the most important piece of information they can be derived from a surface is the slope. Mathematically defined as the amount of rise (vertical distance) over the amount of run (horizontal distance), slope is an expression of the amount of potential energy that is available in a surface. Because the slope of a surface acts to modulate the force of gravity, the steeper the slope, the greater the effective force of gravity on any particle found on that surface. This means that as slopes get steeper, the amount of friction that is required to keep a particle in place increases. Steep slopes often are bare of soil for this reason, and only solid rock may be present, since it is most resistant to the effects of gravity.

Slope is calculated by determining a vector that is normal (perpendicular) to the surface. This vector provides the compass bearing of the line of steepest descent; whose angle relative to the XY-plane represents the slope.

Maps of slope (Figure 65) can be very useful for the analysis of landslide and avalanche risk, for assessing the difficulty of travel and construction. Slope values can be expressed as either a percentage or as degrees. The layer that is created through slope analysis can be used in more sophisticated analyses of surfaces, such as in the Revised Universal Soil Loss Equation (RUSLE), which is used for determining the amount of soil loss from agricultural fields.

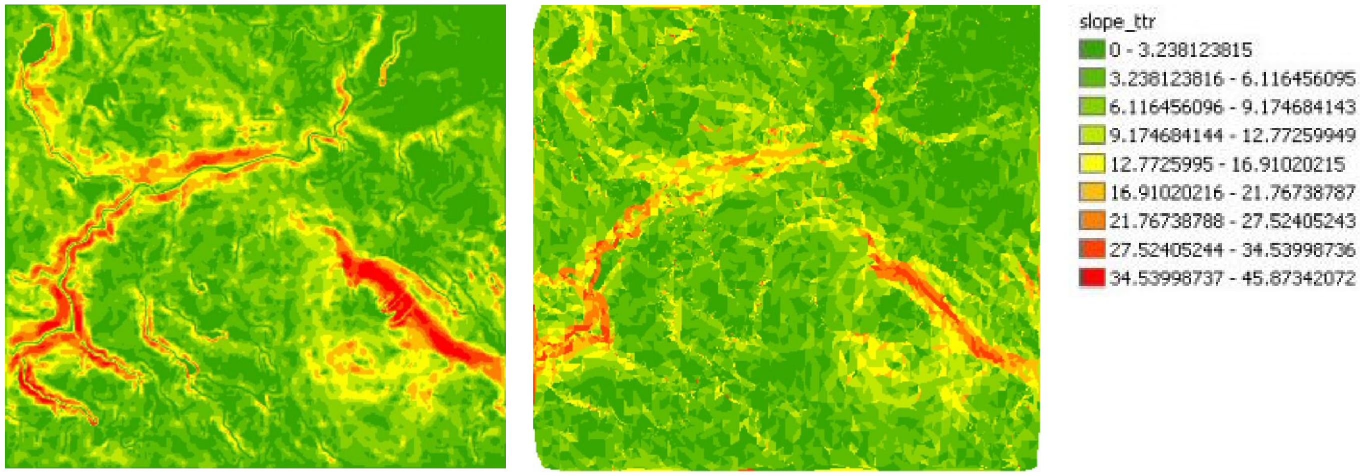


Figure 65. Slope calculations on a DEM (left) and on a TIN (right). Notice the difference in the amount of high slope (red) between the two techniques[3]. Slope classes are in percentages.

5.5.3 Aspect

Slope is an expression of the orientation of a surface into Z-plane, and aspect is an expression of the surface orientation in the XY-plane. In other words, aspect expresses the direction of steepest descent on a surface. When the sun is perpendicular to the surface in both the X-Y and the Z-plane, the surface receives the maximum amount of solar radiation possible. Thus, the aspect is important factor in determining the amount of illumination a particular surface receives, which has a direct impact on the vegetation that can grow there.

Aspect is calculated by determining the compass bearing of the vector that is normal (perpendicular) to the surface.

Maps of aspect (Figure 66) are not particularly useful on their own, however the aspect layer is very useful in the modeling of vegetation and wildlife habitat. Slope and aspect may be combined with calculations of the Sun's location on any particular day, to determine instantaneous solar illumination, and the amount of illumination over time.

[3]The differences in areas of steep slopes shown may have ethical implications. If forestry is banned on steep slopes because of the risk of landslides or soil erosion, using a TIN representation instead of a DEM may significantly increase the amount of area that can be harvested.

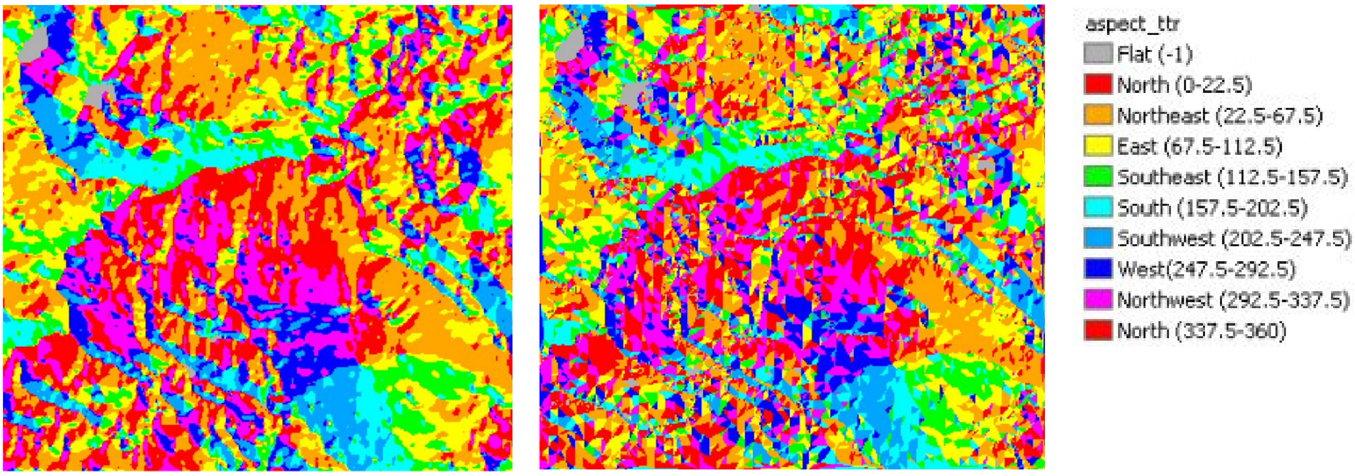


Figure 66. Aspect calculations on a DEM (left) and on a TIN (right). Units are in degrees.

5.5.4 Solar Radiation

Because it is fairly difficult to combine slope, aspect, and a model of the Sun's movement over time, these tools may be combined into a single command to model solar radiation.

Solar radiation modeling computes total, direct, and diffuse solar radiation (that is, the light that is refracted by the sky) in modeling (Figure 67). Also, the hours of illumination are computed (Figure 68). Radiation values are measured in units of Watt-hours per square metre.

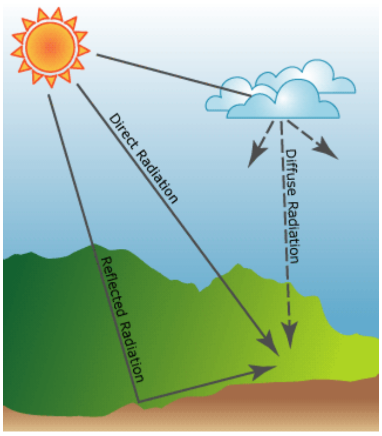


Figure 67. Direct, Diffuse, and Reflected Solar Radiation (Source: ESRI).

