COMPUTATION WITH IMPRECISE GEOSPATIAL DATA

MICHAEL WORBOYS

Department of Computer Science, Keele University, Staffs ST5 5BG UK
email: michael@cs.keele.ac.uk

ABSTRACT

Imprecision in spatial data arises from the granularity or resolution at which observations of phenomena are made, and from the limitations imposed by computational representations, processing and presentational media. Precision is an important component of spatial data quality, and a key to appropriate integration of collections of data sets. Previous work of the author provides a theoretical foundation for imprecision of spatial data resulting from finite granularities, and gives the beginnings of an approach to reasoning with such data using methods similar to rough set theory. This paper develops the theory further, and extends the work to a model that includes both spatial and semantic components. Notions such as observation, schema, frame of discernment and vagueness are examined and formalised.

Keywords: GIS, imprecision, vagueness, resolution, spatial reasoning, data quality, uncertainty.

1. INTRODUCTION

We now live in a world of distributed systems, where the data that are needed to serve particular applications may be scattered globally and distributed through many systems. Absolutely fundamental to effective distributed computation is the notion of *data integration*; we must be able to bring data sets together and provide appropriate levels of interoperability between nodes of a globally distributed network. For interoperability to succeed, appropriate attention needs to be paid to the characteristics of each data set, described for example in terms of its data model and quality indicators. However, a major obstacle to achieving effective data integration lies in the various data sources being *semantically heterogeneous,* lacking uniformity in the 'meaning, interpretation or intended use' of the data in the distributed multi-database collection (Sheth and Larsen, 1990).

The primary motivation for this work is the need (Goodchild 1993) for a well-founded theory of data quality that is applicable to distributed collections of geospatial data. Imprecision is clearly a major dimension of spatial data quality, arising from the granularity or resolution at which observations of phenomena are made, and from limitations imposed by computational representation, processing and presentation. Imprecision here is to be distinguished from inaccuracy, in the sense of error or deviation from some notional real world value. Also, we exclude from consideration in this paper imprecision resulting from inherent vagueness in the geographic phenomena themselves. Scope is limited to discussion of imprecision resulting from limitations imposed by the contexts of observations of the phenomena.

An approach is needed that will provide robust measures of imprecision associated with geospatial data sets that can be used with both single data sets and distributed heterogeneous collections. Indeed, given such a collection, each with its own data model and associated levels of imprecision, the challenge is to determine what can be known about the integrated product. In order to meet this challenge, we must understand the components of imprecision in individual geospatial data sets, and how they work in combination; we must be able to make inferences about the integrated whole from knowledge of properties of its parts. Imprecision leads to vagueness, and part of the work concerns the link between these two concepts, and the ways in which

reasoning with vague geospatial entities may be approached. Previous work of the author (Worboys 1998) provides the beginnings of a theoretical foundation for reasoning with uncertain information due to spatial imprecision. This paper continues to develop a general theory of spatial imprecision and multi-resolution, concentrating on incorporation of semantic and geometric components.

The process of generalization is closely connected with these concerns, and may be thought of as applying transformations that effect integration between finer and coarser levels of detail. In this way, generalization will provide special cases for consideration below. Cartographic generalization has been the subject of a great deal of research by the cartographic and GIS communities, particularly on the geometric aspects of the generalization process (see for example Buttenfield and McMaster 1991, Müller *et al.* 1995). It has been recognised (Müller *et al.* 1995) that there are at least two dimensions of geospatial data generalization, described in (Dettori and Puppo 1996) as cartographic and model-oriented. In the work that follows, we shall make a similar dichotomy between spatial and semantic components of imprecision in a geospatial data set. Previous work on multi-resolution spatial data models includes that of Puppo and Dettori (1995), and Rigaux and Scholl (1995).

2. OBSERVATIONS, CONTEXTS AND SCHEMATA

The starting point for our work is the notion of an *observation* of a phenomenon in a source domain. The term ‘observation’ is intended to be interpreted by the reader in a rather general setting, certainly including more than just the data collection phase. Observations may or may not change the phenomena that are being observed. The source domain might be the domain of application, in which case an observation might be an on-site measurement or data collection for input to the computer. The target domain could be within the computational system and an observation could be data manipulation, data retrieval, or visualization of results of a computational analysis.

An observation always takes place in a *context*, which provides the framework within which the phenomenon is observed. It is the limitation placed on the observation by association with its context that leads to imprecision. A context is represented in this work by a *schema*, and we will be particularly concerned with schemata representing

spatial and semantic contexts. A context may be determined by the observer's location in space-time and the nature of the observation process; and schemata provide specifications of contexts. Figure 1 shows the various categories of schemata that will be discussed. Spatial data models are traditionally divided into field-based and object-based approaches (Couclelis 1992). Field-based approaches consider a spatial phenomenon as a collection of variations of attributes over a spatial framework. The object-based approach applies equally to spatial and non-spatial dimensions, and the phenomenon is viewed as a collection of related elements (objects) with spatial and non-spatial properties. There are two aspects of a schema that have a direct effect upon the precision of the associated operation, namely *extent* and *granularity*.

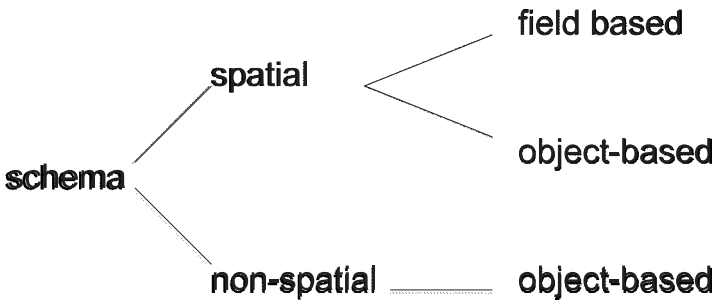

Figure 1: Categories of schemata for observations of geographic phenomena

The *extent* of a schema is the totality of elements expressible in the schema. For example, in the case of a semantic schema, its extent would include the collection of taxonomic constructs in the model, maybe expressed as a collection of object types. The semantic extent provides the 'vocabulary' in which the observation may be expressed. For a spatial context, the extent will be the total area spanned by the observation, maybe a country, road network or the domain of a remotely sensed image. The spatial extent specifies the global spatial 'world' (usually a region) in which the phenomenon is placed. Aspects of the phenomenon outside the boundaries of that world will not be visible to the observer. *Incompleteness* arises when pertinent data is missed by an observation due to the limitations of the extent of its schema. For a simple example, a census would miss the (hypothetical) trend of young people to move out of cities if there was no question about age on the census form. In this case, the semantic schema does not include attribute **age** of class **person** in its extent. Spatial incompleteness arises in a rather obvious similar manner, as it is clear that an observation will not include information about a location outside the spatial extent of

its context. A survey of English soccer teams will provide no direct information about soccer teams in Scotland.

