A KNOWLEDGE-BASED APPROACH TO DATA INTEGRATION FOR SOIL MAPPING

by A-XING ZHU • LAWRENCE E. BAND

RÉSUMÉ

L'intégration de données spatiales provenant de sources multiples est une méthode très courante pour une foule d'analyses géographiques. Cette méthode peut être divisée en deux étapes fondamentales : l'intégration des données sur le plan syntaxique et l'intégration des données sur le plan sémantique. L'intégration des données sur le plan syntaxique concerne l'alignment et la structuration des données spatiales provenant de différentes sources, tandis que l'intégration sur le plan sémantique s'attache à l'analyse et à l'interpretation des données spatiales de sources multiples utilisées pour répondre à des questions précises. Cet article présente une méthode d'intégration de données de sources multiples fondée sur les connaissances, dans un contexte sémantique. Cette méthode fait appel à une méthodologie propre aux systèmes experts pour intégrer des connaissances empiriques à d'autres données environnementales afin d'en tirer de l'information sur un phénomène spatial donné. La méthode incorpore également la logique floue au processus d'intégration des données afin de tenir compte de la nature continue dans l'espace des phénomènes géographique. L'utilisation de cette méthode d'intégration des données est illustrée à l'aide d'un exemple d'inférence des sols. Les résultats d'un relevé de terrain effectué dans le cadre de cette étude ont démontré que la méthode fournissait de l'information sur les sols de meilleure qualité que l'information tirée de la carte des sols produite à partir d'un relevé de terrain classique.

SUMMARY

Multi-source spatial data integration is a very common process in many geographic analyses. Integration of multi-source spatial data can be divided into two basic steps: syntax data integration and semantic data integration. Syntax data integration is concerned with the alignment and formatting of spatial data from different sources, while semantic data integration deals with the analysis and interpretation of multi-source spatial data when they are used to answer specific questions. This paper presents a knowledge-based approach for integrating multi-source spatial data in a semantic context. The approach employs an expert system methodology for integrating empirical knowledge with other environmental data for the derivation of information about a given spatial phenomenon. The approach also incorporates fuzzy logic into the data integration process to accommodate the spatially continuous nature of geographic phenomena. The use of this knowledge-based semantic data integration is illustrated through a soil inference example. Results from a field survey in the soil inference example have shown that the approach produces higher-quality soil information than the soil map produced from a conventional soil survey.

INTRODUCTION

Integration of multi-source spatial data has been the ultimate tool for many geographical analyses. Geographica analysis and a considerable number of management decisions require the use of data from field surveys, paper maps and human knowledge about the phenomenon in questio These data are often in different forms (digital or analog different storage formats (different raster or vector formats and have different meanings for different analyses. Therefore, they must be aligned, formatted, and integrated t support management decisions and research conclusions With the development of remote sensing technology, th increasing availability of remotely sensed data, the development of Global Positioning Systems (GPS), and the development of inter-disciplinary research, the integration of spatia

• A-Xing Zhu is with the Department of Geography, Miami University, Oxford, Ohio 45056.
• Lawrence E. Band is with the Department of Geography, University of Toronto, Toronto, Ontario M5S 1A1.

data from different sources becomes an increasingly important issue.

Integration of multi-source data can be divided into two major parts: syntax integration and semantic integration. Syntax integration is concerned with the alignment of spatial data in terms of their format and geographic registration (georeference). The tasks of syntax integration include analog-to-digital conversion (A/D), format transformation, and map transformation (including georeference). Paper maps have been the ultimate data source for geographic information. Often these paper maps need to be converted into digital forms (digital data layers) through digitizing or scanning processes (see Chapter 4 in Aronoff, 1989; Chapter 4 in Burrough, 1986). This process of converting paper maps into digital maps (data layers) is referred to as analog-to-digital conversion (A/D). Sometimes, the required data may be in different digital forms (such as different raster or vector formats). These data may have to be transformed into the format (data structure) that is being employed in the system currently used for the project. This process of converting existing digital data from one format (data structure) to another is called format transformation (Chapter 7 in Aronoff, 1989; Chapter 10 in Clarke, 1990; Nichols, 1981; Pavlidis, 1979; Peuquet, 1981a, Peuquet, 1981b; Piwowar *et al.*, 1990; Van Der Knaap, 1992; Van Roessel, 1985). Once the needed data layers are converted into the useable digital forms, the user must check if these data layers are correctly aligned. In other words, a point on one data layer should have the same coordinates as on the other data layers. If not, these data layers have to be registered to either a common universal coordinate system (such as the Universal Transverse Mercator (UTM)) or a same local coordinate system. This registration process is called map transformation (georegistration) (Chapter 7 in Aronoff, 1989; Chapter 9 in Clarke, 1990; Maling, 1991). Once these data layers are georegistered, they are ready to be processed, which is done through the semantic integration.

Semantic data integration deals with the meaning of these multi-source data when they are added (overlaid) together to answer a specific question. In the past, two basic approaches have been taken for data integration: the classification approach and the constraint approach. Under the classification approach, different data layers (input data layers) are overlaid together. In a vector data model, polygon boundaries on the input data layers are intersected to form a set of new polygons and then a new data layer is produced. The newly formed polygons are then classified (using supervized or unsupervized classification techniques) based on the attribute values that are carried over from the input data layers. Each class of polygons is then given a specific meaning and serves as an answer to the problem under concern. In a raster data model, the area is dissected into numerous small cells or pixels. Since these cells (pixels) have the same geometric shape, correspondent cells at different data layers match each other very well after georegistration. Therefore, the integration of these data layers simply becomes a classification problem of the pixel attributes. Pixels with a similar configuration of attribute values are assigned to the same class. Each of these resultant classes is then given a specific meaning with respect to the problem under concern.

There are four major problems with the classification approach. The first problem is that all data layers are treated with equal importance during the integration process regardless of the nature of the spatial phenomenon under study. In reality, information from some data layers are more crucial and important than those from other data layers with respect to a specific phenomenon. For example, soil information has very different meanings if forest productivity is being studied as opposed to the study of bird habitats. Therefore, different data layers may have to be treated differently in terms of their contribution to the spatial distribution of a phenomenon under study. The second weakness is that the influence of individual environmental factors (expressed as individual data layers) on a particular geographic phenomenon was given at the time when the resultant classes were assigned labels (meanings). There is no information on how these environmental factors influence the phenomenon. For example, bird ecologists have some knowledge about how forest types affect bird habitats and soil scientists know generally how elevation affects soil profile development through field observation and studies. However, this type of knowledge is not incorporated into the integration process. Instead, it is grossly assigned to the final classes. The third potential drawbacks of the classification approach is that the final results are expressed in the form of categorical classes. In other words, the characteristics of a phenomenon at a given location are categorically said to be those of a specific class, when in reality, the characteristics might well be between the characteristics of two or more classes. The assignment of a phenomenon with intermediate values to one of the predefined classes results in a significant loss of information. The fourth potential problem is a result of the delineation of classes in the attribute domain. When the ranges of attribute values are grouped into categories, many areas with attribute differences smaller than the differences of two consecutive categories are grouped together to form a uniform region (represented by a polygon). In other words, the spatial variation of a geographic phenomenon is assumed to occur only at the boundaries of polygons. Within each of these polygons, the property of the phenomenon is uniform. In reality, very few geographic phenomena exhibit this type of spatial variation. Most of the natural geographic features show at least some spatial gradation in terms of their property values.

The constraint semantic data integration approach differentiates itself from the classification approach by introducing the influence of a particular environmental factor on the phenomenon under study (such as influence of elevation on forest distribution and growth). The influence of each environmental factor on the phenomenon under study is expressed as a constraint, often as a binary function (such as "at an elevation over 1500 m, there are Douglas fir forests; below 1500 m, there are ponderosa pine forests"). Each of the input data layers is then preprocessed using its respective constraint to generate its respective intermediate data layer. Each of these intermediate data layers expresses the limitation or predilection for a particular environmental condition for the phenomenon under study. Finally, these intermediate data layers are overlaid to produce a final layer about the distribution of the phenomenon under study. The constraint approach has an advantage over the classification approach. In the constraint approach, knowledge of the relationship between the phenomenon under study and its environment is explicitly defined and incorporated into the data integration process. A researcher has some controls on the importance of each environmental factor in relation to the phenomenon under study by relaxing or tightening the respective constraints. However, the constraints are often defined in a crisp fashion. In other words, the influence of an environmental factor on the phenomenon is grouped into categories and expressed as some type of step function. Under the constraint approach, the final result is still expressed as uniform polygons, as it is under the classification approach.