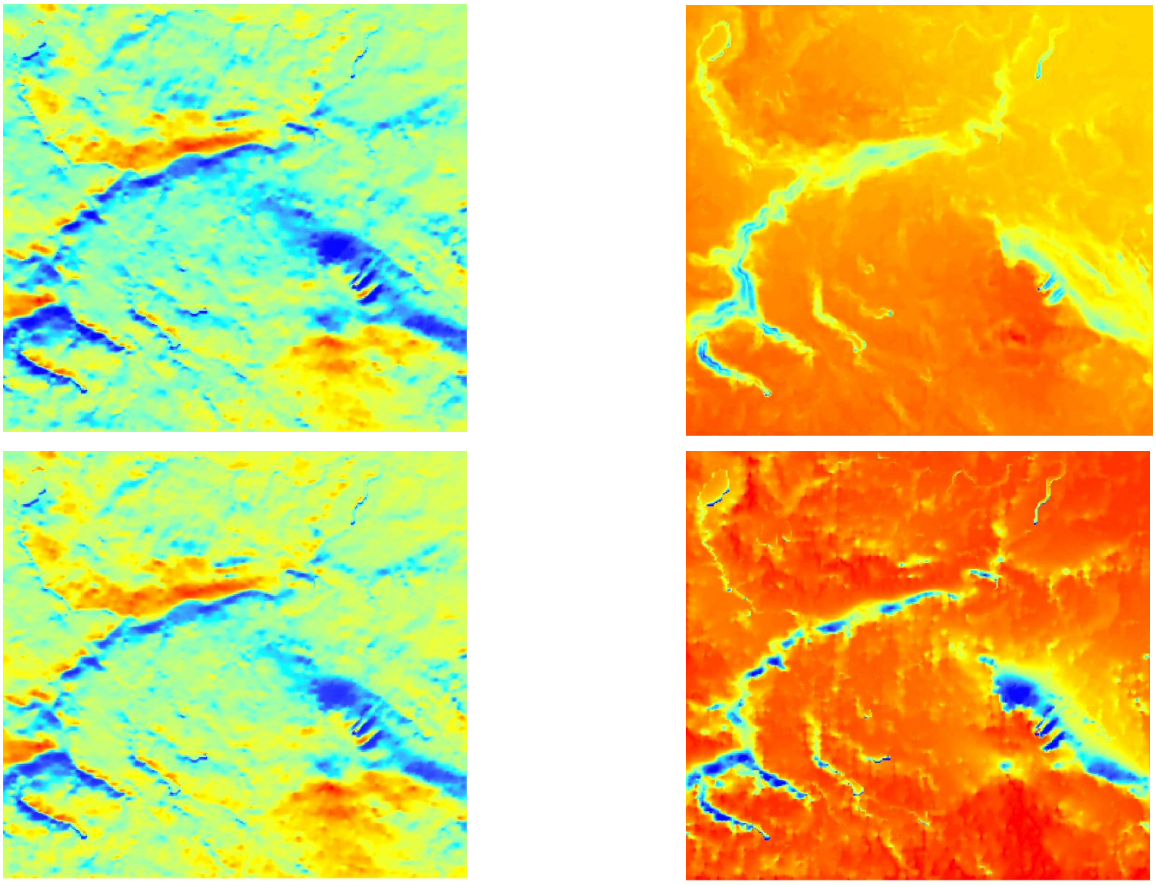


Figure 68. Solar illumination calculations for January 1 to February 28, 2007. Direct solar heating (upper left), diffuse solar heating (upper right), total solar heating (lower left), and cumulative hours of illumination (assuming no clouds, lower right).

The calculation works by determining the instantaneous amount of direct and diffuse solar radiation at a number of times (say once per hour), measured in watts per square metre. These are multiplied by the amount of time that the sun shines during each time slice to calculate watt-hours per square metre. The energy calculations for each time slice are then summed to calculate an overall amount of energy for the time period specified. In ArcGIS, this command operates only on DEMs, not on TINs.

Modeling solar radiation on a surface can help in studies of wildlife habitat (amount of food available, energy expenditure to stay warm), agriculture and vegetation (amount of solar radiation available for plant growth, potential for frost), building heating requirements (number of heating-degree days), forest fire risk (amount of evaporation), and hydrology (amount of evaporation and snowmelt).

5.5.5 Hillshading

Calculating solar radiation is highly analytical, and is very useful as an input for physical models, but a simpler version can calculate the amount of illumination on any particular area of slope based on the position of the sun. For more realistic hillshading, shadows from nearby hills can be modelled (Figure 69). This produces a surface showing shaded hills, which is an excellent tool for depicting terrain on topographic maps.

For the given sun azimuth and elevation, this calculation determines the orientation of the slope relative to the sun. If the slope points directly towards the sun, it is given a high value (255), and if the slope points directly away from the sun, it is given a low value (0). Slopes in between these extremes are given a value based on the orientation of the slope relative to the orientation of the sun to produce the appearance of illumination. If shadows are being modelled, those portions of slope that lie behind obstructing terrain have their illumination values reduced.

Psychological studies have shown that humans understand terrain better when the solar illumination is coming from the Northwest. This is opposite (at least in the northern hemisphere) to the southerly position where the sun normally appears on aerial photographs. This may be because we are accustomed to viewing three-dimensional objects with the source of illumination in front of, and to the side of us, rather than behind us. If the source of illumination were behind us, then the object being studied would be in shadow.

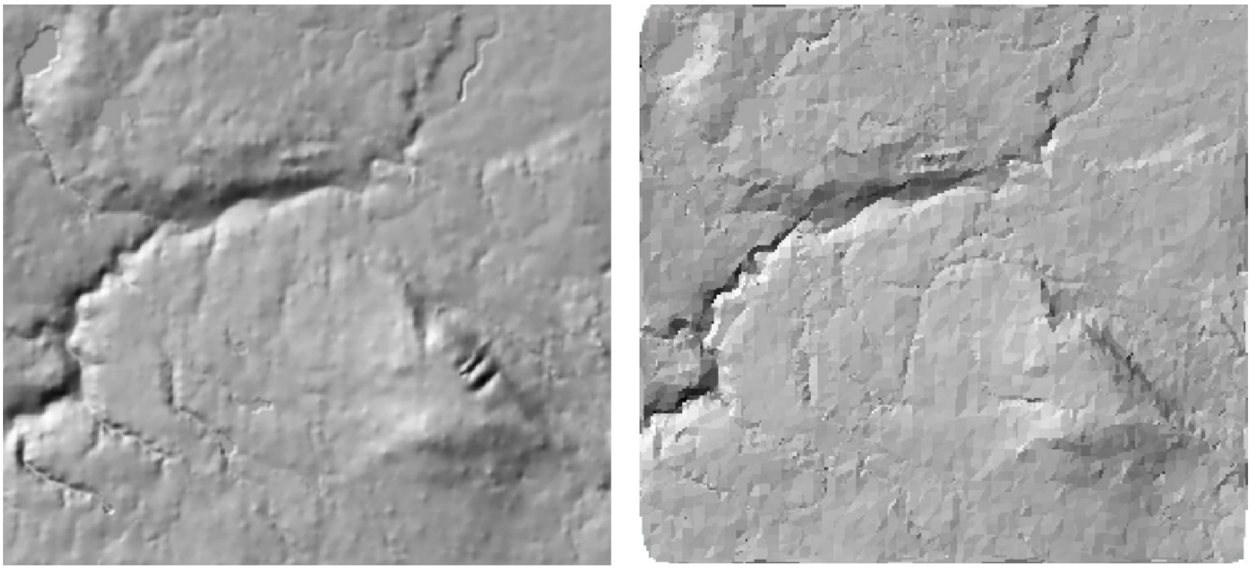


Figure 69. Hillshading on a DEM (left) and on a TIN (right). Both sources of illumination are from the NW at a 45-degree angle. Notice the greater amount of detail visible on the TIN.

5.5.6 Viewsheds

If it is possible to model where the sun is visible on the surface using hillshading, then it should be possible to model what areas are visible from a point on the surface. This is called viewshed, or inter-visibility analysis (the word "viewshed" is a play on the word "watershed"). The amount of the surface that you can view from a particular point is dependent upon your height above the surface. When you are high is located near the ground, you can see very little; conversely, when you are in an aircraft overhead, you can see almost the entire surface.

This command requires a layer showing the location(s) of observers. Each point can be customized by adding attributes to represent parameters such as its height above the ground (OF1), the height that other viewpoints must be in order to be seen (in order to lie above vegetation) (OF2), the start angle and end azimuth of the viewing cone (AZ1 and AZ2), the top and bottom angle of the viewing cone (V1 and V2), the innermost and outermost distance than an object can be viewed (R1 and R2). These values (except for OF1) act to restrict the amount of terrain that can be seen by the observer, so if they are omitted, it is assumed that the observer can see everything. In most cases, it is wise to set the R2 value to something large but reasonable if you are working with a large DTM, to account for the limitations caused by atmospheric haze.

This command works by modeling lines of sight from the observation point to all pixels on the surface of the DEM. Each pixel's X, Y, and Z position is translated into an altitude and azimuth value relative to the observer. Moving outward from the observer, each pixel is analyzed in turn, and if its altitude and azimuth have not already been assigned to another pixel, it is marked as "viewable." As pixels are examined at greater distances from the observer, it is increasingly unlikely that they will be visible because of obscuring terrain.

Viewshed analysis creates a layer showing all areas that are visible from one or more viewing points (Figure 70). We can use this to determine whether industrial facilities or areas of forest harvesting are visible from particular roads. The location of observation posts can also be optimized using this technique as well.

This type of analysis is very important for any sort of application that uses visible light, or radio waves that travel by "line of sight," such as microwaves, and high frequency radio waves. Many modern radio devices use line of sight transmission, including cellular telephones, various types of radio, Global Positioning System receivers, wireless computer communication devices, and some satellite communication devices.

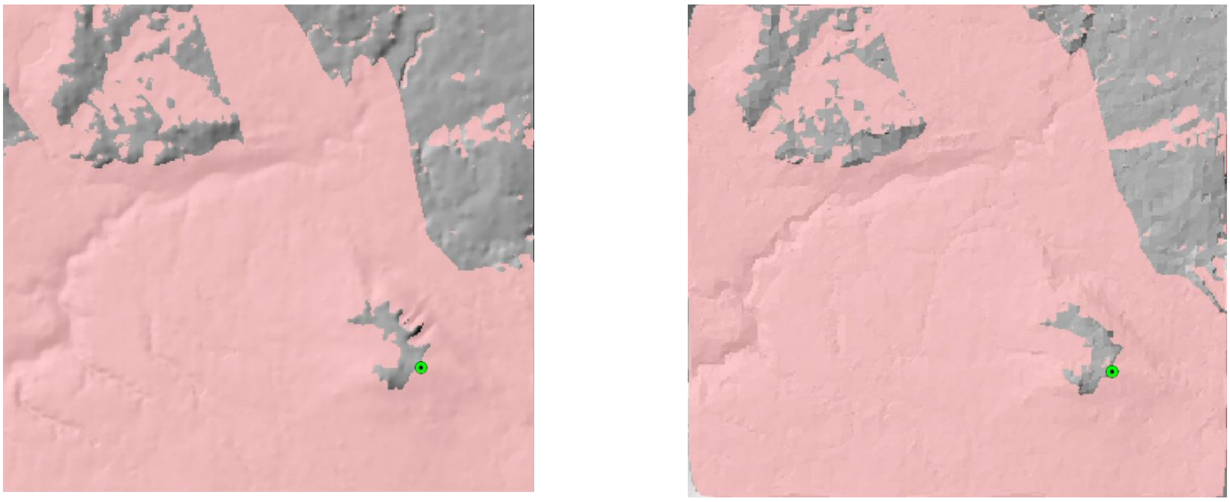


Figure 70. Viewshed analysis on a DEM (left) and on a TIN (right). The unshaded area can be seen, and the red shaded area cannot. The viewer (green dot) is assumed to be 2m high, and a 10 km limit is set on the maximum viewable distance. Not surprisingly, providers of cellular telephone service have been strong users of this technology, since it allows areas of reception of particular cell phone towers to be modeled. For larger DTMs, the amount of the Earth's curvature can be calculated so that locations over the horizon cannot be seen.