The *granularity* of a schema specifies the levels of detail obtainable for an observation using that schema. For the object-based semantic component of a context, the granularity may be expressed as the level of detail provided by the object classes available in the object inheritance hierarchy. In a spatial context, the granularity might relate to the resolution provided by the spatial framework, below which spatial elements become indistinguishable, one from another. *Imprecision* in an observation thus arises from limitations on the granularity of the schema under which the observation is made. For example, an object-based topographical data model whose only object classes for dwellings are **house** and **cottage** would not be able to distinguish a terraced house from a detached house. Or, a quadtree representing data to a certain granularity would not be able to be used to distinguish two elements contained within the finest grain.

So, speaking loosely, the extent of a schema specifies the size of the window on the observation - which classes of entities are visible, while the granularity determines the level of detail observable through the window - which entities are distinguishable from one another. Limitations in the extent and granularity of a schema lead to incompleteness and imprecision respectively in the observation made with respect to the schema. The next few sections will develop these themes further and more formally, with special emphasis on the semantic and spatial dimensions. In particular, we will develop a formal theory of imprecision (and to some degree incompleteness) based upon the extent and granularity of schemata that provide contexts for observation of geographic phenomena. The temporal dimensions of geographic space will not be considered in this paper.

Before leaving this section we make the important point that a phenomenon may itself be an observation of another phenomenon. This relativist position allows us to dispense with the notion of an absolute ‘real world’ with respect to which all observations are made, and to provide a framework in which all observations are relative having varying degrees of associated precision.

3. FRAMES OF DISCERNMENT

The key formal concept that provides a suitable framework for discussion of the imprecision inherent in an observation is that of *indiscernibility*. Intuitively, a finite collection[1] of elements is indiscernible with respect to a particular observation if any pair of elements in the collection cannot be distinguished from each other by the observation. So, for example, in a remotely sensed image of a region, any two locations on the ground which fall within the same pixel cannot be distinguished in characteristics by the observation.

Let us assume that we make an observation Ω of a particular collection S of elements. Indiscernibility of elements is dependent upon Ω, for it is clear that two elements which are distinguishable under one observation may not be distinguishable under another. As stated earlier, it is the context in which the observations are made that is crucial, and it is the nature of the indiscernibility relation that gives some idea of the precision of Ω and its associated context. Given an observation Ω of set S, we can treat indiscernibility as a binary relation ρ_Ω on set S, where $t \rho_\Omega s$ can be read as 't is indiscernible from s by observation Ω'. If no ambiguity results, we omit the subscript and refer to the indiscernibility relation as ρ.

The relation ρ will have certain natural properties, based upon our intuitive understanding of what it means for entities to be indiscernible. Relation ρ may be assumed to be *reflexive*, as from the meaning of the concept it is difficult to imagine an entity not being indiscernible from itself. It will in most cases be *symmetric*, in that if t is indiscernible from s, then it will usually be correct to treat s as indiscernible from t under the same observation. However, there are examples of observations whose associated indiscernibility relation is not symmetric, particularly when indiscernibility takes on characteristics of 'similarity'. For example, we might want to say that a figure is indiscernible from its ground, but that the ground is not indiscernible from the figure. Asymmetric indiscernibility relations are discussed by Slowinski and Vanderpooten (1995, 1996), and are not considered further here.

It will not always be the case that relation ρ is *transitive*, where if u is indiscernible from t and t is indiscernible from s then u is indiscernible from s. For example, when

[1] All sets considered in this paper are assumed to be finite.

indiscernibility is thought of as being related to 'nearness', if *u* is near to *t* and *t* is near to *s* then *u* may not be near to *s*. However, transitivity will be assured when indiscernibility is induced by a partition of the underlying set *S*, for example in the pixellation of a image, and transitivity will be assumed in much of what follows.

An indiscernibility relation ρ on set *S* with respect to observation Ω leads to a collection of subsets of *S*, $R_\Omega(S) = \{R_\Omega(s) \mid s \in S\}$ (or just $R(S)$ if there is no ambiguity), where each element $R_\Omega(s)$ is defined by $R_\Omega(s) = \{t \in S \mid t\ \rho_\Omega\ s\}$.The set $R_\Omega(S)$ is termed the *frame of discernment* of *S* with respect to Ω, following the usage in Shafer (1976). This term indicates an epistemological dimension to the concept, and this is quite appropriate as an observer of a phenomenon can only represent propositions concerning it within the frame of discernment: any finer detail would not be discernible within the observation Ω. In the case when indiscernibility is reflexive, symmetric and transitive, that is an equivalence relation, then the frame of discernment provides a partition of the set of elements of the phenomenon under observation.

So, given a phenomenon under observation, the granularity of the observation's schema induces a mapping (in the mathematical sense) of subsets of a source set (being the collection of constituents of the phenomenon) into a finite target set, where elements that are indiscernible by observation are identified in the target set. We may note that in general a single constituent of the phenomenon may belong to more than one element of the frame of discernment, when indiscernibility is an equivalence relation, the frame of discernment provides a partition of the source set.

3.1 SEMANTIC FRAMES OF DISCERNMENT

This section shows how semantic data models provide contexts for observations of phenomena. As with all contexts, imprecision is introduced and is describable using an indiscernibility relation and frame of discernment.

The semantic context of an observation is provided by the semantic data models that the observation is made with respect to. A *semantic data model* provides an implementation-independent representation of the structure of the information comprising the phenomenon under observation. The are several types of such model - see Peckham and Maryanski (1988) for a review - but maybe the one in most common

use is the so-called *entity-relationship* (ER) model (Chen 1976). In the ER approach, the structure of the information space is seen as composed of entities, each of which has an independent and uniquely identifiable existence in the application domain. Entities are describable by means of their attributes (for example, the name, boundary and population of a country), and have explicit relationships with other entities. Entities are grouped into entity types or classes, where members of the same class have the same attribute and relationship structure.