There is a need for developing a new data integration approach that will allow the expression of the environmental influence in a continuous function and that will preserve the spatial gradation of the phenomenon in the final results. Such an approach, presented in this paper, is knowledge-based semantic data integration, which combines expert system techniques and fuzzy set theory for spatial data integration. In a broader sense, this knowledge-based semantic data integration approach (referred to as knowledge-based data integration) is a special type of constraint-based data integration. The difference between this knowledge-based data integration and conventional constraint-based data integration is the definition of the constraints and the representation of the results. With the knowledge-based data integration, the effects of environmental factors on the physical processes of a spatial phenomenon are expressed by a set of constraints. However, each constraint is defined as a fuzzy membership function (more or less a continuous function), which is based on the knowledge of a domain concerned. The difference between the definitions of "soil A occurs at middle elevation (2000 m to 4000 m)" using a conventional crisp function and of a fuzzy membership function is shown in **Figure 1**. The result from the knowledge-based data integration is expressed in terms of fuzzy membership values. For example, if we were to infer the soil at a location, the result from the knowledge-based data integration would express the membership for the soil at the location being the prescribed soil type instead of assigning or not assigning the soil type to the location. In reality, the soil at the location may exhibit some characteristics of the prescribed soil type but may not be exactly the same. The fuzzy membership representation of integration results may be more realistic. In the next section we discuss the knowledge-based data integration in depth, which is followed by a case study using the described method for soil inference.

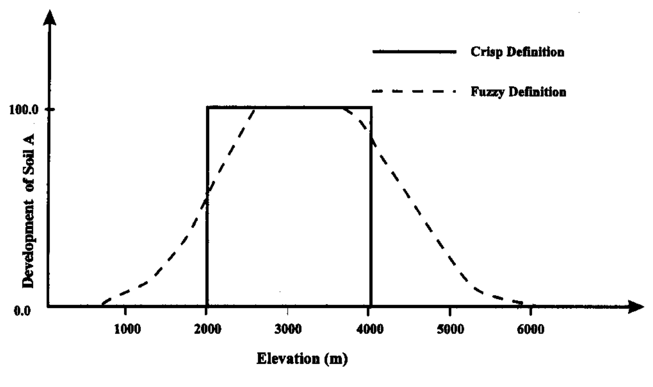

Figure 1.
Crisp and fuzzy definitions of "soil A occurs at middle elevation."

KNOWLEDGE-BASED DATA INTEGRATION

In many management activities, the distribution of a geographic phenomenon is often derived from a set of available data on the environment that the phenomenon is very tightly associated with or influenced by. This derivation conforms to Equation 1,

$$O = f(E) \tag{1}$$

where O is the spatial distribution of a geographic phenomenon, E is the environment associated with the phenomenon, and f is the function that specifies how the environmental conditions influence the distribution of a given geographic phenomenon.

In many cases, the exact form of f is unknown due to a limited understanding and complicated relationships between a given phenomenon and its environmental conditions. Different approaches of approximating f result in different methods of semantic data integration. With a knowledge-based data integration approach, f is approximated using the empirical knowledge of local experts who, having studied the phenomenon and its distribution for many years, have accumulated a great deal of information on the relationships of the phenomenon and its environmental conditions. Knowledge about the phenomenon and its environment from existing reports and maps can also be used to approximate f.

In the knowledge-based data integration, the = in Equation 1 is interpreted under fuzzy logic (referred to as fuzzy inference). Under crisp logic, a geographic phenomenon at a given point can belong to only one type. Therefore, the phenomenon at a given point carries only the properties of the assigned type. In reality, many geographic phenomena, such as soils, often vary continuously. The phenomenon at a given point might have some but not all of the property values of the assigned type, and might also possess some properties of other types or have the property values intermediate to typical values of prescribed types. This gradation of spatial phenomena cannot be represented with crisp logic. Under fuzzy logic, the phenomena at a given point can be classified as more than one type with different degrees of assignment depending on the similarity between the phenomenon at the site and the prescribed type (**Figure 2**). In other words, the phenomenon at a given point (i, j) can be represented as a vector of similarity values (S_{ij}). S_{ij} is also referred to as a vector of memberships or simply as a membership vector and consists of $S_{ij}^1, S_{ij}^2, \ldots, S_{ij}^k, \ldots, S_{ij}^n$, with n being the total number of prescribed types for the phenomenon. The similarity value S_{ij}^k represents the similarity between the phenomenon at the point (i, j) and the prescribed type k with a similarity value of 1 being exactly the

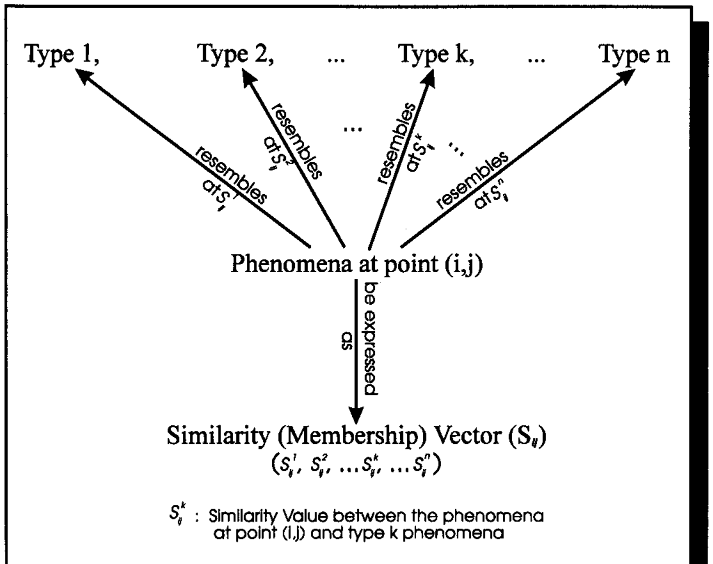
Type 1,
Type 2,
...
Type k,
...
Type n
resembles at S_{ij}^1
resembles at S_{ij}^2
...
resembles at S_{ij}^k
...
resembles at S_{ij}^n
Phenomena at point (i,j)
be expressed as
Similarity (Membership) Vector (S_{ij})
(s_{ij}^1, s_{ij}^2, ...s_{ij}^k ...s_{ij}^n)
S_{ij}^k : Similarity Value between the phenomena at point (i,j) and type k phenomena

Figure 2.
Representation of geographic phenomena under fuzzy logic.

same as the prescribed type *k* and 0 not being the prescribed type *k* at all. Values between 1 and 0 represent different degrees of similarity to the prescribed type *k*. It can be seen that fuzzy logic is more appropriate than crisp logic for representing the spatial variation of a geographical phenomenon such as soils or wildlife habitats.

Once both the environmental conditions at a site and the relationships between a given phenomenon and its environment are known, the membership vector can then be obtained for the site. This membership vector (S_{ij}) shows the degree of fit of the phenomenon (such as soil) at the site to various types (such as different soil types). In other words, the membership vector of a given site contains several elements. Each element expresses the belonging of the phenomenon at the site to a particular type. If an area is made of M by N pixels (sites), we would have an M by N array (S) of membership vectors. If we display the membership values of the first element in each of the membership vectors for the entire area, we would have an image of memberships for type *1* phenomenon. This image is different from the crisp logic-based image since the membership image shows the degrees of occurrence of a particular type of phenomenon over the area. In other words, the distribution of a given type of phenomenon varies in degrees, not in a black/white fashion, as depicted in crisp logic-based images. It is these varying degrees that provide information about the gradation of spatial phenomena over space.

A CASE STUDY: SOIL INFERENCE

The Problem

Knowledge of the distribution of soil properties over the landscape is required for a number of hydrological, ecological, and land management applications. In detailed hydroecological and other environmental modelling applications, continuous soil properties over an area are very much desired to approximate the resolution of other e
vironmental parameters gathered from remote sensing an
digital terrain analysis (Band *et al.*, 1991, 1993).

Currently, soil maps produced from soil surveys a
digitized to provide the soil information required for
variety of land analyses and management activities. Hov
ever, soil maps often contain a great deal of uncertain
since much of the quantitative and qualitative knowled
of the soil scientist regarding the occurrence of given sc
categories (for example, series) or properties is not mai
tained in the map. There are some major problems regar
ing the use of current soil maps in geographic analys
(Chapter 6 of Burrough, 1986). These problems inclu
limited coverage at a fixed scale, location errors, attribu
errors, and insufficient information in the mapping uni
due to the discrete spatial model (Bregt, 1992; Heuvelin
1993) with which soil maps are produced. The discre
spatial model is based on crisp logic. Under crisp logic, t
soil of an area belongs to one, and only one, soil catego
even if the soil might resemble more than one soil catego
since spatial variations of soil are more or less continuou
Under the discrete spatial model, soils of an area are tran
lated onto soil maps as discrete polygons. Areas within ea
of the polygons are assumed to have the same type of so

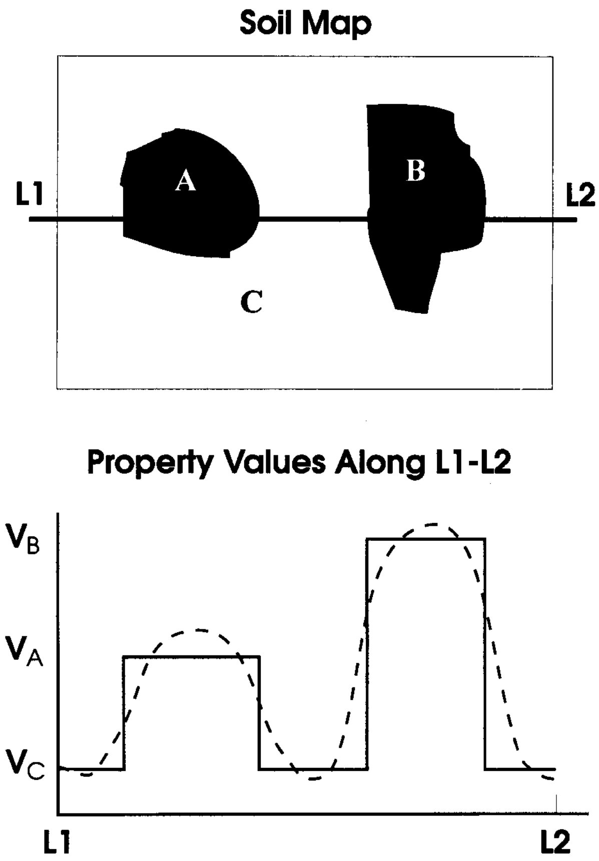

Figure 3.
Soil property on soil maps and in reality; solid line for property profile from soil map, dash line for real property profile along L1-L2.