5.5.7 Contours

Although DEMs and TINs are the primary means by which terrain is represented in a GIS, we can also use these to generate contours for the production of topographic maps. Contours can be generated at any interval, with a fixed base contour elevation; this can be used to generate index and intermediate contours. This is the technique that will be used to generate contours for the new Lithuanian 1:50,000 topographic maps.

When TINs are used to create contours, the contours will sometimes appear to be "unnatural," since the edges of the triangles may become incorporated into the contour boundary, causing triangular shapes to be formed. Similarly, contours created from DEMs may show evidence of pixel boundaries (Figure 71). New algorithms, however, have greatly improved the quality of contours that are generated from TINs and DEMs. However, for serious cartographic use, such as the publication of topographic maps, some editing of computer-generated contours will be required to ensure that they are of "cartographic" quality.

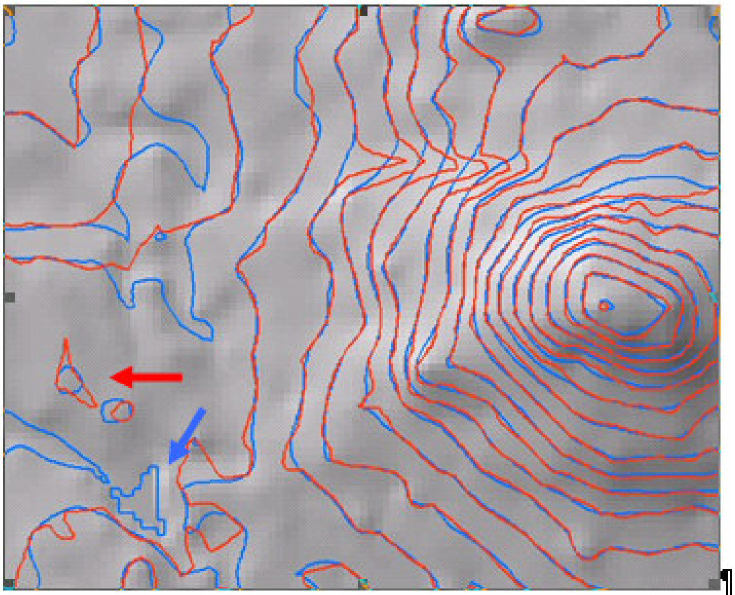


Figure 71. Contours generated from a TIN (red) and a DEM (Blue). Notice the triangular contour shown with the TIN contours (red arrow), and the pixel boundaries showing up in the DEM contours (blue arrow). Background hillshading is based on DEM.

5.5.8 Volume

People who analyze digital terrain models may be interested in the volume of subsurface features. It is relatively easy to determine the volume of material beneath a modelled surface using a GIS. If we look at our surface as a continuous, curving form, then we must use calculus to determine integrals to find the volume below the curve. Fortunately, however, we do not store our surfaces using splines (although we may use the splines to generate a DEM), so we can use simpler mathematics to determine the volume beneath a DEM or a TIN.

The TIN presents the easiest case. Each triangle has a particular orientation, so we have a sloping surface, which can be thought of as an irregular pyramid lying on top of a triangular prism (Figure 72). The irregular pyramid has a flat portion at the lowest point of elevation in the triangle, and it extends upwards to the highest point in the triangle. This pyramid sits on top of a triangular prism that extends from the lowest point of the triangle down to sea level (or some other datum). For each triangle, you need to calculate the volume of the pyramid and add to it volume of the prism. By summing the volumes beneath each triangle, you can determine the volume beneath the entire TIN (Maguire, 2005).

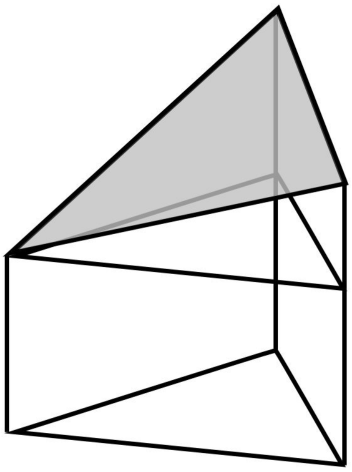


Figure 72. Determining the volume below a triangle in a TIN. The volume can be calculated by adding the volume of an irregular pyramid, which extends from the lowest elevation of the 3 points making up the triangle, to the highest elevation of the 3 points, to a rectangular prism, which extends from the lowest elevation of the 3 points making up the triangle to the vertical datum.

The DEM appears to be more difficult, but it can be approached in the same manner. The pixels that make up the DEM do not form a continuous surface, however a group of four pixels can be used to create two triangles (Figure 73). We then calculate the volume beneath each triangle in exactly the same way that we do for TINs.

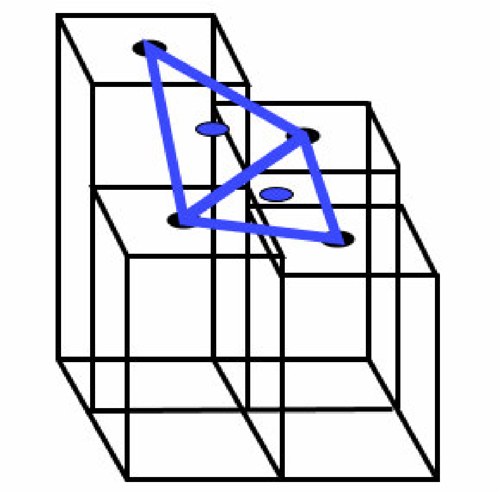


Figure 73. The centre of each pixel can be used to define a series of triangles. These triangles can then be analyzed in the same fashion as TINs.

For most natural features, knowing the amount of volume that they occupy maybe of academic interest, but it is of little practical value. If we have an irregularly shaped stockpile of gravel on a flat surface, then knowing the volume beneath the surface may be of some value. However, for most applications, volume analysis is important when we have a *before* surface and an *after* surface. This can be used for mining excavation applications. Before excavation begins, we determine the volume beneath the area to be excavated. Once excavation is complete, we read measure the surface and calculate the volume beneath the new surface. The difference in volume represents the amount excavated.

Unlike the other tools that we have discussed in this section, no layers are produced with volume analysis; only a report showing the volume is produced (Figure 74). This may not be particularly useful in construction applications, because we want to know *where* to remove material from and *where* to place it (after all, this is a GIS). We solve this problem by using cut and fill analysis, which reports volume calculations in map form.

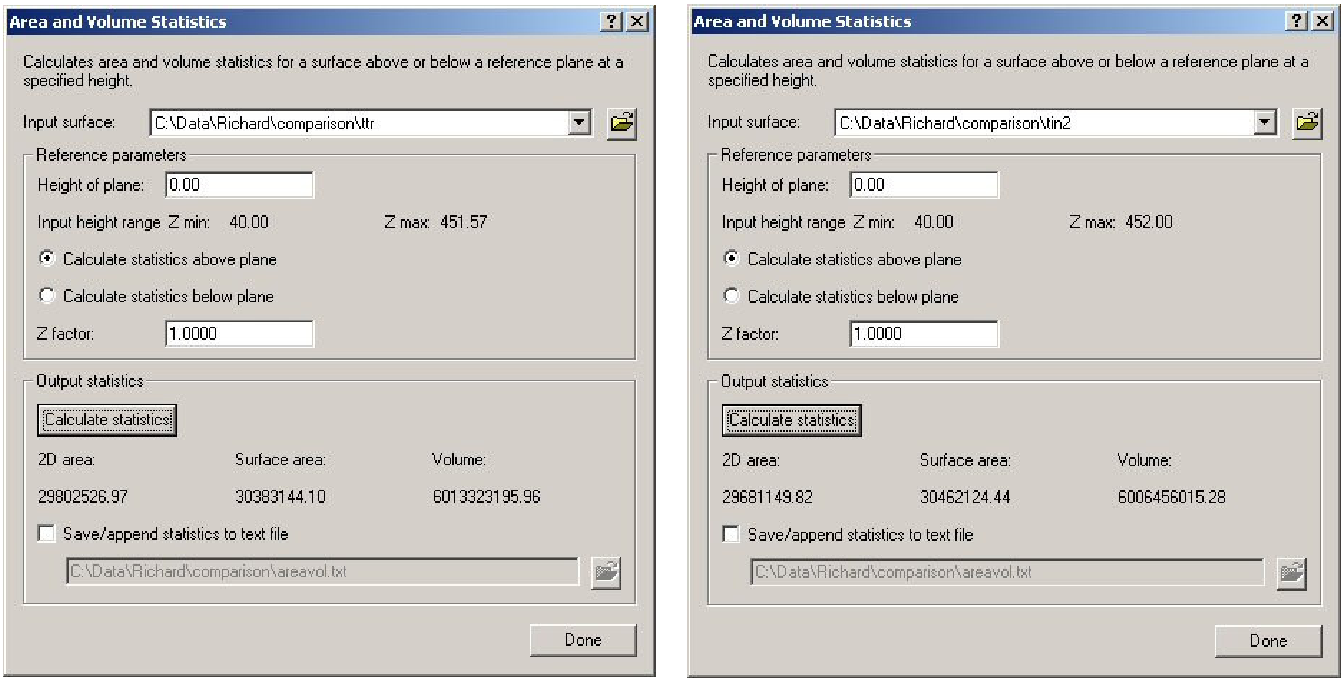


Figure 74. Volume statistics for a DEM (left) and a TIN (right) created from the same source data.