Imprecision will arise if the granularity of the semantic data model does not match the semantic detail of the phenomenon under observation. For example, if the phenomenon being studied consisted of flows on a heterogeneous transportation network (road and rail, say), and the model had only the single entity class **road**, then only total flows on roads could be observed, even if in the underlying phenomenon there were important distinctions between flows on major trunk roads, minor roads and inner-city roads. Such a level of detail just could not be captured, due to the imprecision imposed by the semantic model.

Reasoning with imprecision in semantic schemata is an important aspect of geospatial uncertainty management. It is only when imprecision in different schemata structuring distinct observations of the same phenomenon is understood can we be confident about the results of integration and generalisation. A simple example of semantic integration is shown in figure 2, where two small entity-relationship diagrams have been integrated into a larger diagram. In order to perform this integration, we would need to know that class **car** in the left-hand schema is a subclass of class **vehicle** in the right-hand schema, and that **road** and **rail** are both subclasses of class **transport medium**. Such additional information would be provided by an *integration schema* maintained in a metasystem.

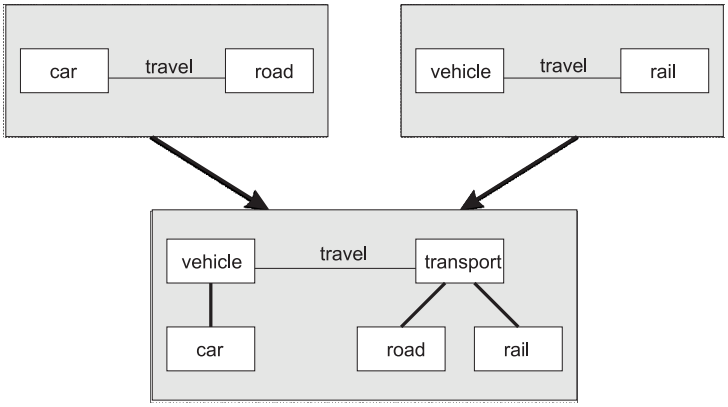

Figure 2: Integration of semantic data models specified by ER diagrams

There is a large body of work on mechanisms for integration of schemata of this kind and we will not pursue the subject further here. For an early review of schema integration methods, see Batini *et al.* 1986. Baader *et al.* (1997) contains several papers on formal approaches, include use of logics, for information integration.

In order to address the question of the degrees of imprecision introduced by semantic schemata, we consider a special category of semantic schemata, namely inheritance hierarchies related to taxonomies, and develop a formal theory of semantic imprecision in this case. Inheritance hierarchies (Smith and Smith 1977) are an important component of almost all contemporary semantic data modelling methodologies. Classes are arranged in a hierarchy, with classes lower in the hierarchy modelling finer details of the taxonomic classification of objects in the information space. The details modelled at each stage depend on the category of semantic model being used. In the case of the (enhanced) entity-relationship model, properties inherited by subclasses from superclasses might be attributes and relationships to other classes. In an object-oriented model (Worboys *et al.* 1990, Worboys 1994), subclasses might also inherit behaviour in the form of operations from superclasses. In the case of multiple inheritance, where a subclass can inherit from more than one superclass, some form of conflict resolution is required if inherited properties conflict.

An inheritance hierarchy as a first approximation may be modelled as a partially ordered set (poset) $< S, \leq >$, where S is a set of classes and for classes $s, s' \in S$, $s \leq s'$ expresses the fact that class s is a subclass of class s'. We can assume that the partially ordered set has a top element, representing the most general class, called here **feature**. The definitions above allow the hierarchy to have both single and multiple

inheritances. An example of such a hierarchy is shown in figure 3. It should be noted that the formal model of an inheritance hierarchy developed here is simplified in the following ways:

- Only inheritance of attributes is considered; inheritance of methods, as found in an object-oriented data model, is not accounted for here.
- No constraints are placed on multiple inheritance. For example, it would be natural to specify that an observed entity might simultaneously be of types **artificial feature** and **water-related feature**, but could not be of types **canal** and **dwelling** at the same time. These constraints are not expressible in the simplified model.

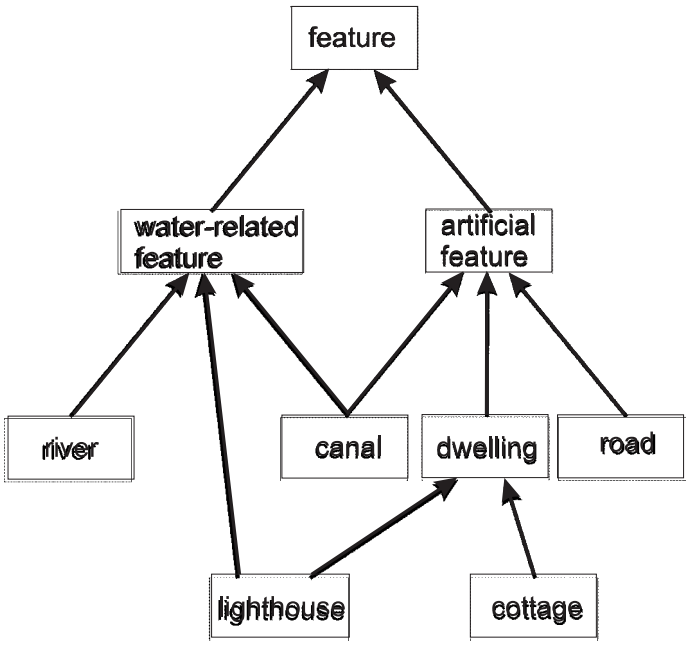

Figure 3: Semantic data model as a taxonomic inheritance hierarchy

Taxonomic imprecision is introduced when an observation of a phenomenon (which remember may itself be an observation) is made in a context whose semantic schema contains a taxonomy that is coarser than the taxonomy inherent in the phenomena itself. The notion of coarser and finer taxonomies may be formally expressed using the concept of an upper set, whose definition is now given.

An *upper subset U* of a partially ordered set *S* is a subset of *S* with the property that for all elements *u* of *U* and *s* of *S*, if $u \leq s$, then *s* is an element of *U*.

Figure 4 shows two upper subsets of the inheritance hierarchy in figure 3. There is clearly a sense in which these hierarchies are less precise than the original. The set of all upper subsets of a partially ordered set is closed with respect to set union and intersection. An upper subset of a hierarchy is itself a hierarchy (that is why we use upper subsets), and the space of all upper subsets of a hierarchy models all coarser subhierarchies of a given hierarchy.