Delineation of soil continuum into discrete polygons also assumes that soil properties change dramatically at the boundaries of these polygons (**Figure 3**). As shown in **Figure 3**, the actual value of the continuous soil property (*v*) might exhibit variation outlined by the dashed curve.

Studies have shown that even soil maps produced from systematic surveys might contain large amounts of "impurities" (errors) within the units delineated (Beckett, 1971; Beckett and Burrough, 1971; and Wilding *et al.*, 1965). These "impurities" might propagate through geographical analysis in a GIS and make the results from these analyses, particularly GIS-based simulation modelling, unpredictable.

While the soil scientists conducting the survey and mapping process may be aware of the degree of uncertainty in the mapping and may be able to infer more details on expected soil properties at a given location on the map from their knowledge of soil/landscape relations, this information is generally not translated onto soil maps. It is feasible to use the knowledge-based data integration approach to incorporate this information into the soil information generation processes.

Study Site

The study area for testing the knowledge-based data integration is the south part of the Lubrecht Experimental Forest, located about 50 km northeast of Missoula, Montana, USA. The forest was established in 1937 to foster research on natural resources. The study area is centred on the North Fork of Elk Creek, and is 4.8 km in the north/south direction and 7.5 km in the east/west direction (160 rows and 250 columns with a pixel size of 30 m by 30 m).

The elevation in the area ranges from 1,160 m to 1,930 m with high elevation in the east and southwest and low elevation in the northwest (**Figure 4**). The majority of areas are at an elevation between 1,200 m to 1,510 m. There is no evidence of glaciation in the study area (Brenner, 1968). Ross and Hunter (1976) estimated that the mean annual precipitation at 1,830 m is between 50 cm and 76 cm. Annual precipitation at lower elevations is lower. Annual precipitation at 1,200 m is between 37 cm and 56 cm. Approximately 44% of the precipitation falls during the winter (November through March) and 24 percent falls during the summer (June through August) (Nimlos, 1986). The mean annual air temperature at 1,200 m is about 4°C. The mean winter air temperature (December through February) is about -6°C, while the mean summer air temperature (June through August) is about 15°C. The study area can be considered as semi-humid to semi-arid.

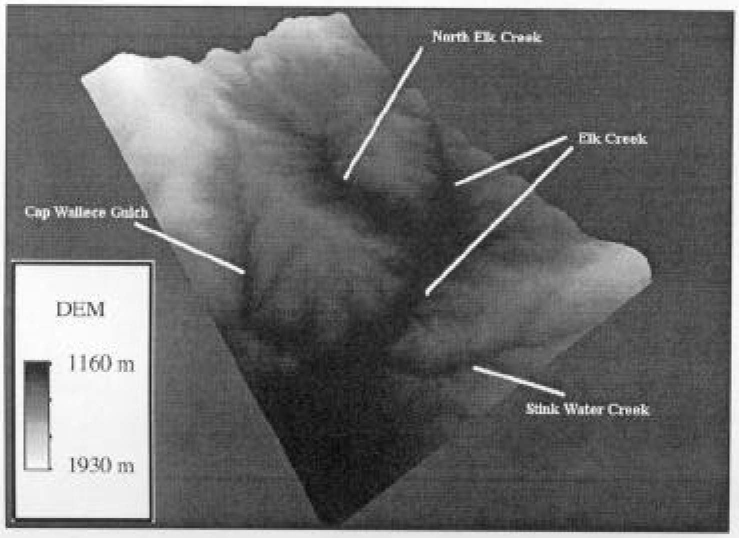

Figure 4.
A Digital Elevation Model of the study area.

Most of the mountain slopes in the study area are forested, dominated by Douglas-fir (*Pseudotsuga menziesii*), although lesser amounts of Western Larch (*Larix occidentalis*) and Ponderosa Pine (*Pinus ponderosa*) are present. A small portion of the study area (the northwest) is covered by fescue/bluebunch wheatgrass. Only a few areas of virgin forest remain, most having been logged, mainly during the periods of 1904 to 1906, 1925 to 1935 (Cauvin, 1961), and 1961 to 1962 (Brenner, 1968). Much of the timber is second-growth. There have been no large wild fires in the study area since 1937.

Four major types of parent materials are in the area: belt rocks, granite, limestone, and recent alluvium (Brenner, 1968; and Nimlos, 1986). The alluvium materials occur only in limited areas along the North and South Forks of Elk Creek. The first three parent materials make up the majority of the study area, with belt rocks in the north, granite in the south, and limestone through the central part.

Soils on these three materials are formed from a mantle of colluvium. There are four soil orders in the general area of the Lubrecht Experiment Forest: Alfisols, Entisols, Inceptisols, and Mollisols (Nimlos, 1986), and only two orders — Entisol and Inceptisol — are found in the study area. Entisols are weakly developed soils with very little organic matter accumulation and no illuvial clay or sesquioxides. In Lubrecht, Entisols are usually found on ridge crests. Inceptisols are young soils with little or no illuviated clays but brown subsoil horizons that indicate some translocations of sesquioxides. About 90% of the soils (in terms of areal extent) in the study area are Inceptisols. The 12 mapped soil series in the study area are: Ambrant, Elkner, Evaro, Ovando, Repp, Rochester, Sharrott, Tevis, Trapps, Whitore, Winkler, and Winkler Cool. Brief descriptions of these soil series are given as follows.

Ambrant, Elkner, Ovando, and Rochester are the soil series occurring on the granite materials that are located in the southeast part of the study area. Ambrant and Rochester are low-elevation soil series and often occur intermittently. Ambrant has a better developed and a deeper profile, while Rochester has a poor profile development and contains a large amount of rock fragments. Rochester often occurs at the dry sites where limited moisture conditions restrict the development of the soil profile. Elkner is a middle- to high-elevation soil with a well-developed profile, while

Delineation of soil continuum into discrete polygons also assumes that soil properties change dramatically at the boundaries of these polygons (**Figure 3**). As shown in **Figure 3**, the actual value of the continuous soil property (*v*) might exhibit variation outlined by the dashed curve.

Studies have shown that even soil maps produced from systematic surveys might contain large amounts of "impurities" (errors) within the units delineated (Beckett, 1971; Beckett and Burrough, 1971; and Wilding *et al.*, 1965). These "impurities" might propagate through geographical analysis in a GIS and make the results from these analyses, particularly GIS-based simulation modelling, unpredictable.

While the soil scientists conducting the survey and mapping process may be aware of the degree of uncertainty in the mapping and may be able to infer more details on expected soil properties at a given location on the map from their knowledge of soil/landscape relations, this information is generally not translated onto soil maps. It is feasible to use the knowledge-based data integration approach to incorporate this information into the soil information generation processes.

Study Site

The study area for testing the knowledge-based data integration is the south part of the Lubrecht Experimental Forest, located about 50 km northeast of Missoula, Montana, USA. The forest was established in 1937 to foster research on natural resources. The study area is centred on the North Fork of Elk Creek, and is 4.8 km in the north/south direction and 7.5 km in the east/west direction (160 rows and 250 columns with a pixel size of 30 m by 30 m).

The elevation in the area ranges from 1,160 m to 1,930 m with high elevation in the east and southwest and low elevation in the northwest (**Figure 4**). The majority of areas are at an elevation between 1,200 m to 1,510 m. There is no evidence of glaciation in the study area (Brenner, 1968). Ross and Hunter (1976) estimated that the mean annual precipitation at 1,830 m is between 50 cm and 76 cm. Annual precipitation at lower elevations is lower. Annual precipitation at 1,200 m is between 37 cm and 56 cm. Approximately 44% of the precipitation falls during the winter (November through March) and 24 percent falls during the summer (June through August) (Nimlos, 1986). The mean annual air temperature at 1,200 m is about 4°C. The mean winter air temperature (December through February) is about -6°C, while the mean summer air temperature (June through August) is about 15°C. The study area can be considered as semi-humid to semi-arid.

Most of the mountain slopes in the study area are forested, dominated by Douglas-fir (*Pseudotsuga menziesii*), although lesser amounts of Western Larch (*Larix occidentalis*) and Ponderosa Pine (*Pinus ponderosa*) are present. A small portion of the study area (the northwest) is covered by fescue/bluebunch wheatgrass. Only a few areas of virgin forest remain, most having been logged, mainly during the periods of 1904 to 1906, 1925 to 1935 (Cauvin, 1961), and 1961 to 1962 (Brenner, 1968). Much of the timber is second-growth. There have been no large wild fires in the study area since 1937.