5.5.9 Cut and Fill

Cut and fill analysis compares two surfaces to determine the difference in volume between the first surface and the second surface. Volumes may be negative if material has been removed from the first surface, or they may be positive if material has been deposited. Layers which show simply where material has to be removed or added can be produced, or more sophisticated analysis can create layers that show how deep excavations need to be or how much material has to be added (Figure 75).

Cut and fill analysis is frequently used in construction projects, particularly for roads. In order to build a road for the least cost, it is necessary to move a minimal amount of material. Using cut and fill analysis, the route for the road can be adjusted to minimize the amount of material that needs to be moved in order to make the road flat or gently sloping. Once the final route has been chosen, and detailed plans have been made, cut and fill analysis can be used to determine exactly how much material needs to be removed, where it needs to be moved to, and how much material needs to be filled to build the final road. If regular updates of topographic data are available (perhaps from flying a LIDAR sensor over the road construction area), the progress of construction and the amount of material remaining to be moved can be tracked closely.

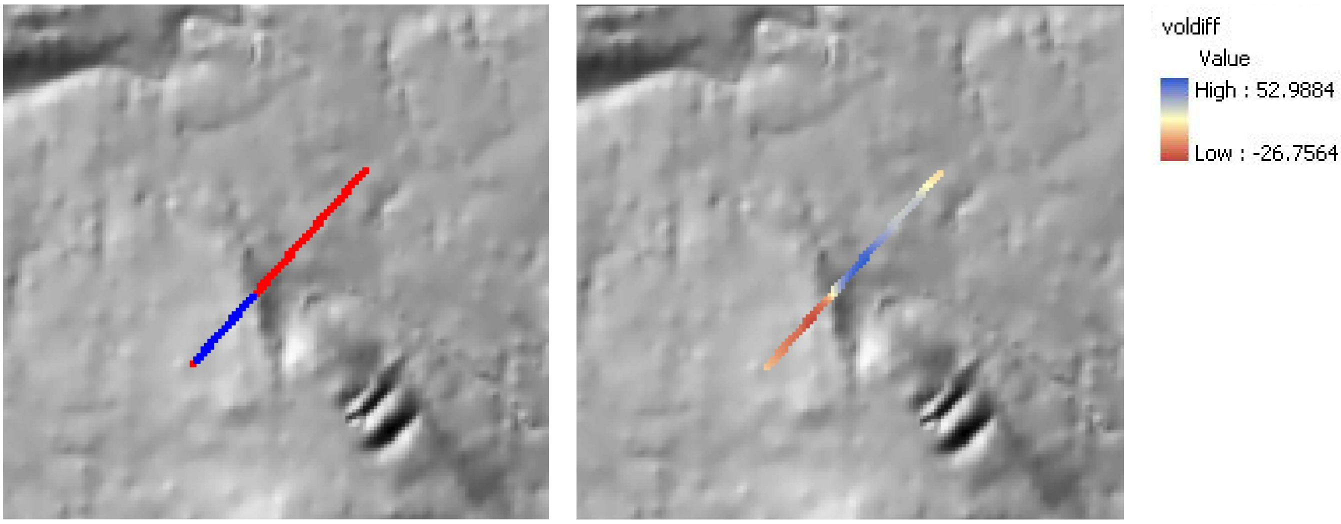


Figure 75. Cut and fill analysis for a hypothetical 1.3 km long section of road with a uniform slope constructed over uneven terrain. The image at left shows areas that need to be excavated (blue) and filled (red). The image at right shows depth of excavation or fill required. Clearly this section of road is impractical to build, as over 50m of fill needs to be deposited at some points to create a uniform slope.

5.5.10 Curvature

Determining the curvature of a piece of land can be valuable in the study of environments, both human and natural. The curvature of a piece of land affects such factors as the local climate, the amount of solar illumination, or the degree to which the area is "closed-in" or "expansive." The horizontal (plan) and vertical (profile) curvatures are combined into an overall measure of curvature in the ArcGIS 9.2 tools. A negative value indicates that the area is sheltered, and a positive value indicates that the area is exposed. Frequently, we want to split the horizontal and vertical curvature of slope into separate parts (Figure 76). Each of these affects the physical geography and geology of an area.

Curvature can be calculated by comparing a triangle with the triangle or pixel above and below it to determine the vertical curvature, and the triangle or pixel to the left and right to determine the horizontal curvature. The curvature tools in ArcGIS 9.2 work only on the DEMs, not on TINs.

In the vertical plane, a convex slope is termed "water shedding," whereas a concave slope is termed "water collecting." In the horizontal plane, slopes that are concave are frequently in shadow, whereas slopes that are convex are exposed to sunlight throughout much of the day. When combined, these properties help to determine the physical properties of the land surface.

Studies of curvature have application in the physical, human and natural environment. For example, a surface that is both concave in the horizontal and vertical planes makes an ideal location for a pocket glacier if it is located at sufficient altitude and has a northerly aspect. Not only is it shaded from sunlight for most of the day, but it also collects water (or snowfall due to avalanches), so there is a build-up of frozen water in a cool pocket. Curvature studies have been used extensively in Lithuania for studies of soil, and have been used in Russia to identify promising areas of oil exploration.

Curvature analysis may be used for studies of the human environment. A positive horizontal curvature indicates an area of broad views, which might be suitable for a house, a restaurant or an observation point. A negative horizontal curvature indicates an area that is enclosed. Such a location might the appropriate for something that should *not* be widely visible, such as an industrial facility.

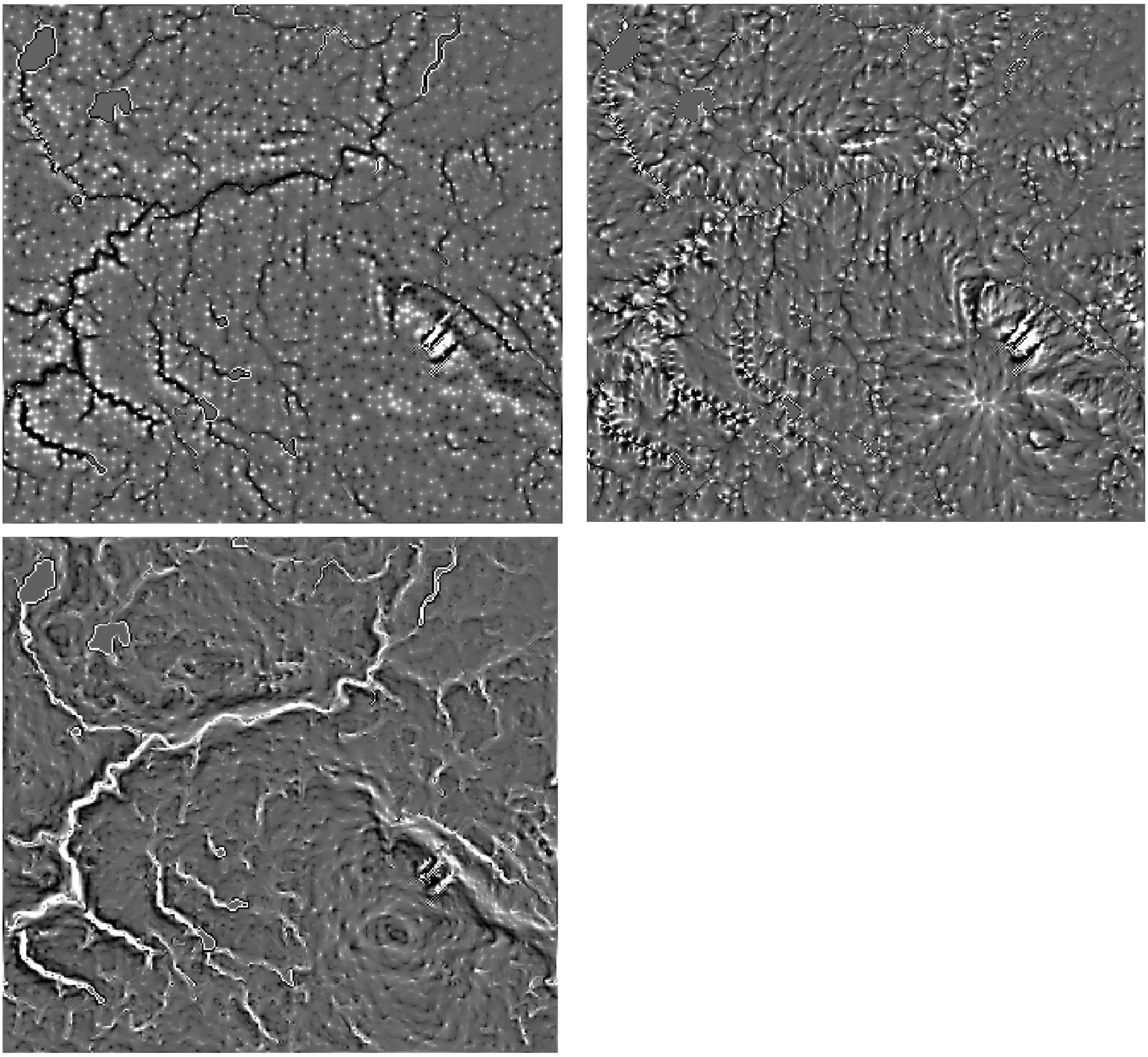


Figure 76. Combined (upper left), horizontal (upper right) and vertical curvatures (lower left).