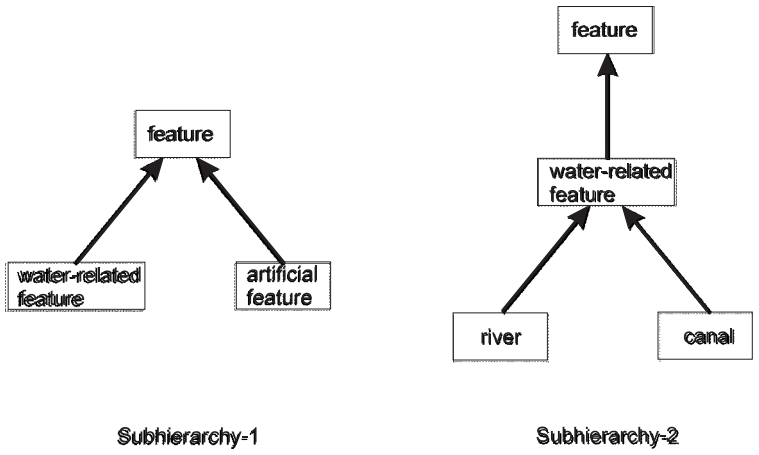

Figure 4: Two subhierarchies of the hierarchy in figure 3

The next step is to show that a coarsening of a hierarchy leads to a partitioning of the original classes forming the semantic entities of the phenomenon. To show this formally, let s be an element of partially ordered set S, and for upper subset U of S and element s of S, define $U(s) = \{u \in U \mid u \geq s\}$. We note that $U(s)$ is itself an upper subset of U. Now use U to define an indiscernibility relation on S as follows:

For all $s, s' \in S$, $s \sim_U s'$ if and only if $U(s) = U(s')$

This indiscernibility relation $\sim_U$ is an equivalence relation that induces a partition $P_U(S)$ of S, whose blocks, $[s]_U$, where $s \in S$, consist of subsets of elements all equivalent to each other (and to s). The partition is referred to as a *semantic frame of discernment*, and is an expression of the taxonomic imprecision introduced into the observation of the phenomenon by the subhierarchy U.

It can be observed that each block of $P_U(S)$ contains at most one member of U, for suppose that $u, u' \in U$, and u and u' are in the same block of $P_U(S)$. Then $U(u) = U(u')$, $u \geq u'$ and $u' \geq u$, so $u = u'$.

To continue with our example, the two subhierarchies shown in figure 4 induce partitions of the classes shown in figure 5. Subhierarchy-1, whose elements are shown in thick-edged boxes in the left-hand part of the figure, partitions the set of classes into the blocks: {**feature**}, {**water-related feature**, **river**}, {**canal**, **lighthouse**}, {**artificial feature**, **dwelling**, **road**, **cottage**}. Subhierarchy-2, whose elements are shown in thick-edged boxes in the right-hand part of the figure, partitions the same set of classes into the blocks: {**feature, artificial feature**, **dwelling**, **road**, **cottage**}, {**water-related feature**, **lighthouse**} {**river**}, {**canal**}. The next section discusses relationships between and operations on partitions in more detail, and the examples shown here will be used to illustrate the ideas for semantic schemata.

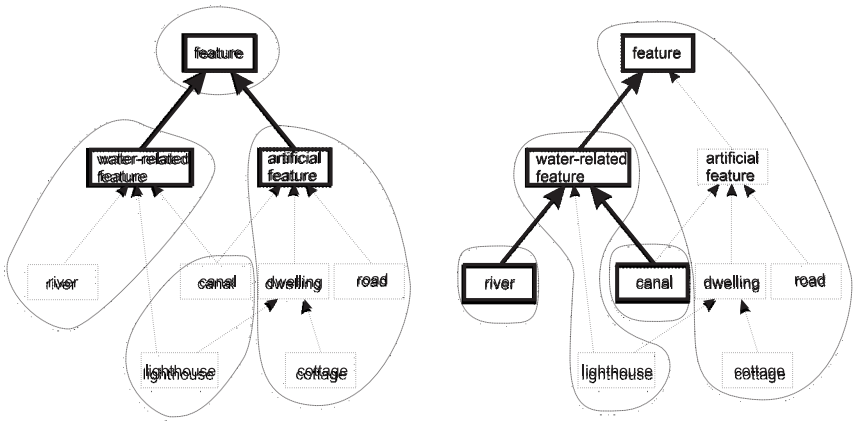

Figure 5: Partition frames induced by the two subhierarchies in figure 4

3.2 SPATIAL FRAMES OF DISCERNMENT

The spatial granularity associated with an observation relates to the level of spatial or geometric detail at which the phenomenon may be observed, and may be variable over the spatial extent. Spatial granularity used to be closely coupled with scale in the representation of geographic detail, but digital spatial data make this relationship rather more indirect (Goodchild and Proctor 1997). Any observation of data that are referenced to spatial locations, whether point, line or area, has associated with it a degree of geometric imprecision dependent upon the context in which it is made. This spatial context is determined by several factors:

- The nature of the geographic phenomenon (e.g., the context for a collection of administrative areas would be different from a context for a road network).

- The nature of the data collection mechanism (e.g. determined by a surveyor's measurement device or based on a framework of pixels of radiation received by a remote sensing device.
- The form of computational modelling paradigm (e.g. field-based or object-based).
- The form of computational representation and spatial data structure
- The medium in which results are presented and visualised.

The combination of one or more of these factors will result in an indiscernibility relation between elements of the source domain, as before induced by the schema of the context in which the observation is taking place. As with the earlier cases, the indiscernibility relation induces a *spatial frame of discernment* on the collection of spatial elements of the geographic phenomenon.

Imprecision is a property of an observation that leads to uncertainty, an epistemological property, in the observer. Clearly, imprecision and uncertainty are related notions. Couclelis (1996) provides a summary of the nature of uncertainty and imprecision in a spatial setting. There is a considerable body of related previous work on spatial imprecision. A computer-based representations of geographic phenomena are, by the nature of the paradigm, discrete. Examples of explicit discrete spatial data representations include the notion of a 'realm' (Güting and Schneider 1993), hierarchical spatial data structures (for example, Samet 1989), and hierarchical terrain models of De Floriani and her colleagues (1994).

Another approach to spatial uncertainty has been through fuzzy spatial objects. The fuzzy approach to geographic regional boundaries (and thus to the regions themselves) has been researched by several authors, for example, Davis and Keller (1997), Mark and Csillag (1989), Wang and Brent Hall (1996). The essence of the approach is that uncertainty of membership of a location in a region is indicated by a real number between 0 and 1, where a membership value of 0 indicates that the location is definitely not in the region, a value of 1 that the location definitely is in the region, and the magnitude of an intermediate value indicating a level of certainty that the location is in the region.