Four major types of parent materials are in the area: belt rocks, granite, limestone, and recent alluvium (Brenner, 1968; and Nimlos, 1986). The alluvium materials occur only in limited areas along the North and South Forks of Elk Creek. The first three parent materials make up the majority of the study area, with belt rocks in the north, granite in the south, and limestone through the central part.

Soils on these three materials are formed from a mantle of colluvium. There are four soil orders in the general area of the Lubrecht Experiment Forest: Alfisols, Entisols, Inceptisols, and Mollisols (Nimlos, 1986), and only two orders — Entisol and Inceptisol — are found in the study area. Entisols are weakly developed soils with very little organic matter accumulation and no illuvial clay or sesquioxides. In Lubrecht, Entisols are usually found on ridge crests. Inceptisols are young soils with little or no illuviated clays but brown subsoil horizons that indicate some translocations of sesquioxides. About 90% of the soils (in terms of areal extent) in the study area are Inceptisols. The 12 mapped soil series in the study area are: Ambrant, Elkner, Evaro, Ovando, Repp, Rochester, Sharrott, Tevis, Trapps, Whitore, Winkler, and Winkler Cool. Brief descriptions of these soil series are given as follows.

Ambrant, Elkner, Ovando, and Rochester are the soil series occurring on the granite materials that are located in the southeast part of the study area. Ambrant and Rochester are low-elevation soil series and often occur intermittently. Ambrant has a better developed and a deeper profile, while Rochester has a poor profile development and contains a large amount of rock fragments. Rochester often occurs at the dry sites where limited moisture conditions restrict the development of the soil profile. Elkner is a middle- to high-elevation soil with a well-developed profile, while

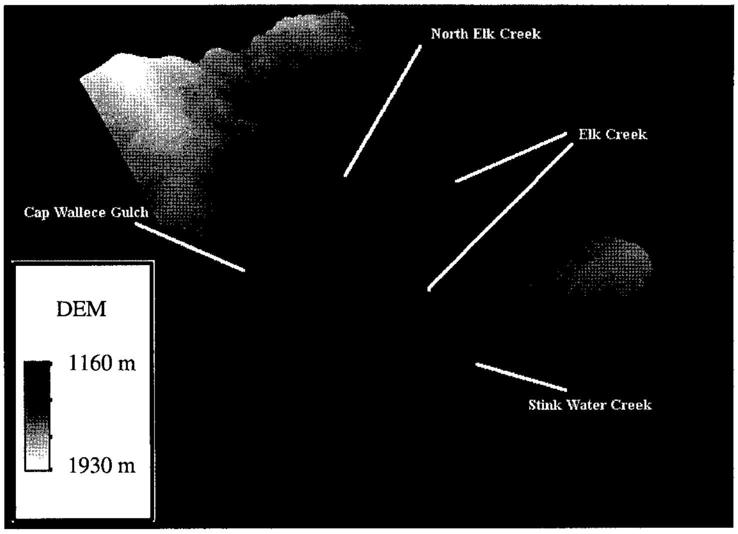

Figure 4.
A Digital Elevation Model of the study area.

Ovando is located at high elevation where cool temperatures limit the soil formation processes.

Soil series Evaro, Sharrott, Tevis, Winkler, and Winkler Cool developed on the belt materials, which are located in the north and west part of the study area. Evaro is a high-elevation soil with a well-developed profile. Sharrott and Winkler are low-elevation soils, which often occur intermittently on the belt parent materials. They are the counterparts of Rochester and Ambrant on belt materials. Tevis is a mid-elevation soil with a reasonably deep profile. Winkler Cool is a soil on north-, northeast-, and northwest-facing slopes at low elevation. It has a better profile development than Winkler.

Repp, Trapps, and Whitore are limestone soils, which occur in a narrow zone between the belt soils in the north and the west and the granite soils in the southeast. Repp is the least-developed soil on the limestone material and is often found on the south- and southwest-facing slopes at low elevations. Whitore is the best-developed soil on limestone material and is found at high elevations. Trapps, a mid-elevation soil, occurs on the north-facing slopes at low elevations. Its profile development is between Repp and Whitore.

Data Source

Six data variables (elevation, parent material, aspect, canopy coverage, gradient, and surface profile curvature) were employed in this study to characterize the soil-forming environment. The inclusion of these variables in this study was determined during the knowledge-acquisition process where the knowledge engineer and the soil expert discussed the importance and feasibility of each data variable and formed a list of variables to be used (Zhu, 1994). In the data variable list, there are no data variables that directly measure climatic factors. Although the study was conducted on a small drainage basin, great differences in terms of micro-climate do exist within the basin. However, these differences in micro-climate are well expressed by variations in elevation, aspect, and gradient. Therefore, climate variables are not included in the data variable list.

Information on elevation, aspect, gradient, and surface profile curvature were obtained from a Digital Elevation Model (DEM) of the study area. The DEM was supplied by the GIS laboratory of the School of Forestry, University of Montana. The accuracy of the DEM is comparable with the accuracy of the level-one USGS seven and one-half-minute DEMs (USGS, 1990). The canopy coverage information was approximated with an index derived from Landsat Thematic Mapper (TM) (see next section for details). Information on parent material was obtained from the geological map of the study area (Brenner, 1968). The inclusion of soil parent material from geological maps would also affect the performance of the approach because geological maps were produced in the same way as soil maps and therefore potentially contain human errors. The bedrock geology cannot truly represent the surficial geology from which the soils were developed. However, the soil parent materials in the study area play such an important role in the development of soils in the area that the inclusion of this parent material information is essential for soil inference. The other reason for the use of bedrock geology as soil parent material information is that the bedrock geology was the only geological information available for the area to be used as soil parent materials.

These data have by no means exhausted the soil formation factors and the interaction of these factors on soil development. They were used to demonstrate the potential of the new methodology of soil information-gathering and representation. These environmental data are currently feasible to be derived from a GIS data base or using some GIS techniques, and are meaningful to the soil expert.

In this study, the knowledge of the relationships between soils and the environment was drawn from a certified soil scientist, Barry Dutton, President of Land & Water Consulting Inc. Mr. Dutton was chosen as the domain expert for this study because he has worked extensively in the study area and understands the interaction between soil and the environment very well. The knowledge acquisition was performed using interactive, stepwise-structured interviews. The entire process of knowledge acquisition is fully discussed in Zhu (1994).

Syntax Data Integration

In this study, six environmental variables (parent material, elevation, slope aspect, slope gradient, canopy coverage, and surface profile curvature) were employed to characterize the soil-forming environment. These variables can be grouped into three general categories: the geology variable, terrain variables, and the vegetation variable.

Information on elevation, aspect, gradient, and surface profile curvature were obtained from a Digital Elevation Model (DEM) of the study area. The aspect and gradient were calculated using a third-order finite difference method (Horn, 1981) (see Chapter 3 of Burrough, 1986; Skidmore, 1989).

Canopy cover was approximated with an index derived from the TM images of the study area. Recent research suggests that reasonable estimates of canopy closure can be gained with middle infrared wavelengths (MIR) (Baret *et al.*, 1988; Butera, 1986). Nemani *et al.* (1993) have used the changes in MIR (TM band 5, 1.55–1.75 μm) response to canopy closure in combination with red to infrared ratios to estimate Leaf Area Index (LAI) in the study site. In this study, the MIR was used to estimate an index of canopy cover, where CC stands for canopy cover index, and MIR_{min} and MIR_{max} are middle infrared radiance from completely

$$CC = 100\,(1 - \frac{MIR - MIR_{\min}}{MIR_{\max} - MIR_{\min}}) \qquad (2)$$

closed and completely open canopies in/around the study area, respectively. Note that Equation 2 was not correlated with actual ground estimates of canopy cover and therefore should be considered as an index rather than an estimate of actual percentage of canopy cover.

Information on parent material was obtained by digitizing the bedrock geology map (with a scale of 1:50,000) of the study area (Brenner, 1968). The digitized map was converted to a raster image by overlaying a grid system on top of the vector map with each grid cell having a ground resolution of 30 m by 30 m.

All of these data layers were registered to Universal Transverse Mercator (UTM) coordinate systems (see Clarke, 1990, for details about UTM). The registration process was done in GRASS (Geographical Resources Analysis Support System) (U.S. Army Corps of Engineers, 1991). The data layers are stored in a raster-based data base since the raster model is more suitable for representing the continuous spatial variation of geographic phenomenon (such as soil). The pixel resolution is 30 m by 30 m. The values of each data layer are stored as an *n* x *m* matrix, where *n* is the number of rows (lines) and *m* is the number of columns of the grid used in a study area.

Semantic Data Integration

The theoretical basis for soil inference is based on Jenny's (1941, 1980) classic concept that a soil is a product of interaction among its forming factors over time. Since the time factor is often implicitly expressed by other environmental conditions (such as landform positions and vegetation), Jenny's concept can be simplified to "soil is a product of the interaction of physical environmental conditions." This simplified concept can be qualitatively expressed by

$$S = f(E) \qquad (3)$$

where S is soil information, E represents the vector of environmental conditions, and f represents the relationship between soils and the formative environments. It is possible to obtain information on the soil at a given location if the local environmental conditions and the relationships between soils and the environment are known.