Positive vertical curvatures imply that areas have an open view of the sky and surrounding terrain. Such sites might be appropriate for observatories, observation towers, radio and television transmission towers, or cell phone towers. Conversely, negative vertical curvatures imply areas that have poor views of the sky. These areas tend to be enclosed, have uphill slopes on both sides, and are poor locations for the transmission and reception of radio waves. Search and rescue teams need to pay special attention to such areas, since they tend to "trap" people.

Overall curvature might be used to identify building sites that are sheltered from the elements (those that have a negative overall curvature), or that have been expansive view, but which are exposed to the elements (those having a positive overall curvature). Wildlife also responds to similar conditions, so an area with a positive overall curvature might be favourable to animals such as mountain goats, which prefer exposed rocky areas. Animals that wish to be sheltered from the elements might prefer an area with a negative overall curvature.

5.5.11 Path Distance

In Module 3, we discussed the Path Distance algorithm, which allows overland travel to be modelled based on a cost raster, a horizontal factor (i.e. wind), and a vertical factor (i.e. a DEM). We will not discuss this further in this section, except to note that this command does not accept a TIN as an input for the vertical factor.

5.6 Examples

The following case studies present some examples of how digital terrain modeling is being used to solve problems in Lithuania, Europe, around the world, and even on Mars.

5.6.1 Sea Level Analysis

Determining the effects of global warming is an ideal application for digital terrain models. The European Environment Agency has produced a very rough analysis showing which areas in Europe are most susceptible to a sea-level increase of up to 5m. Using Digital Chart of the World (DCW) and Digital Terrain Elevation Data (DTED) data for Europe, the agency built a DEM and showed the areas that might be flooded (European Environment Agency, 2005). Fortunately, Lithuania is not affected strongly, according to their results, although Latvia, Estonia and Poland are all expected to experience moderate levels of flooding (Figure 77).

A much more detailed analysis is currently under way for the Southeast Baltic Sea to determine the effects of sea level rise over the short and long-term (Gelumbauskaitė & Šečkus, 2005). By combining elevation data from three different sources (GIS Centras data for Lithuania, Baltic Sea bathymetry, and GTOPO30 data) a single DEM with a resolution of 50 m horizontally and 1 m vertically was created for Lithuania, Kaliningrad, and Poland, using kriging as the interpolation technique. Sea level changes were modelled based on the effects of sediment level changes (erosion or deposition), eustatic changes (due to changes in ice levels due to global warming), and isostatic changes caused by the land surface gradually rebounding after the last Ice Age.

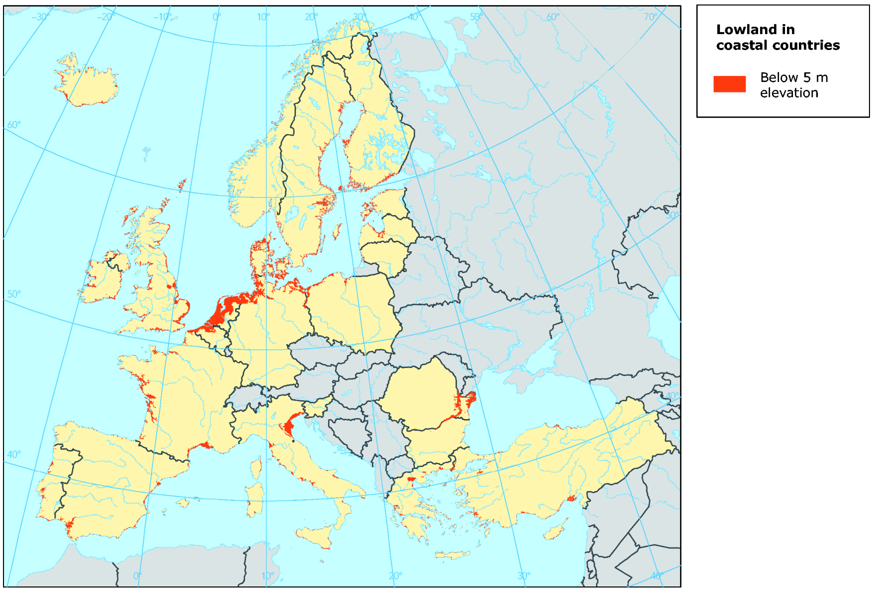


Figure 77. European Environment Agency map of areas susceptible to increases in sea level, based on DEM analysis (European Environment Agency, 2005).

5.6.2 Removal of Contaminated Soils

In Milwaukee, Wisconsin, USA, a group of 80 land parcels that were formerly occupied by industrial developments is now being redeveloped for housing. Various levels of soil contamination occur on the site, although it was unknown whether a particular patch of soil was uncontaminated or highly contaminated. Two backhoes and 10 dump trucks were used to remove material, and the whole operation was coordinated using a GIS performing cut and fill analysis and a survey-grade (sub-centimetre accuracy) GPS receiver.

To solve the problem, a series of soil samples were taken at various places throughout the site, and the depths and levels of contamination were recorded. Soils were divided into three contamination levels:

1. Contaminant levels that are above residential limits, but below industrial limits
2. Contaminant levels that are above industrial limits
3. Contaminant levels that are below residential limits

The three types of soil needed to be separated with the first type being moved to industrial sites, and the second being moved to a landfill. The (relatively) uncontaminated soils need to be temporarily removed and then replaced, once the more contaminated soils have been removed.

To accomplish this task, the site was first surveyed the GPS receivers to create a DEM showing the original ground surface. Next, polygons were laid out on the ground outlining the areas of contamination, and these areas were excavated and moved. At the end of each day, the GPS was used to re-survey the areas being excavated to make sure that the right amount of material was being removed (Figure 78).

Figure 78. Cut and fill analysis to show depth of excavation required in Milwaukee, Wisconsin, USA (Misky et al., 2004).

At the end of the project, the final amount of soil that was removed was within 1% of the amount planned during the original engineering study. Needless to say, this helped to minimize the cost of the project, by allowing as little material to be moved as possible, and by ensuring that the project could be completed as quickly as possible (Misky *et al.*, 2004).

5.6.3 Wireless Computer Access Planning

Ball State University, in Muncie, Indiana, USA wanted to establish a wireless computer network both on-campus, and in the surrounding community where many students live. The "Digital Middletown Project" required some way to model signal reflection, refraction, and attenuation characteristics in the area surrounding the University.

To ensure that the wireless computer network worked as planned, the University made use of a product called Cellular Expert, which is produced by HNIT-Baltic Geoinfoservisas, UAB in Vilnius. Cellular Expert is a GIS application that allows radio reception by cellular telephones, wireless computers, and digital radios to be modeled based on the techniques of viewshed analysis, which were discussed in Section 5.5.6.

Based on a detailed digital elevation model that includes buildings, Ball State was able to determine the best frequency to use and the height of towers required, based on the minimum acceptable field strength level for wireless computer cards. This Cellular Expert tool is able to model sector antennas that are used for cellular telephones using the AZ1, AZ2, V1 and V2 parameters used by the viewshed analysis command, and can produce maps showing wireless reception as well as profiles showing wireless signal strength at various heights in the buildings surrounding the University (Figure 79) (HNIT-Baltic, 2005).

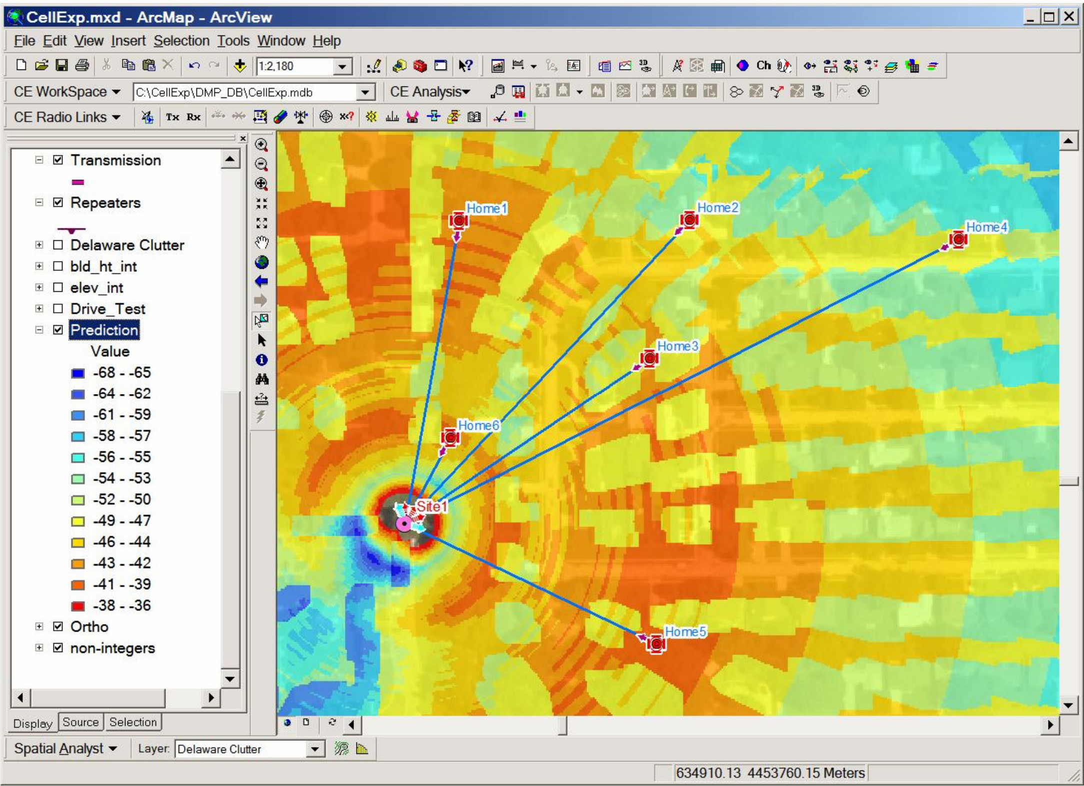


Figure 79. Modeling of wireless computer network signal strength surrounding central distribution tower at Ball State University in Muncie, Indiana, USA.