Related to the fuzzy formalism is the notion of *vagueness* in geographic phenomena. This is distinguished from the work using fuzzy approaches by the qualitative nature

of the reasoning methods. Recent examples of literature using this approach are Cohn and Gotts (1996), Erwig, and Schneider (1997), Fisher (1997) and Worboys (1998).

In a sense, spatial imprecision is easier to formalise than semantic precision, at least if we assume a simple model of the underlying space. In this paper, we assume that the spatial frame of discernment partitions the space in which the phenomenon occurs into a finite set of connected subregions. This type of spatial frame is very common in many types of observations of geospatial phenomena. For example, a remotely-sensed image will partition the extent of the observed space into a set of pixels. It is easy to see that in these cases, elements of the phenomenon within the same pixel are indiscernible from each other, indiscernibility is an equivalence relation, and the associated spatial frame of discernment is just the resulting partition. However, it is possible to imagine richer kinds of indiscernibility relation, for example resulting from topological generalization, and these would lead to wider classes of spatial frames.

4. FRAME SPACES

We have seen that semantic inheritance hierarchies and spatial schemata both lead to an indiscernibility relation that is an equivalence relation. In this case, the frame of discernment forms a partition P of S, called a *partition frame of discernment*, and as this is clearly an important subcase, we now examine it in further detail.

When we come to consider data integration and generalisation, it will be important to move between schemata that in general will have different associated frames of discernment. Assume that the data sets all arise from observations of the same geographic phenomenon, that is, the source set S is the same in each case. We now consider the space $\mathbf{P} = \mathbf{P}(S)$ of all partition frames of discernment over S.

The space $\mathbf{P}$ is a lattice, when a partial order $\leq$ on $\mathbf{P}$ is defined as follows. For P_1 and P_2 belonging to $\mathbf{P}$, $P_1 \leq P_2$ if and only if $\forall x \in P_1, \exists y \in P_2$ such that $x \subseteq y$ (Grätzer 1978). The idea behind this definition is that a partition frame is less than another in the partial ordering if it provides a finer granularity. For any two partition frames P_1 and P_2 of S, it is always possible to form their greatest lower bound $P_1 \wedge P_2$ and least upper bound $P_1 \vee P_2$ (see figure 6 for an example).

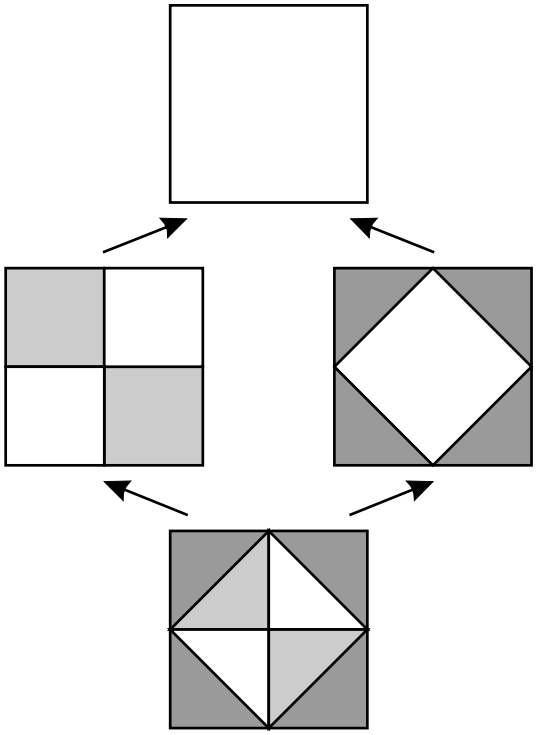

Figure 6: Meet and join of two partition frames

The greatest lower bound $\wedge P$ of any finite set $P = \{P_1, \ldots, P_n\}$ of partition frames is given by:

$$\wedge P = \{x_1 \cap \ldots \cap x_n \mid x_1 \in P_1, \ldots, x_n \in P_n \text{ and } x_1 \cap \ldots \cap x_n \neq \varnothing\}$$

The construction for the upper bound $\vee P$ of any finite set $P = \{P_1, \ldots, P_n\}$ of partitions is a little more complicated, and may be better described using the equivalence relations $\rho_1, \ldots, \rho_n$ that induce the partition frames $P_1, \ldots, P_n$, respectively. Construct a new equivalence relation ρ as with the following definition:

For all $u, v \in S$, $u \rho v$ if and only $\exists\ w_1, \ldots, w_m \in S$ such that

$u \alpha_0 w_1, w_1 \alpha_1 w_2, \ldots, w_m \alpha_m v$

where $\alpha_i \in \{\rho_1, \ldots, \rho_n\}$ for $0 \leq i \leq m$.

The partition frame $\vee P$ induced from ρ can be shown to be the least upper bound of the finite set $P = \{P_1, \ldots, P_n\}$ of partition frames (Grätzer 1978).

Thus, **P** is a lattice. Furthermore, **P** has top and bottom elements, the top element $\top$ being the partition frame consisting of the single block S, and the bottom element $\bot$ the partition whose blocks are the singleton sets $\{s\}$ for each $s \in S$.

Any subset of **P** with a top element the same as the top element of **P**, and that is closed under partition meets will be termed a *frame space*.

4.1 SPATIAL AND SEMANTIC FRAME SPACES

Spatial frames of discernment are discussed in a companion paper [34] and will not be considered further in this section. With regard to semantic frames, in the special case that the schema is based upon semantic hierarchical data model, semantic frames of discernment have been shown in section 3.1 to arise from subhierarchies of taxonomic inheritance hierarchies. Clearly, not every partition of the set of semantic classes arises from a subhierarchy of the set, so in general the subset of frames of discernment based upon subhierarchies of a given semantic hierarchy is a proper subset of the full set of partitions. The question arises whether this subset is a frame space. This is indeed the case, and to show this all we need is the following theorem.