E in Equation 3 was perceived to be made up of parent material, elevation, slope aspect, slope gradient, surface profile curvature, and canopy coverage (**Figure 5**). These data can be drawn from the raster GIS data base. However, the details of the relationship (f) are different at different places. It is very difficult at this stage to derive a mathematical formula for the relationship because of the complexity and limited understanding of both soil-forming processes and the paleo-environment. It was believed that local soil scientists who study and map soils in their respective regions have accumulated a detailed knowledge of soil/environment relationships, and this empirical knowledge can be used to approximate f in Equation 3 for soil inference. This empirical knowledge can be obtained through the application of some knowledge-acquisition techniques (Zhu, 1994). Soils show a great deal of gradual change over space, and the discrete spatial model (crisp logic) is not appropriate for the representation of soil information. The "=" in Equation 3 is perceived under fuzzy logic in this study, and the inference processes and the presentation of results were conducted under fuzzy logic. S in Equation 3 represents the membership vector (called Soil Similarity Vector in this case) for a set of prescribed soil taxonomic units, soil series.

The inference process is illustrated in **Figure 6**. At any given point in the study area, the inference system retrieves a set of environmental conditions (E) from the GIS data base and combines these conditions with their respective functions (f^k) from the knowledge base for soil series k to calculate a set of membership values (O^k)

$$O^k = f^k(E) \qquad (4)$$

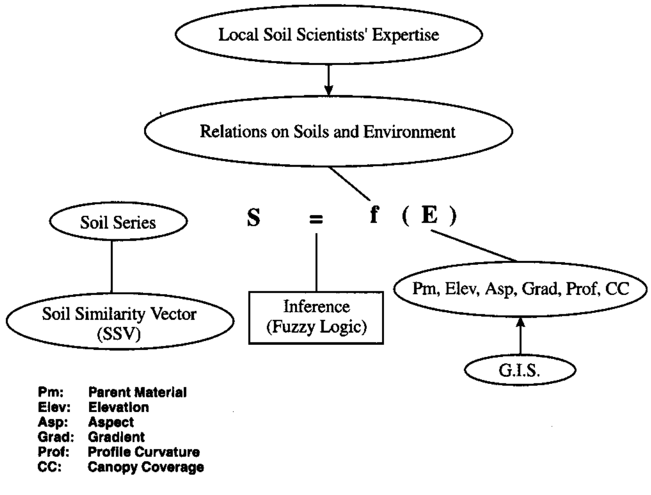

Figure 5.
Implementation of knowledge-based semantic data integration for soil inference.

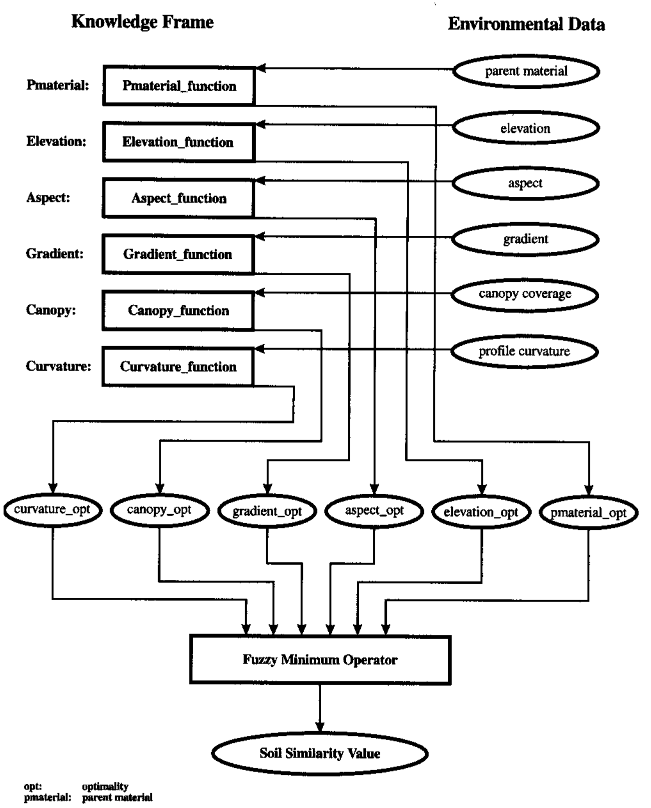

Figure 6.
The inference process.

where O^k is a vector of membership values and its *i*th element, o_ik, represents the membership for soil series k to develop under the *i*th environmental condition (e_i) in environmental vector E. Therefore, o_ik is also called the partial membership (optimality value), and f^k represents a vector of membership functions defined in the knowledge base for soil series k with respect to each of the environmental variables. Each (o_ik) of the partial membership values was calculated by substituting the environmental condition (e_i) into its respective membership function (f_ik). The resultant value o_ik from this calculation is considered to be the similarity between the soil at the site and soil series k with respect to environmental condition e_i. If there are m functions in the knowledge base for soil series k, then there would be m such optimality values (that is, o_1k ... o_mk). The fuzzy minimum operator was used to integrate these partial membership values and to produce the final membership value for the soil at the given point being soil series k

$$S^k = \min(O_1^k, O_2^k, \ldots, O_i^k, \ldots, O_m^k) \qquad (5)$$

where s^k is the membership value (similarity value) of the soil at a point being soil series k and m is the number of environmental variables involved. This membership expresses the degree of similarity of the soil at the given point to soil series k. The reason for using the fuzzy minimum operator for the derivation of the membership value is that the development of a soil is assumed to be controlled by the least favourable environmental factor (the limiting factor), and the smallest optimality value from the set of optimality values should be used to represent the membership value. It is understood that the interaction among the soil formation factors may be more complicated than can be modelled by the fuzzy minimum operator. However, little is known about the exact forms of the interactions and the interactions vary greatly from one environment to the other. Therefore, for simplicity, the limiting factor assumption was employed and the fuzzy minimum operator was used to model the limiting factor interaction.

The inference process is repeated for the next soil series until all soil series are exhausted. At this point the inference system stores the membership vector and continues on to the next pixel in the area. The process stops when all pixels in the area are visited.

Soil information generated from the inference system is represented in the form of S (Soil Similarity Vector). Each element (s^k) of S is an image of membership values that represent the similarities of the soils in the area to the respective soil series (k). At any given point (i, j) in the study area, the information about the soil at this point is represented by S_{ij}, that is, (s_{ij}^1, s_{ij}^2 ... s_{ij}^k ... s_{ij}^n), where n is the number of prescribed soil series. Element $s_{ij}k$ of S_{ij} represents the similarity of soil at point (i, j) to a prescribed soil series (k). This representation of soil information is useful because the similarities of soil information to different soil series can be maintained and soil properties that are intermediate to the typical values of the prescribed soil series may be derived from the vector. The information in this vector also allows the expression of soil gradation over space because the subtle difference between soils at neighbouring pixels can be reflected in the changes of membership distribution in the vector. This is significantly different from the information in standard soil maps in which the difference between

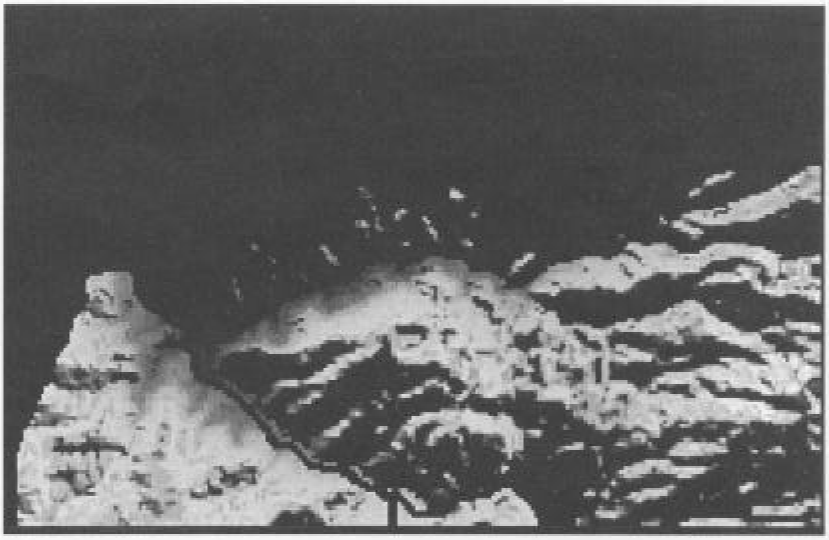

Figure 7.
Membership map of Elkner-Ovando complex.