5.6.4 Exploration of Mars

A little further from home, the United States Geological Survey is making use of viewshed analysis to determine what areas of Mars can be photographed from the two Mars Exploration Rovers (Spirit and Opportunity). Each rover has a 1.5 m tall mast on which a camera is mounted for photographing surrounding terrain. On Mars, it should be noted, the planetary curvature is greater than on Earth, so the horizon is only 3.2 km away (as opposed to 4.4 on Earth). This can be modelled using viewshed analysis. The images are used to help plan where the rover should be directed to acquire further images (Figure 80) (USGS, 2007).

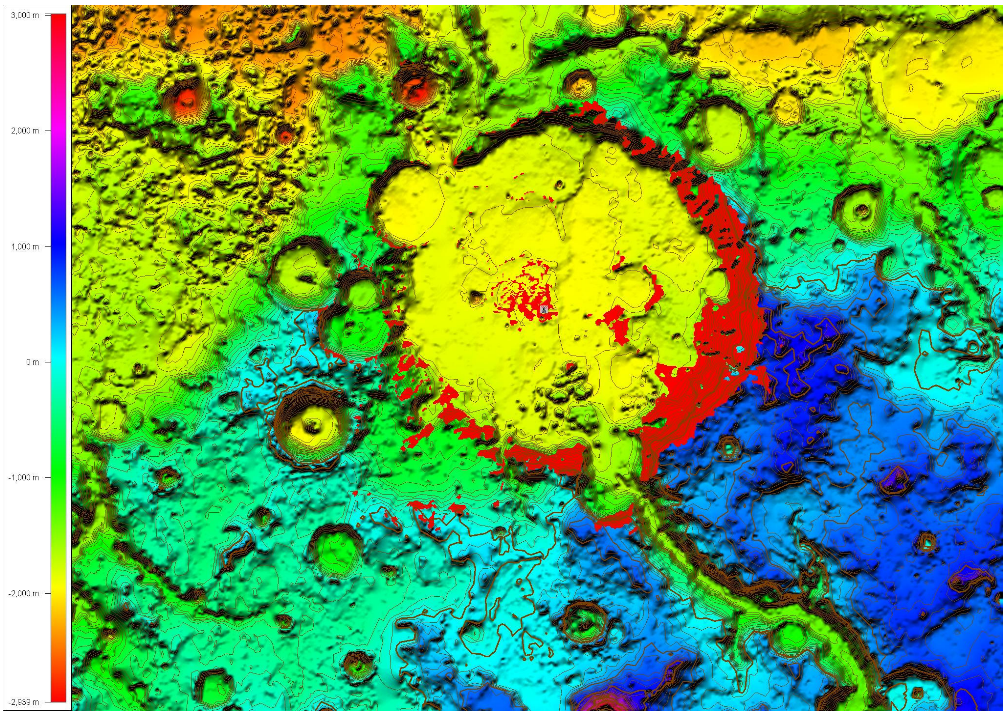


Figure 80. Visibility analysis for the Spirit Rover in Gusev Crater on Mars. Bright red areas are visible from rover. The size of the crater, combined with the short distance to the horizon, mean that the rover can see little of its local environment, but can see far-away hills, such as the crater rim, because they project above the horizon.

5.7 Conclusion

Digital Terrain Models are one of the most exciting areas for GIS applications. For centuries, flat paper maps have allowed only a very limited form of terrain analysis. DTMs now allow the full impact of the third dimension to be considered in GIS modeling through many applications such as slope and aspect analysis, modeling of solar radiation, viewshed analysis, cut and fill analysis, and the visualization of terrain.

Appendix A

This appendix discusses the types and uses for different interpolation techniques that are used to create digital elevation models. The differences between these techniques are fairly subtle, and an animation has been created in the PowerPoint presentation for this lecture to help see the differences between the results of these different techniques. For additional detail on the algorithms used in these techniques, please refer to course GII-07, Spatial Analysis and Modeling.

A.1 Local Interpolation Techniques

Local interpolators use a subset of the total number of points to predict a value at an unknown location. Because they use fewer points to make this prediction than do global interpolators, local interpolators tend to run faster.

Natural Neighbours is in interpolation technique that automatically determines and neighbourhood for each point based on the surrounding points. In this technique, also known as Sibson or "area-stealing" interpolation, the size of the neighbourhood is variable, depending on the number of points in the area. To determine the number of points in the neighbourhood, this technique calculates a Voronoi polygon surrounding the point whose value will be interpolated. Voronoi polygons are calculated by determining the distance between a point and all surrounding points. The mid-point between the point and each surrounding point is used to define the edge of a polygon. The Voronoi polygon is then overlaid on a set of Voronoi polygons that were created for the original data points (Figure 81). The area of overlap between the new Voronoi polygon and the existing ones is used to determine the amount of weighting that is used for each point when calculating the value for the interpolated point.

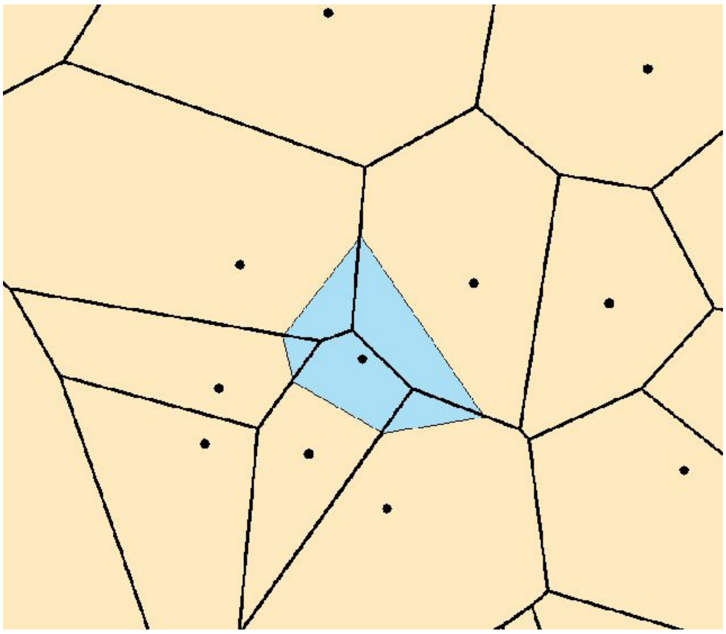


Figure 81. The weighting of a new point (centre of blue polygon) in Natural Neighbours interpolation is based on the amount of overlap between the Voronoi polygon for the new point (blue) and the previously existing Voronoi polygons (beige).

The computation of the initial Voronoi map for all the data points is very rapid, as is the calculation for the new Voronoi map, so this technique is completely automated and very rapid. This enables this technique to be used on very large data sets.

Because this technique uses weighted averages, it is an inexact interpolator, and is deterministic in nature. The technique of using weighted averages ensures that the predicted value will always be between the minimum and maximum value in the neighbourhood. This means that extreme values in the neighbourhood are removed, resulting in a smoothed appearance to the data set. This technique has the advantage in that it uses the density of local points to define the neighbourhood. The next interpolator, in contrast, requires the number of points to be defined *a priori*.

Inverse Distance Weighting (IDW) is another common local interpolator, which is similar in concept to Natural Neighbours interpolation. The difference is that the user defines the number of points that are used in the interpolation. The algorithm then locates a number of surrounding points, based on two techniques. In the variable-radius technique, the algorithm chooses the nearest *n* points, and determines their distance from the point that is being interpolated. In the fixed-radius technique, those points that lie within a circle having a specified radius are used. The closest points have the greatest effect on the prediction, and the farthest points have the least, hence the name *inverse* distance weighting. Inverse distance weighting works best for dense, evenly spaced data. In addition to being a local interpolator, IDW is also inexact and deterministic, just like Natural Neighbours.

Kriging is a local geostatistical technique that was developed in the 1950s by a South African mining engineer named Danie G. Krige, and was formalized in the 1960s by Georges Matheron. The technique was designed to predict the depths of geological strata based on depths where the strata were found in samples.

Because this is a stochastic interpolation technique, it attempts to make predictions based on spatial autocorrelation and the properties of the surface that is being modelled. Unlike many other interpolators, it does not assume that the input data is isotropic. Because it attempts to model the nature of the surface, this technique also does a reasonable job of extrapolating data for short distances outside the boundaries of the input data. It is an inexact interpolator, and so does not necessarily pass through all of the sample points. There are many variants of kriging, including ordinary, simple, universal, indicator, probability, and co-kriging.