Theorem

Let $< S, \leq >$ be a poset and U and V be upper subsets of S. Then, the following equality holds:

$$P_U(S) \wedge P_V(S) = P_{U \cup V}(S)$$

Proof

The following chain of equalities provides a brief indication why the equality holds.

$$\begin{aligned} P_U(S) \wedge P_V(S) &= \{[s]_U \cap [t]_V \mid s,t \in S, [s]_U \cap [t]_V \neq \varnothing\} \\ &= \{[s]_U \cap [s]_V \mid s \in S\} \\ &= \{[s]_{U \cup V} \mid s \in S\} \\ &= P_{U \cup V}(S). \end{aligned}$$

■

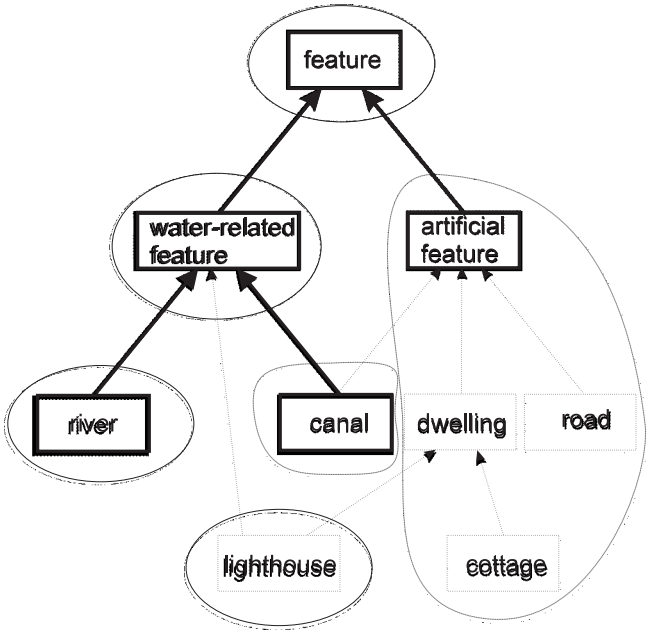

Figure 7: Partition induced by the join of the upper subsets in figure 5

The example in figure 7 illustrates this result using the subhierarchy-1 and subhierarchy-2 from figure 5. The upper set of nodes (call it U) of subhierarchy-1 is {**feature**, **water-related feature**, **artificial feature**}, and the upper set of nodes (call it V) of subhierarchy-2 is {**feature**, **water-related feature**, **river**, **canal**}. In this example, $U \cup V$ = {**feature**, **water-related feature**, **artificial feature**, **river**, **canal**}, as shown as the collection of thick-edged boxes in figure 7, and it can be checked that the induced partition frame of discernment consists of blocks {**feature**}, {**water-related feature**}, {**river**}, {**lighthouse**}, {**canal**}, {**artificial feature**, **dwelling**, **road**, **cottage**}. It can be checked further that this partition is the meet of the partition frames of U and V.

Suppose we are given two partition frames of discernment $P_U(S)$ and $P_V(S)$ over a given source set S, induced by subhierarchies represented by upper sets U and V. Then the above theorem shows that their meet is given by the partition $P_{U \cup V}(S)$. Therefore, $P_{U \cup V}(S)$ is itself a partition frame of discernment over S, induced by the subhierarchy represented by the upper set $U \cap V$. Thus, partition frames of discernment induced by subhierarchies of a fixed hierarchy are closed under partition meet, and so form a frame space.

We may note in passing that the corresponding equality concerned with partition joins does not hold, for in general $P_U(S) \vee P_V(S) \neq P_{U \cap V}(S)$. However, it is possible to show the weaker result that $P_U(S) \vee P_V(S) \geq P_{U \cap V}(S)$.

5. THEORY OF IMPRECISE OBSERVATIONS

We are at last in a position to formally define an observation with respect to an imprecise context specified by a schema. The given schema, whether semantic or spatial will consist of a collection of schema elements, partitioned by the indiscernibility (equivalence) relation. In the case of a spatial schema, the elements will be blocks in a partition of the underlying spatial framework for the observation; and in the case of the inheritance hierarchy that provides an example of a semantic schema, the elements will be blocks of a partition of the poset of semantic classes. As above, we will call partitions of the set of schema elements induced by an indiscernibility relation, frames of discernment.

An *observation* of a phenomenon with respect to a schema is an assignment of values from the truth set **T** = {**yes**, **maybe**, **no**} to the frame of discernment F. The meaning of this assignment is that for each element $f \in \mathbf{F}$, the observation does one of the following:

1. Observes the phenomenon as definitely instantiated at $f \in F$.
2. Observes the phenomenon as possibly instantiated at $f \in F$.
3. Observes the phenomenon as definitely not instantiated at $f \in F$.

In the case of spatial frames, the instantiation indicates that the phenomenon is definitely, maybe, or definitely not located at that element of the spatial framework. In the case of semantic frames, the instantiation indicates that the phenomenon is definitely, maybe, or definitely not an instance of the class or collection of classes in the inheritance hierarchy. In general the observation is not totally unconstrained. For example, if the spatial extent of a phenomenon is known to be a connected region, then the blocks of the spatial schema must be contiguous. We will return to this question for semantic schemata shortly.

Given an observation Ω, we may define the following two subsets of the frame of discernment F.

$L_\Omega(F) = \{f \in F \mid \Omega(f) = \textbf{yes}\}$

$U_\Omega(F) = \{f \in F \mid \Omega(f) = \textbf{yes} \text{ or } \Omega(f) = \textbf{maybe}\}$

$L_\Omega(F)$ is called the *lower approximation* for the observation Ω, and contains all those frame elements which are definitely involved with the phenomenon, while $U_\Omega(F)$ is called the *upper approximation* for the observation Ω, and contains all those frame elements which are possibly involved with the phenomenon. This approach is reminiscent of the theory of rough sets (Pawlak 1982, 1991, 1993), which is a formal approach to reasoning under conditions of imprecise information. The sets $L_\Omega(F)$ and $U_\Omega(F)$ may be thought of as providing respectively pessimistic and optimistic approximations to properties of the phenomenon under observation, given the imperfect granularity provided by the frame. In general $L_\Omega(F) \subseteq U_\Omega(F)$. It may be noted that $L_\Omega(F) = U_\Omega(F)$ if and only if the relevant properties of the phenomenon are observed with total precision (crisply).

An observation thus results in a pair $<L, U>$ of sets, where $L \subseteq U \subseteq F$, and called an *F-granular object* (*granular object).* We are thinking here of the frame F as a granulation of a finer set S (granularity being imposed by the imprecision of the observation). To the degree of precision admitted by the frame, $<L, U>$ may represent any subset X of F such that $\cup L \subseteq X \subseteq \cup U$. It will often be useful to think of $<L, U>$ as actually being the set of all possible such subsets of S. For an example of a vague spatial region, figure 8 shows on the left a vague granular object (definite area in darker grey and possible area in lighter grey) and on the right the boundaries of three of the regions that it might represent.