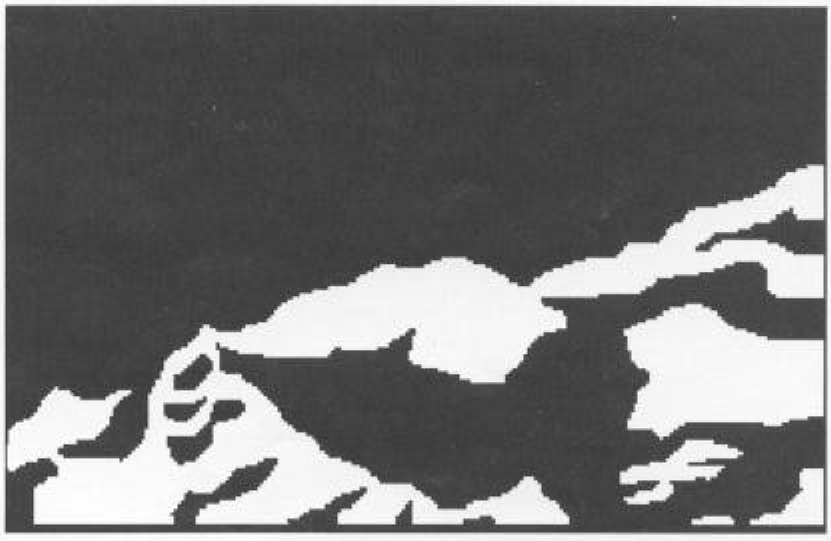

Figure 8.
Soil map of Elkner-Ovando complex.

where O^k is a vector of membership values and its ith element, o_ik, represents the membership for soil series k to develop under the ith environmental condition (e_i) in environmental vector E. Therefore, o_ik is also called the partial membership (optimality value), and f^k represents a vector of membership functions defined in the knowledge base for soil series k with respect to each of the environmental variables. Each (o_ik) of the partial membership values was calculated by substituting the environmental condition (e_i) into its respective membership function (f_ik). The resultant value o_ik from this calculation is considered to be the similarity between the soil at the site and soil series k with respect to environmental condition e_i. If there are m functions in the knowledge base for soil series k, then there would be m such optimality values (that is, o_1k ... o_mk). The fuzzy minimum operator was used to integrate these partial membership values and to produce the final membership value for the soil at the given point being soil series k

$$S^k = \min(O_1^k, O_2^k, \ldots, O_i^k, \ldots, O_m^k) \tag{5}$$

where s^k is the membership value (similarity value) of the soil at a point being soil series k and m is the number of environmental variables involved. This membership expresses the degree of similarity of the soil at the given point to soil series k. The reason for using the fuzzy minimum operator for the derivation of the membership value is that the development of a soil is assumed to be controlled by the least favourable environmental factor (the limiting factor), and the smallest optimality value from the set of optimality values should be used to represent the membership value. It is understood that the interaction among the soil formation factors may be more complicated than can be modelled by the fuzzy minimum operator. However, little is known about the exact forms of the interactions and the interactions vary greatly from one environment to the other. Therefore, for simplicity, the limiting factor assumption was employed and the fuzzy minimum operator was used to model the limiting factor interaction.

The inference process is repeated for the next soil series until all soil series are exhausted. At this point the inference system stores the membership vector and continues on to the next pixel in the area. The process stops when all pixels in the area are visited.

Soil information generated from the inference system is represented in the form of S (Soil Similarity Vector). Each element (s^k) of S is an image of membership values that represent the similarities of the soils in the area to the respective soil series (k). At any given point (i, j) in the study area, the information about the soil at this point is represented by S_{ij}, that is, (s_{ij}^1, s_{ij}^2 ... s_{ij}^k ... s_{ij}^n), where n is the number of prescribed soil series. Element $s_{ij}k$ of S_{ij} represents the similarity of soil at point (i, j) to a prescribed soil series (k). This representation of soil information is useful because the similarities of soil information to different soil series can be maintained and soil properties that are intermediate to the typical values of the prescribed soil series may be derived from the vector. The information in this vector also allows the expression of soil gradation over space because the subtle difference between soils at neighbouring pixels can be reflected in the changes of membership distribution in the vector. This is significantly different from the information in standard soil maps in which the difference between

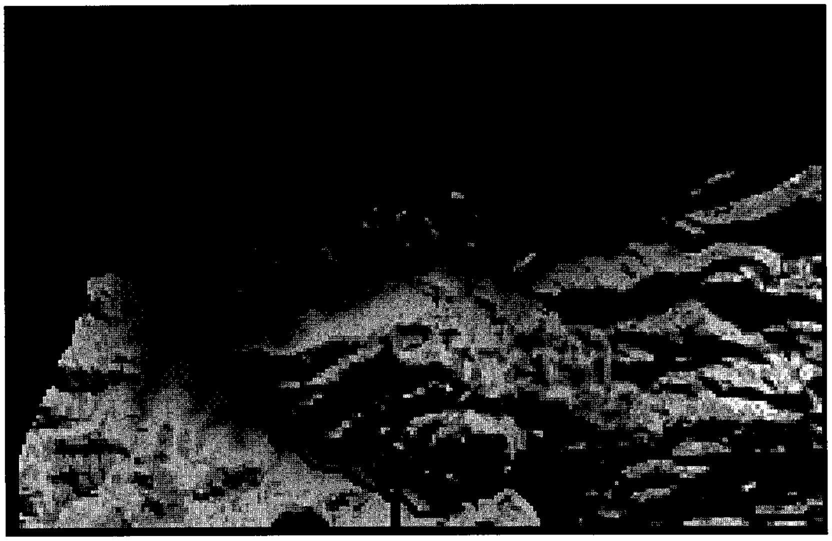

Figure 7.
Membership map of Elkner-Ovando complex.

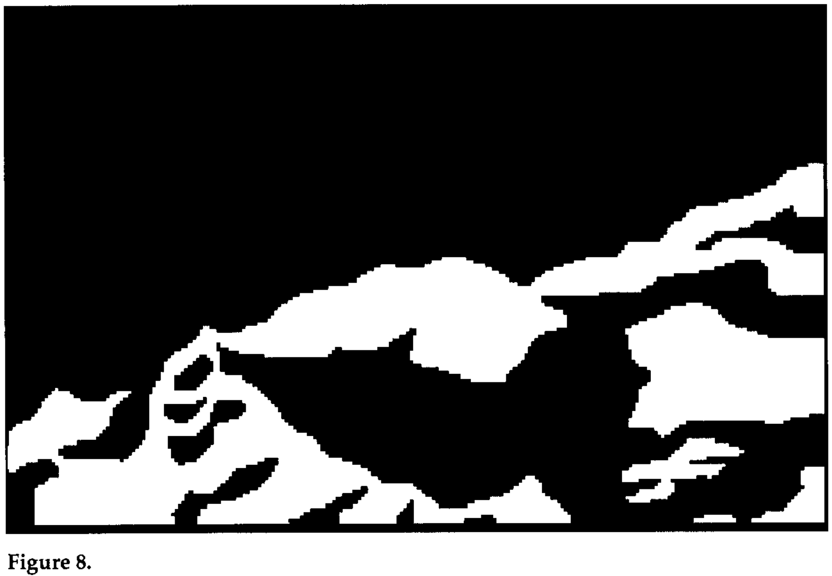

Figure 8.
Soil map of Elkner-Ovando complex.

neighbouring pixels is either exaggerated to the difference between two different soil series or is completely ignored.

Inference Results and Discussion

The membership values over space for soil series complex Elkner-Ovando are illustrated in **Figure 7**. The difference between **Figure 7** and the soil map of Elkner-Ovando (**Figure 8**) is clear. The membership image shows the gradation of Elkner-Ovando over space, while there is no difference in the areas where the soil is mapped as Elkner-Ovando in **Figure 8**.

As *S* gives a set of similarity measures to a prescribed set of soil series, it may be possible to approximate values of specific soil properties for each pixel using the information in *S*. In other words, *S* can be used to derive other soil information at the values intermediate to the values assigned to each of the prescribed soil series. In this demonstration, the *S* values generated from the inference system were used to derive depths of soil *A* horizons for demonstrating the usefulness of *S* and validating the presented semantic data integration approach.

For this demonstration, it was assumed the depth of *A* horizon at a given point is "influenced" by all of the soil series in the area. The degree of influence (contribution) depends on the similarity of the soil at the site to the given soil series. The more similar it is to the soil series, the more contribution that soil series has in the depth of *A* horizon of the soil at the site. It was also assumed, for a first approximation, that the contribution from different soil series in the depth of *A* horizon at the site is linearly additive, which can be expressed in Equation 6:

$$D_{ij} = \frac{\sum_{k=1}^{n} S_{ij}^{k} \cdot d^{k}}{\sum_{k=1}^{n} S_{ij}^{k}} \qquad (6)$$

where D_{ij} is the depth of *A* horizon at site (*i*, *j*), s_{ij}^{k} is the similarity of the soil at site (*i*, *j*) to the prescribed soil series *k* and is an element of $\mathbf{S}_{ij}$, d^{k} is the prescribed depth of *A* horizon of soil series *k*, and *n* is the total number of prescribed soil series in the area. The prescribed depth of *A* horizon for the soil series in the study area was extracted from the soil series descriptions (Missoula County Soil Survey, 1983).

The inferred and the mapped depths of *A* horizons of the soils in the study area are shown in **Figures 9** and **10**. The contrast between the two maps is strong and clear. The image of inferred depth shows a spatially continuous pattern of *A* horizon depth, while the image of mapped depth inherits the exact spatial pattern o the soil map. Since the study area is in the semi-humid tc semi-arid area of western Montana, the soils on the north-facing slopes and at high elevations develop better than the soils on the south-facing slopes and at lower elevations due to limited moisture conditions at low elevations and or south-facing slopes. The *A* horizons are deep for these better-developed soils. The inferred depth map shows this characteristic of the depth of *A* horizon in this area, but the change of depth is gradual over space and is not like the change depicted in the depth image derived from the soi map.