Splines, also known as radial basis functions or piecewise polynomials, differ from the previous three interpolation techniques, in that it is an exact interpolator. A spline is a polynomial function that creates a surface, which passes through each data point in a dataset. Like the previous two functions, this is a deterministic interpolator, meaning that a mathematical formula (a polynomial function) is used to make a prediction for unknown point. Because splines use mathematical curves, they produce smooth surfaces, which are ideal for the modeling of rolling terrain. Unfortunately, because the surface produced by splines is so smooth, it is unable to model abrupt discontinuities, such as cliffs, canyons, or geological faults. To solve this, a variant of the spline command has been developed which allows the user to input lines representing discontinuities. At a discontinuity, the polynomial functions on each side of the discontinuity are abruptly discontinued, and the surfaces are joined so that a smooth surface is not formed across the discontinuity.

Local Polynomial Interpolation is the last of our local interpolation techniques. This is similar to splines in that it is a deterministic interpolator, but it is not and exact interpolator, however. The algorithm applies polynomial functions to neighbourhoods to model local curvature. Many polynomial curves are combined to generate the complete surface from the input points. The user specifies the order of polynomial can be used; the higher the order, the more variation in the surface that can be modeled. The resulting surface reflects local terrain variation quite well. Like kriging, local polynomial interpolation can account for anisotropic input data, because the user to specify the size and shape of the neighbourhood that is used for the computation of the polynomial surface.

A.2 Global Interpolation Techniques

Global interpolators are generally used when it is important to understand the overall trends of a surface. These operators use all of the input data points in computing the value for each unknown point. For this reason, they provide a higher degree of smoothing and do not represent local topography very well. They also tend to operate more slowly than local interpolators.

Global Polynomial Interpolation, also known as trend surface analysis, uses the same techniques as local polynomial interpolation, except that only a single polynomial function is applied to the entire surface. Depending on the order of the polynomial function, the surface may range from completely flat to curved; it will not reflect local variations in terrain. It is not surprising that this deterministic interpolator is also an inexact interpolator.

If only a first order polynomial is applied to the surface, a sloping plane will be generated; second order polynomials will generate a valley, and third order polynomials will generate a saddle or a bowl shape. In each case, regression analysis is used to create a least-squares fit to minimize the amount of error for the generated surface.

Topo to Raster is an inexact, deterministic interpolator, which uses a combination of techniques to produce a hydrologically correct raster. Unlike the other theory-based interpolation techniques that we have discussed, this technique is data-based and was developed to best represent real topography. The algorithm, which is based on ANUDEM software (Hutchinson, 1988, 1989) uses splines, but has been modified to allow for abrupt terrain changes and discontinuities, such as streams and ridges. Although this is a global interpolator, the algorithm has been optimized so that its performance is similar to local interpolators.

The algorithm is based also on geographical experience, which teaches us that it is much more likely that an area will have peaks (i.e. hill and mountain tops) than pits (i.e. sinkholes). Generally, when elevation data is collected, the surface elevation of water bodies is collected, so lakes and parts of the ocean appear to be flat. In fact, except in areas of karst topography, all pits should be filled with water. Because this technique "understands" how topography should appear, it is the only interpolation technique that is able to directly make use of contours as an input source. Normally, contours produce undesirable results when used as inputs for interpolators (see Section 5.4.6).

The beauty of this algorithm is that it allows streams to flow downhill without interruption. In many interpolations, streams may have small pits and peaks, which affect the modeling of water flow over the interpolated surface. A pit or a peak leads to the creation of a small "lake" in modelled downstream flow, so the model will not reflect reality because of small elevation differences. By adjusting input parameters, it is possible to create pits, provided that an elevation threshold is exceeded. This algorithm also generates a list of pits and their locations in the final DEM, so that these can be verified and corrected, if necessary.

Appendix B

The following is a list of de jure (developed in advance by a committee of interested parties) and de facto standards (developed in response to an immediate need by a software developer, and then widely adopted) that may be used when transferring DTM data. Each class is divided into raster and vector formats.

- **De jure standards**
  - **Raster**
    - Digital Terrain Elevation Data (DTED)
      - Publisher: National Imagery and Mapping Agency (NIMA)
      - Extension(s): .dt0, .dt1
      - Description: 16 bit binary file
      - Advantages: Use for legacy files
      - Disadvantages: Being replaced by SDTS
      - For more information: http://www.fas.org/irp/program/core/dted.htm, http://www.nga.mil/ast/fm/acq/89020B.pdf
    - Digital Elevation Model (DEM -- Note that, despite the name, this is a raster data exchange format, not a type of DTM)
      - Publisher: United States Geological Survey
      - Extension(s): .dem
      - Description: 16 bit binary file
      - Advantages: Used for many legacy files. Aggressive updating to SDTS is now under way.
      - Disadvantages: Being replaced by SDTS
      - For more information: http://erg.usgs.gov/isb/pubs/factsheets/fs04000.html, http://rockyweb.cr.usgs.gov/nmpstds/acrodocs/dem/2DEM0198.PDF
    - Spatial Data Transfer Standard (SDTS)
      - Publisher: United States Geological Survey (USGS)
      - Extension(s): .ddf
      - Description: binary files. Each consists of a "profile" that supports transfer of a particular data type
      - Advantages: Designed to support all types of spatial data
      - Disadvantages: Files encode other formats, including DLG and DTED. Somewhat confusing
      - For more information: http://data.geocomm.com/sdts/, http://mcmcweb.er.usgs.gov/sdts/whatsdts.html
    - Band Sequential, Band Interleaved by Line, Band Interleaved by Pixel
      - Publisher:
      - Extension(s): .bsq,.bil,.bip .hdr
      - Description: 8 or 16 bit binary file, header information separate
      - Advantages: Simple format
      - Disadvantages: Uncompressed
      - For more information: http://glcf.umiacs.umd.edu/data/guide/fileformat/

  - **Vector (Point and Line)**
    - Digital Line Graph (DLG)

- Publisher: United States Geological Survey
- Extension(s): .dlg
- Description: Vector data ASCII file
- Advantages: Used for many legacy files. Aggressive updating to SDTS is now under way.
- Disadvantages: Being replaced by SDTS
- For more information: http://www.safe.com/reader_writerPDF/dlg.pdf http://rockyweb.cr.usgs.gov/nmpstds/dlgstds.html

▪ Spatial Data Transfer Standard (SDTS)
  - Publisher: United States Geological Survey (USGS)
  - Extension(s): .ddf
  - Description: binary files. Each consists of a "profile" that supports transfer of a particular data type
  - Advantages: Designed to support all types of spatial data
  - Disadvantages: Files encode other formats, including DLG and DTED. Somewhat confusing
  - For more information: http://data.geocomm.com/sdts/, http://mcmcweb.er.usgs.gov/sdts/whatsdts.html

- **De facto standards**
  - **Raster**
    - Arc Export
      - Publisher: ESRI
      - Extension(s): .e00
      - Description: create an ASCII file from an ArcInfo coverage
      - Advantages:
      - Disadvantages: Very large, coverages have now been superseded by more modern file formats
      - For more information: http://tdr.uoguelph.ca/GEOG/gisystem96.htm#Arc/Info%20Export%20Format, http://avce00.maptools.org/docs/v7_e00_cover.html
    - Enhanced Compression Wavelet
      - Publisher: Leica Geosystems (ERMapper)
      - Extension(s): .ecw
      - Description: Wavelet compression
      - Advantages: Highly compressed: 1/20 to 1/50 of original size, can support very large files
      - Disadvantages: Lossy compression, Proprietary format
      - For more information: http://www.leica-geosystems.com/
    - Multiresolution Seamless Image Database (MrSID)
      - Publisher: LizardTech
      - Extension(s): .sid
      - Description: Wavelet compression
      - Advantages: highly compressed, widely supported format in commercial products
      - Disadvantages: Lossy compression, proprietary format

- For more information: http://docs.unh.edu/nhtopos/mrsid.htm, http://www.lizardtech.com

▪ Tagged Interchange File Format
- Publisher: Adobe Systems
- Extension(s): .tif, (.tiff)
- Description: Uncompressed binary file
- Advantages: Can be read by nearly all software
- Disadvantages: Uncompressed (very large), not georeferenced
- For more information: http://home.Earthlink.net/~ritter/tiff/, http://www.faqs.org/faqs/graphics/fileformats-faq/part3/section-147.html

▪ GeoTIFF
- Publishers: GeoTIFF Working Group/Adobe Systems
- Extension(s): .tif, (.tiff) .tfw
- Description: Georeferenced TIFF file
- Advantages: Can be read by nearly all software
- Disadvantages: Uncompressed data (very large)
- For more information: http://remotesensing.org/geotiff/spec/geotiff1.html, http://www.remotesensing.org/geotiff/spec/geotiffhome.html

▪ Arc Raster ASCII file
- Publisher: ESRI
- Extension(s): .asc, .txt
- Description: ASCII file generated by Raster to ASCII command
- Advantages: Human-readable file
- Disadvantages: Very large files
- For more information:

▪ Intergraph Raster
- Publisher: Intergraph
- Extension(s): .rgb, .cot, .cit, .rle, tg4, or .cfl
- Description: Multiple format supporting different colour depths
- Advantages: Well supported, file format publicly available
- Disadvantages:
- For more information: http://www.faqs.org/faqs/graphics/fileformats-faq/part3/section-72.html

▪ ERDAS Imagine file
- Publisher: Leica Geosystems (ERDAS)
- Extension(s): .img
- Description: 4, 8 or 16 bit unsigned, 16 bit signed, or 32 bit real binary file
- Advantages: widely used for image processing
- Disadvantages: Proprietary format, no information on file format found on web
- For more information: http://www.leica-geosystems.com/

o **Vector (Point and Line)**

▪ Arc Export
- Publisher: ESRI
- Extension(s): .e00
- Description: ASCII file
- Advantages:

    - Disadvantages: Large file size, depreciated by ESRI at ArcGIS 8
    - For more information:
- Shape files
    - Publisher: ESRI
    - Extension(s): .shp, shx, and .dbf
    - Description: Binary files
    - Advantages: Widely recognized by different software
    - Disadvantages:
    - For more information: http://en.wikipedia.org/wiki/Shapefile,
    - http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf
- Coverages
    - Publisher: ESRI
    - Extension(s): None
    - Description: Arc Coverage files will come as a directory of files, hopefully compressed
    - Advantages: Much archival data still exists in this format
    - Disadvantages: Proprietary format, unlikely to be ready by non ESRI software; has been deprecated by ESRI
    - For more information: http://mapserver.gis.umn.edu/docs/reference/vector_data/ArcInfo
- Geodatabase
    - Publisher: ESRI
    - Extension(s): .mdb + ?
    - Description: Spatial data encoded in database files. For personal geodatabases, this will be delivered in MS-Access format; for file geodatabases, it may be delivered in Oracle of other database file format
    - Advantages: Compatible with Enterprise DBMS system
    - Disadvantages: Readable only by ArcGIS
    - For more information: http://www.esri.com/software/arcgis/geodatabase/index.html
- Drawing Exchange Format (.dxf)
    - Publisher: Autodesk (AutoCAD)
    - Extension(s): .dxf
    - Description: ASCII or Binary CAD file
    - Advantages: Widely read
    - Disadvantages: Not georeferenced, has been superseded by DWG File
    - For more information: http://en.wikipedia.org/wiki/AutoCAD_DXF
- Drawing Format (.dwg)
    - Publisher: AutoDesk (AutoCAD)
    - Extension(s): .dwg
    - Description: Binary format
    - Advantages: Widely read
    - Disadvantages: Not georeferenced, Proprietary file format
    - For more information: http://en.wikipedia.org/wiki/DWG
- Intergraph Graphics Design System (.dgn)
    - Publisher: Bentley Systems (Microstation)
    - Extension(s): .dgn (possibly .igds)

- Description: CAD file
- Advantages: Widely read
- Disadvantages: Not georeferenced, Proprietary file format
- For more information: http://en.wikipedia.org/wiki/DGN

▪ GML
- Publisher: Open Geospatial Consortium
- Extension(s): .gml
- Description: XML based file for transfer of spatial data
- Advantages: Human-readable, open file format
- Disadvantages: Very large files
- For more information: http://mapserver.gis.umn.edu/docs/reference/vector_data/gml

▪ SICAD
- Publisher: AED SICAD
- Extension(s): .sqd
- Description: sequential topological vector file
- Advantages: Includes a metadata
- Disadvantages:
- For more information: http://translate.google.com/translate?hl=en&sl=de&u=http://www.gismngt.de/format/sqd.htm&sa=X&oi=translate&resnum=4&ct=result&prev=/search%3Fq%3Dsicad%2B.sqd%26start%3D10%26hl%3Den%26sa%3DN (translated to English) http://www.gismngt.de/format/sqd.htm (original German)

▪ MapInfo Native Format File
- Publisher: MapInfo
- Extension(s): .tab, .dat, (may include the following: .id, .map, .ind, .tda, .tin, .tma, .lda, .lin, lma)
- Description: Native file format for MapInfo
- Advantages:
- Disadvantages: Polygons were limited to 32,000 vertices in older versions of MapInfo
- For more information: http://en.wikipedia.org/wiki/MapInfo_TAB_format

▪ MapInfo Interchange File
- Publisher: MapInfo
- Extension(s): .mif, .mid
- Description: Interchange file format for MapInfo
- Advantages:
- Disadvantages: Polygons were limited to 32,000 vertices in older versions of MapInfo, ASCII file
- For more information: http://www.directionsmag.com/mapinfo-l/mif/Mif_j.htm

▪ Text File, Comma Separated Values File (ASCII)
- Publisher:
- Extension(s): .txt, .asc, csv
- Description: Text file with X, Y, Z triplets, may be delimited by various characters such as tabs, or commas (.csv = comma separated values)
- Advantages: simple to create
- Disadvantages: sometimes difficult to import, format is somewhat variable

- Being used for Lithuanian DTM (10m point spacing)

Module Self-Study Questions

1. Given the advantages and disadvantages of TINs and DEMs, explain which you would use in the following situations:

1. Detailed studies of river bed morphology
   Answer: TIN, because the form of the river bed must be represented accurately

2. Engineering bridge foundations
   Answer: TIN, because the cut and fill analysis must be exact for proper cost estimates

3. Using elevation data as an input for an index model, in which raster values are weighted and combined using Grid Algebra (see Module 2)
   Answer: DEM because the other inputs will also be in raster format

4. Worldwide maps showing sea level increase
   Answer: DEM, because you are analyzing large data sets at a low level of precision; the data are already available as a DEM

5. Classification of the land surface into landforms by their shape
   Answer: TIN, because the shape must be represented accurately

6. Solar Illumination studies for an area
   Answer: DEM, because the software isn't available for TINs!

2. Discuss how viewshed analysis might be used to optimize the location of fire watchtowers.

3. How might cut and fill analysis be applied to an archaeological dig?

4. What terrain analysis techniques might be helpful in determining the optimal location for a new ski resort?

Suggested Readings

- Natural Resources Canada (2007). *Applications: Digital Elevation Models* (http://ccrs.nrcan.gc.ca/resource/tutor/fundam/chapter5/23_e.php)
- Lund University GIS Centre. *File Format (GIS)* (http://www.giscentrum.lu.se/english/whatisgisFileFormat.htm)

ESRI Virtual Campus Module

- Introduction to Urban and Regional Planning Using ArcGIS 9 Module 5: Impact Assessment: An Essential Planning Task

Assignments

- Assignment 5: DTM Analysis*References*

Chang, Kang-tsung (2006). Introduction to Geographic Information Systems, 3rd Ed. New York: McGraw-Hill

DeMers, Michael N. (2005) Fundamentals of Geographic Information Systems, 3rd Ed. Hoboken, N.J.: Wiley.

European Environment Agency, 2005. European coastal lowlands most vulnerable to sea level rise (http://dataservice.eea.europa.eu/atlas/viewdata/viewpub.asp?id=2226, Aug. 17, 2007).

Gelumbauskaitė, L.Ž. & J. Šečkus (2005). Sea Level Changes and Coastal Evolution in the South-Eastern Baltic Sea. (http://www.astra-project.org/cms/sites/download/Coastal_evolution.pdf, Aug. 17, 2007)

HNIT-Baltic Geoinfoservisas (2005). Cellular Expert for High Bandwidth Wireless Network Development (Case study) (http://www.cellular-expert.com/DesktopModules/Articles/ArticlesView.aspx?tabID=0&alias=cellular-expert&lang=en-US&ItemID=90&mid=11895, Aug 17, 2007)

Korte, George (2001). *The GIS Book, 5th Ed.* Albany, New York: OnWord Press

Lund University GIS Centre. *File Format (GIS)* (http://www.giscentrum.lu.se/english/whatisgisFileFormat.htm, Aug.14, 2007)

Maguire, Brad (2005). *Towards a Landform Geodatabase: The Automatic Identification of Landforms*. Master's Thesis, University of British Columbia. (http://web.mala.bc.ca/maguireb/papers/Thesis_Final_Apr21_2005.pdf)

Misky, Dave, David Holmes, and Michael Kumbera (2004). "Maintaining Accurate Data During Brownfield Site Redevelopment Excavation" ArcNews Online, Winter 2006/2007 (http://www.esri.com/news/arcnews/winter0607articles/maintaining-accurate.html)

Natural Resources Canada (2007). *Applications: Digital Elevation Models* (http://ccrs.nrcan.gc.ca/resource/tutor/fundam/chapter5/23_e.php, August 11th, 2007)

United States Geological Survey (2007). MER Landing Site Viewshed Analysis (http://webgis.wr.usgs.gov/mer/viewshed_analysis.htm, Aug 17, 2007)

Terms Used

- 2.5 D Surfaces
- Arc Export File
- Aspect
- ASCII
- Breakline
- Circumcircle
- Coverage
- Cut and Fill Analysis
- De Facto Standards
- De Jure Standards
- Delaunay Triangulation
- Deterministic Interpolators
- Digital Elevation Model (DEM)
- Digital Terrain Elevation Data (DTED)
- Digital Terrain Model (DTM)
- Drawing Exchange File (DXF)
- Enhanced Compression Wavelet
- ERDAS Imagine File
- Exact Interpolators
- Geodatabase
- GeoTIFF File
- Global Interpolators
- Global Polynomial Interpolation
- Hillshading
- Horizontal Curvature
- Inexact Interpolators
- Interpolation
- Inverse Distance Weighting (IDW) Interpolation
- Kriging
- LIDAR
- Local Interpolators
- Local Polynomial Interpolation
- Multiresolution Seamless Image Database (MrSID)
- Natural Neighbours Interpolation
- Photogrammetry
- Photorealistic Rendering
- Piecewise Polynomial Interpolation
- Plan Curvature
- Profile Curvature
- Pyramids
- Radial Basis Functions
- Reduced Resolution Datasets
- Shape File
- Slope
- Spatial Data Transfer Standard (SDTS)
- Splines
- Stereoplotter
- Stochastic Interpolators
- Tagged Interchange File Format (TIFF)
- Terrain
- Topo to Raster
- Trend Surface Analysis
- Triangulated Irregular Network (TIN)
- Vertical Curvature
- Very Important Points Algorithm
- Viewshed Analysis
- Voxel