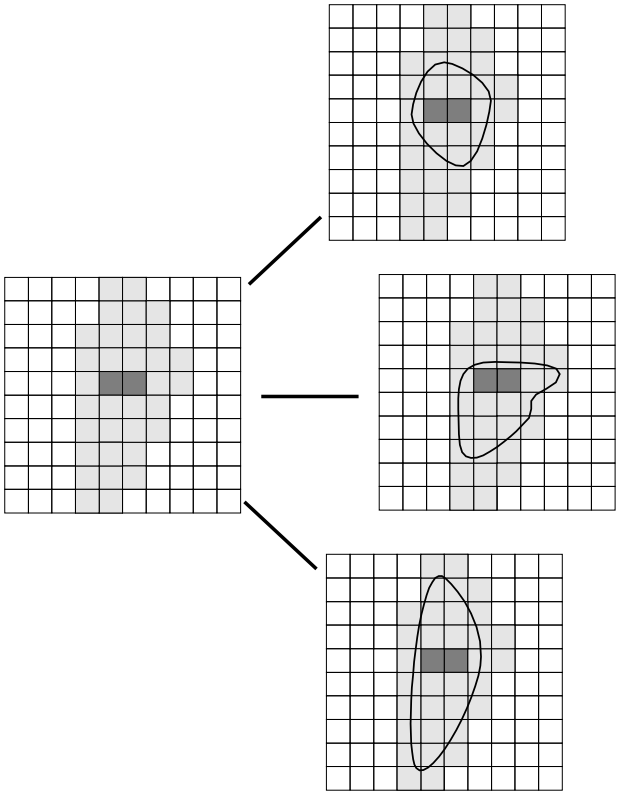

Figure 8: Example of a granular object and some of the regions it might represent

5.1 OBSERVATIONS WITH SEMANTIC FRAMES

This section considers observations with respect to semantic schemata. We have already noted that the assignment that defines an observation is not unconstrained. For example, an observation that notes a phenomenon as a possible instance of a class cannot at the same time note the phenomenon as definitely not an instance of a superclass (an observation cannot both note something as possibly a house but definitely not a building, because assuming the usual definitions of these classes, a house is a building).

Assume we are given a semantic schema provided by an inheritance hierarchy represented by the poset S. If we place an ordering relation on **T** as **no** < **maybe** < **yes**, then we can formally define a *semantic observation* as an order preserving function Ω from S to **T**. That is, $\Omega : S \rightarrow \mathbf{T}$ such that for all $s, s' \in S$, $s \leq s'$ implies that $\Omega(s) \leq \Omega(s')$.

In the case when the observation is imprecise with respect to the original schema, the observation is an assignment Ω_U of values of **T** to elements of an upper subset U of S.

This assignment can be propagated in a unique way to all nodes in S as follows. Define $\Omega_U(s) = \mathbf{inf}\ \{\Omega(u) \mid u \in U(s)\}$, where, for $X \subseteq \mathbf{T}$, $\mathbf{inf}(X)$ is the unique minimum element of X if X is non-empty, and $\mathbf{inf}(\varnothing) = \mathbf{yes}$. It is clear that elements in the same block of the partition $P_U(S)$ will have the same assignment, so Ω_U induces a function (overloaded with the same name) on the partition frame of discernment $P_U(S)$.

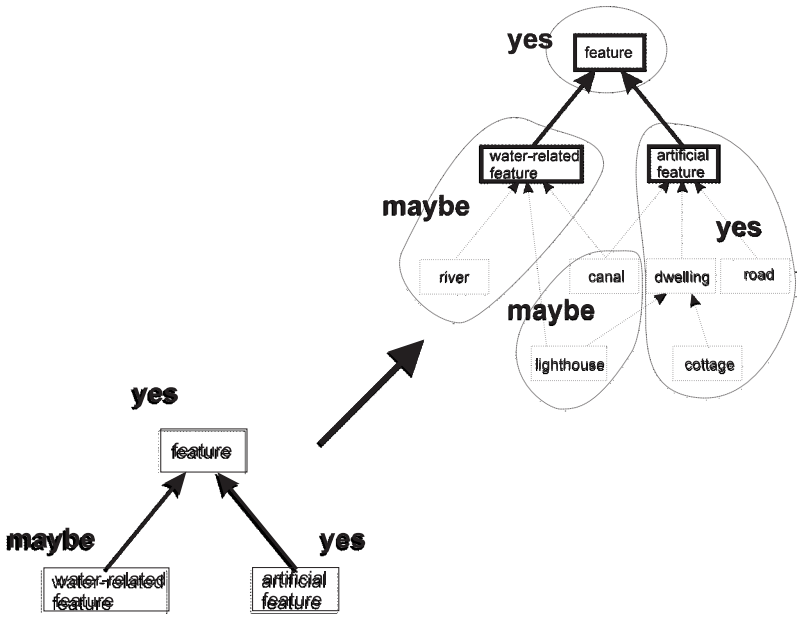

Figure 9: Observation based upon a subhierarchy from figure 4

Figure 9 shows an observation Ω_U, based upon the subhierarchy $U = \{$**feature**, **water-related feature**, **artificial feature**$\}$ of our usual example hierarchy (see figures 3 and 4). In this case:

Ω_U(**feature**) = **yes**

Ω_U(**water-related feature**) = **maybe**

Ω_U(**artificial feature**) = **yes**

We can notice that Ω_U is order preserving, and use the above construction to lift to function Ω_U on the frame, with our theory necessitating the following assignments:

$\Omega_U(\{$**feature**$\})$ = **yes**

$\Omega_U(\{$**water-related feature**, **river**$\})$ = **maybe**

$\Omega_U(\{$**artificial feature**, **dwelling**, **road**, **cottage**$\})$ = **yes**

$\Omega_U(\{$**canal**, **lighthouse**$\})$ = **maybe**

5.2 OBSERVATIONS WITH SPATIAL FRAMES

In this paper, we are assuming that the spatial frame of discernment partitions the space in which the phenomenon occurs into a finite set of connected subregions. The topology of the underlying space induces a topology on the spatial frame. In general, an observation with respect to a spatial frame is the assignment of any elements of set **T** to elements of the frame. However, extra information will constrain the assignment. For example, if the location of the geographic phenomena is known to be connected, then $L_\Omega(F)$ and $U_\Omega(F)$ are not completely unconstrained. Figure 10 shows several observations of an areal extent, and if further question arises as to whether the areal extent is connected, then we can reason that

- Observation A cannot be of a connected region
- Observation B can only be of a connected region if the right-hand 'maybe' area is definitely not part of the region.
- Observation C can be of a connected region, but the region is wholly contained in either the right-hand or left-hand 'maybe' area, but not both.
- Observation D can be of a connected region.