During the summer of 1993, 33 field observations on soi *A* horizon depths were used to compare the results from the

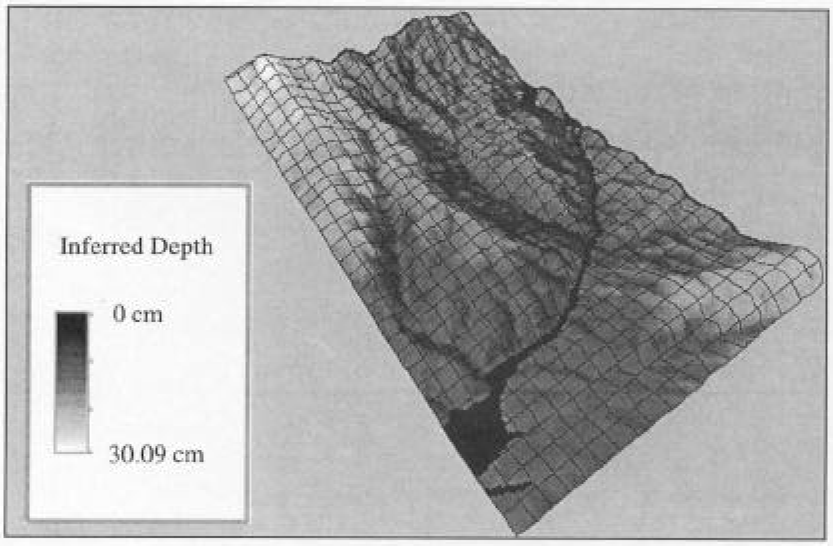

Figure 9.
Inferred *A* horizon depth of the study area.

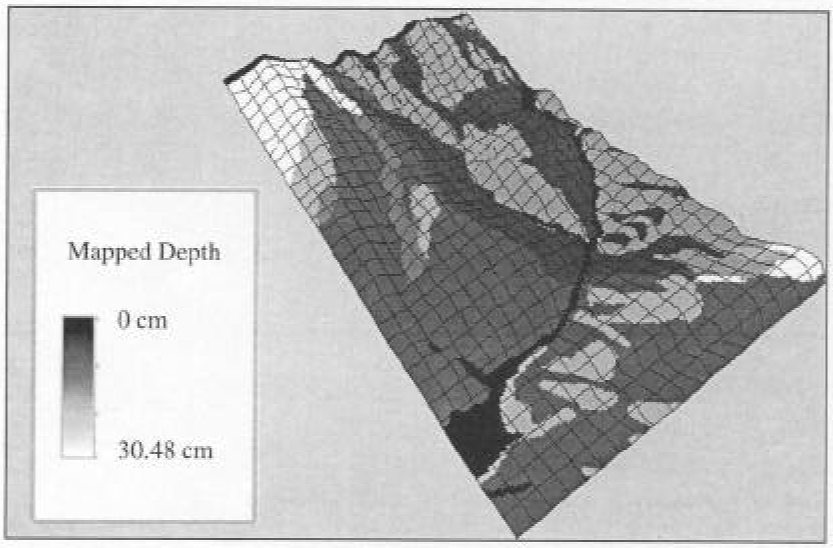

Figure 10.
***A* horizon depth derived from a conventional soil map of the study area.**

neighbouring pixels is either exaggerated to the difference between two different soil series or is completely ignored.

Inference Results and Discussion

The membership values over space for soil series complex Elkner-Ovando are illustrated in **Figure 7**. The difference between **Figure 7** and the soil map of Elkner-Ovando (**Figure 8**) is clear. The membership image shows the gradation of Elkner-Ovando over space, while there is no difference in the areas where the soil is mapped as Elkner-Ovando in **Figure 8**.

As *S* gives a set of similarity measures to a prescribed set of soil series, it may be possible to approximate values of specific soil properties for each pixel using the information in *S*. In other words, *S* can be used to derive other soil information at the values intermediate to the values assigned to each of the prescribed soil series. In this demonstration, the *S* values generated from the inference system were used to derive depths of soil *A* horizons for demonstrating the usefulness of *S* and validating the presented semantic data integration approach.

For this demonstration, it was assumed the depth of *A* horizon at a given point is "influenced" by all of the soil series in the area. The degree of influence (contribution) depends on the similarity of the soil at the site to the given soil series. The more similar it is to the soil series, the more contribution that soil series has in the depth of *A* horizon of the soil at the site. It was also assumed, for a first approximation, that the contribution from different soil series in the depth of *A* horizon at the site is linearly additive, which can be expressed in Equation 6:

$$D_{ij} = \frac{\sum_{k=1}^{n} s_{ij}^{k} \cdot d^{k}}{\sum_{k=1}^{n} s_{ij}^{k}} \qquad (6)$$

where D_{ij} is the depth of *A* horizon at site (i, j), s_{ij}^{k} is the similarity of the soil at site (i, j) to the prescribed soil series *k* and is an element of S_{ij}, d^{k} is the prescribed depth of *A* horizon of soil series *k*, and *n* is the total number of prescribed soil series in the area. The prescribed depth of *A* horizon for the soil series in the study area was extracted from the soil series descriptions (Missoula County Soil Survey, 1983).

The inferred and the mapped depths of *A* horizons of the soils in the study area are shown in **Figures 9** and **10**. The contrast between the two maps is strong and clear. The image of inferred depth shows a spatially continuous pattern of *A* horizon depth, while the image of mapped depth inherits the exact spatial pattern of the soil map. Since the study area is in the semi-humid to semi-arid area of western Montana, the soils on the north-facing slopes and at high elevations develop better than the soils on the south-facing slopes and at lower elevations due to limited moisture conditions at low elevations and on south-facing slopes. The *A* horizons are deep for these better-developed soils. The inferred depth map shows this characteristic of the depth of *A* horizon in this area, but the change of depth is gradual over space and is not like the change depicted in the depth image derived from the soil map.

During the summer of 1993, 33 field observations on soil *A* horizon depths were used to compare the results from the

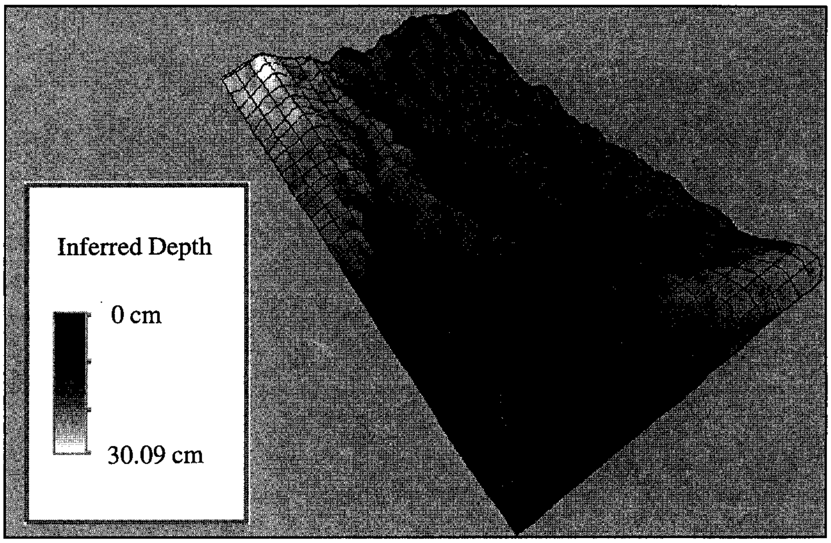

Figure 9.
Inferred *A* horizon depth of the study area.

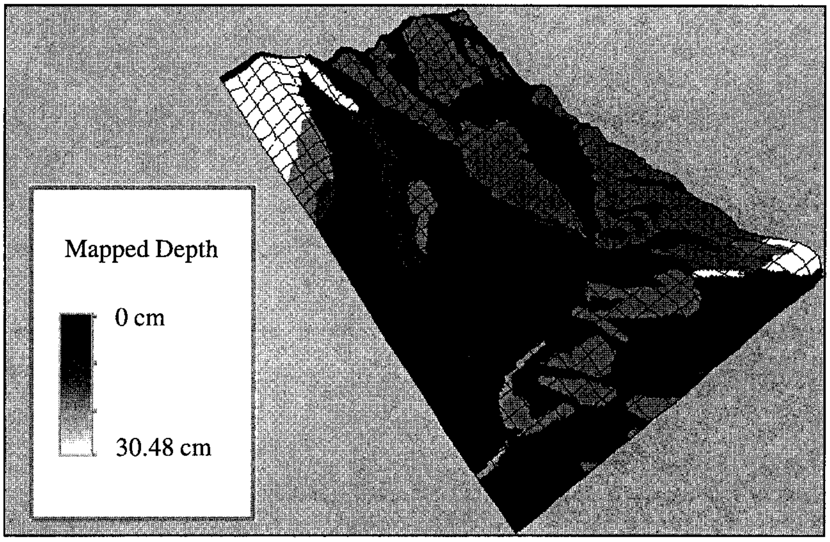

Figure 10.
***A* horizon depth derived from a conventional soil map of the study area.**

knowledge-based data integration approach. Three major statistics, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Agreement Coefficient (AC) (Willmott, 1984; Willmott *et al.*, 1985) were used for the comparison. Statistics for the evaluation of the performance of the data integration method versus the soil map, listed in **Table 1**, show that the prediction of *A* horizon depth derived using the knowledge-based data integration approach is better than using the soil map. In a subsequent paper, a detailed discussion on the comparison between the results from the knowledge-based integration approach and field observations will be presented. The result of this approach when applied to an Australian watershed will also be presented and discussed in detail in the upcoming paper.

Table 1.
Summary of performance measures for the data integration method and the soil map.

Quantitative Measures	MAE	RMSE	AC
The Knowledge-based Approach	3.0448	4.0024	0.8533
Soil Map	4.4064	5.9249	0.8033

SUMMARY

A model for integrating multi-source spatial data for geographical analysis is presented in this paper. The model integrates the semantics of the data layers with the data analysis process. In particular, the model uses empirical knowledge about the relationships between environmental data and a given geographic phenomenon to guide the integration of multi-source data. A soil inference example has demonstrated that the approach is useful and is capable of producing high-quality soil information. The approach does rely heavily on empirical knowledge about the relationships between a given phenomenon and environmental conditions. The success of this approach largely depends on the quality of this empirical knowledge.

ACKNOWLEDGEMENTS

The research conducted in this paper was financially supported by University of Toronto open doctoral fellowships, by teaching assistantships to Dr. A-Xing Zhu while he was a student at University of Toronto, and by grants from NSERC and NASA. We thank the GIS laboratory, School of Forestry, University of Montana for providing the GIS data used in this paper. We also thank Barry Dutton for his co-operation during the knowledge-acquisition processes.

REFERENCES

Aronoff, S. 1989. *Geographic Information Systems: A Management Perspective.* Ottawa, Canada: WDL Publication, 294 pp.

Band, L.E., D.L. Peterson, S.W. Running, J.C. Coughlan, R.B. Lammers, J. Dungan, and R. Nemani. 1991. "Ecosystem Processes at the Watershed Scale: Basis for Distributed Simulation," *Ecological Modeling*, Vol. 56, pp. 151–176.

Band, L.E., J.P. Patterson, R. Nemani, and S.W. Running. 1993. "Ecosystem Processes at the Watershed Scale: Incorporating Hill-Slope Hydrology," *Agricultural and Forest Meteorology*, Vol. 63, pp. 93–126.

Baret, F., G. Guyot, A. Begue, P. Maurel, and A. Podaire. 1988. "Complementarity of Middle Infrared with Visible and Near-Infrared Reflectance for Monitoring Wheat Canopies," *Remote Sensing of Environment*, Vol. 26, pp. 213–225.

Beckett, P.H.T. 1971. "The Cost-effectiveness of Soil Survey," *Outlook Agriculture*, Vol. 6, No. 5, pp. 191–198.

Beckett, P.H.T. and P.A. Burrough. 1971. "The Relation Between Cost and Utility in Soil Survey IV," *Journal of Soil Science*, Vol. 22, pp. 466–480.

Brenner, R.L. 1968. *The Geology of Lubrecht Experimental Forest*, Lubrecht Series One. University of Montana, Missoula: The Montana Forest and Conservation Experiment Station, School of Forestry, 71 pp.

Bregt, A.K. 1992. *Processing of Soil Survey Data*, doctoral thesis. Wageningen, The Netherlands: Wageningen Agricultural University, 167 pp.

Burrough, P.A. 1986. *Principles of Geographical Information Systems for Land Resources Assessment.* Oxford: Clarendon Press, 193 pp.

Butera, C. 1986. "A Correlation and Regression Analysis of Per Cent Canopy Closure and TM Spectral Response for Selected Forested Sites in San Juan National Forest, Colorado," *IEEE Transactions on Geosciences and Remote Sensing*, GE-24(1), pp. 122–129.

Cauvin, D.M. 1961. *Management Plan for Lubrecht Forest*, Master's thesis. University of Montana, Missoula: School of Forestry.

Clarke, K.C. 1990. *Analytical and Computer Cartography.* Englewood Cliffs, New Jersey: Prentice Hall, 290 pp.

Horn, B.K.P. 1981. "Hill Shading and the Reflectance Map," *Proceedings of the IEEE*, Vol. 69, No. 1, pp. 14–47.

Heuvelink, G.B.M. 1993. *Error Propagation in Quantitative Spatial Modeling Applications in Geographical Information Systems.* Utrecht, Nederlands: Faculteit Ruimtelijke, Wetenschappen Universiteit, 150 pp.

Jenny, H. 1941. *Factors of Soil Formation: A System of Quantitative Pedology.* New York: McGraw-Hill, 281 pp.

Jenny, H. 1980. *The Soil Resource: Origin and Behaviour.* New York, Berlin: Springer-Verlag, 377 pp.

Jensen, J.R. 1986. *Introductory Digital Image Processing: A Remote Sensing Perspective.* Englewood Cliffs, New Jersey: Prentice-Hall, 379 pp.

Missoula County Soil Survey. 1983. *Soil Series Description.* Missoula, Montana, 117 pp.

Maling, D.H. 1991. "Coordinate Systems and Map Projections for GIS," *Geographical Information Systems: Principles and Applications,* D.J. Maguire, M.F. Goodchild, and D.W. Rhind (eds.). New York: Longman Scientific & Technical, pp. 135–146.

Nemani, R., L. Pierce, S. Running, and L. Band. 1993. "Forest Ecosystem Processes at the Watershed Scale: Sensitivity to Remotely Sensed Leaf Area Index Estimates," *International Journal of Remote Sensing,* Vol. 14, No. 13, pp. 2519–2534.

Nichols, D.A. 1981. "Conversion of Raster Coded Images to Polygonal Data Structures," *Proceedings of the PECORA VII Symposium held in Sioux Falls, South Dakota,* October 18–24, 1981. Falls Church: American Scociety of Photogrammetry, pp. 508–515.

Nimlos, J.T. 1986. *Soils of Lubrecht Experimental Forest,* Miscellaneous Publication No. 44. Missoula, Montana: Montana Forest and Conservation Experiment Station, 36 pp.

Pavlidis, T. 1979. "Filling Algorithms for Raster Graphics," *Computer Graphics and Image Processing,* Vol. 10, 126 pp.

Peuquet, D.J. 1981a. "An Examination of Techniques for Reformatting Digital Cartographic Data. Part 1: The Raster-to-Vector Process," *Cartographica,* Vol. 18, pp. 21–33.

Peuquet, D.J. 1981b. "An Examination of Techniques for Reformatting Digital Cartographic Data. Part 2: The Vector-to-Raster Process," *Cartographica,* Vol. 18, pp. 34–48.

Piwowar, J.M, E.F. LeDrew, and D. Dudycha. 1990. "Integration of Spatial Data in Vector and Raster Formats in a Geographic Information System Environment," *International Journal of Geographical Information Systems,* Vol. 4, No. 4, pp. 429–444.

Ross, R.L. and H.E. Hunter. 1976. *Climax Vegetation of Montana Based on Soils and Climate.* Bozeman, Montana: USDA Soil Conservation Service, 36 pp.

Skidmore, Andrew K. 1989. "A Comparison of Techniques for Calculating Gradient and Aspect from a Gridded Digital Elevation Model," *International Journal of Geographical Information Systems,* Vol. 3, No. 4, pp. 323–334.

U.S. Army Corps of Engineers, Construction Engineering Research Laboratory (CERL). 1991. *GRASS 4.0 Reference Manual.* Champaign, Illinois.

U.S. Geological Survey, Department of the Interior. 1990. *Digital Elevation Models: Data User's Guide.* Reston, Virginia, 51 pp.

Van Der Knaap, W.G.M. 1992. "The Vector to Raster Conversion: (Mis)Use in Geographical Information Systems," *International Journal of Geographical Information Systems,* Vol. 6, No. 2, pp. 159–170.

Van Roessel, J.W. and E.A. Fosnight. 1985. "A Relational Approach to Vector Data Structure Conversion," *Proceedings of the Seventh International Symposium on Computer-Assisted Cartography held in Washington, D.C., March 11–14,* 1985. Falls Church: American Society of Photogrammetry, American Congress on Surveying and Mapping, pp. 541–551.

Wilding, L.P., R.B. Jones, and G.M. Schafer. 1965. "Variation of Soil Morphological Properties Within Miami, Celina and Crosby Mapping Units in West-Central Ohio," *Proceedings of Soil Science Society of America,* Vol. 29, No. 6, pp. 711–717.

Willmott, C.J. 1984. "On The Evaluation of Model Performance in Physical Geography," *Spatial Statistics and Models,* Gary L. Gaile and Cort J. Willmott (eds.). D. Reidel Publishing Company, pp. 443–460.

Willmott, C.J., S.G. Ackleson, R.E. Davis, J.J. Feddema, K.M. Klink, D.R. Legates, J. O'Donnell, and C.M. Rowe. 1985. "Statistics for the Evaluation and Comparison of Models," *Journal of Geophysical Research,* Vol. 90, No. C5, pp. 8995–9005.

Zhu, A.X. 1994. *Soil Pattern Inference Using GIS Under Fuzzy Logic,* Ph.D. Dissertation. University of Toronto, Toronto, Ontario, 177 pp.