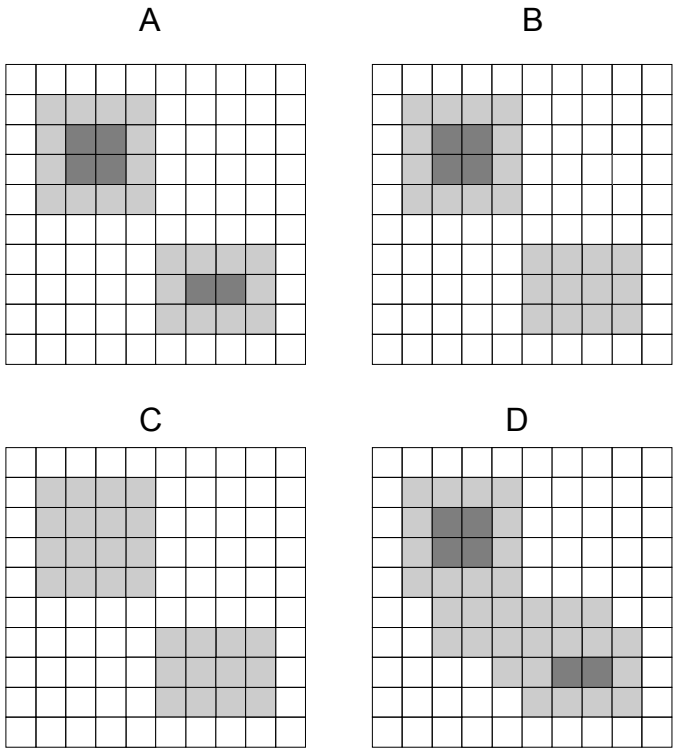

Figure 10: Spatial observations

Further work on the topology of imprecise spatial extents is required to formalise reasoning of this kind.

5.3 INTEGRATING IMPRECISE OBSERVATIONS

The work so far has provided a formal framework for the treatment of single observations of a geographic phenomenon. This section discusses how more than one observation may be amalgamated. A general question that will also be considered is how an observation made with respect to one schema may be represented with respect to another schema (see figure 11). In general, we use the term *transport* for this movement between schemata; in the case when the second schema is more imprecise than the first, then transport becomes generalization.

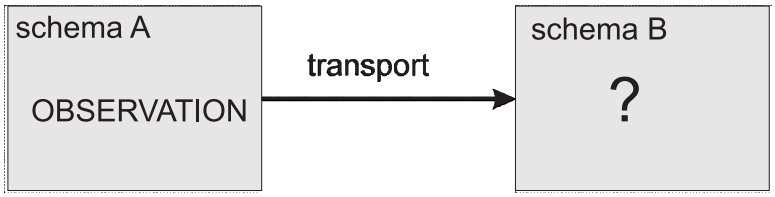

Figure 11: Transport between schemata

We begin with some more fundamental considerations. Assume an underlying fixed semantic or spatial partition frame space **P** relating to observations of a phenomenon with source set *S.* The first step is to provide a partial order on the frame space, that expresses the idea that one observation may be more precise than a second. In order to make comparisons between observations, we need some way of relating them across their corresponding frames. This is done by using the underlying common source set *S*, and an embedding function **emb** that assigns to each entity in the partition space the subset of the set *S* that it occupies.

Formally, let $P \in \mathbf{P}$ be a partition frame with source set *S*, and $p \in P$. Then define the embedding of p in *S* by $\mathbf{emb}(p) = \cup p$. Further, define the embedding of the *P*-granular object $O = \langle L, U \rangle$ as:

$$\mathbf{emb}(O) = \langle \mathbf{emb}(L), \mathbf{emb}(U) \rangle.$$

Different granular objects, with different granularities, may have the same embeddings when the entities that they represent occupy the same portion of the

source set S. This is expressed formally by defining relation ~ on the set of all observations in the frame space **P** as follows. Let $O_1 = <L_1, U_1>$ be a P_1-granular object, and $O_2 = <L_2, U_2>$ be a P_2-granular object, where P_1 and P_2 are partitions in **P**. Then $O_1 \sim O_2$ if and only if $\mathbf{emb}(O_1) = \mathbf{emb}(O_2)$. It is immediate that ~ is an equivalence relation. The set of equivalence classes is called *the embedded object space*. Now, define an ordering in the embedded object space. Let T_1 and T_2 be elements of the embedded object space. Let $T_1 = [\ O_1\ ]$ and $T_2 = [\ O_2\ ]$, where P_1--granular object $O_1 = <L_1, U_1>$ and P_2--granular object $O_2 = <L_2, U_2>$ are representatives of equivalence classes of resolution objects in the embedded object space.

Define $T_1 \leq T_2$ if and only if $\mathbf{emb}(L_1) \supseteq \mathbf{emb}(L_2)$ and $\mathbf{emb}(U_1) \subseteq \mathbf{emb}(U_2)$. It can be shown that $\leq$ is well-defined in the embedded object space, and in that space $\leq$ is a partial ordering. The intuition behind this ordering is that $T_1 \leq T_2$ expresses the fact that T_1 is a more precise observation than T_2.

We now address the transport question: if we know the representation of an object at one granularity, what is its representation at a second granularity? Let $O = <L, U>$ be a P-granular object and let a second partition P' be given. Then, define the P'-granular object $O' = <L', U'>$ to give the best possible representation of the same entity, in the following way. For $x \in P'$:

$x \in L'$ if and only if $\mathbf{emb}(x) \subseteq \mathbf{emb}(L)$

$x \in U'$ if and only if $\mathbf{emb}(x) \cap \mathbf{emb}(U) \neq \varnothing$

We note that $O \leq O'$, and this accords with intuition, as we would not expect to gain precision by transferring the observation of a geographic phenomenon from one frame of discernment to another, without additional information being provided. Figure 12 shows an example, where observation O_1 is represented as observation O_2 in a different frame.

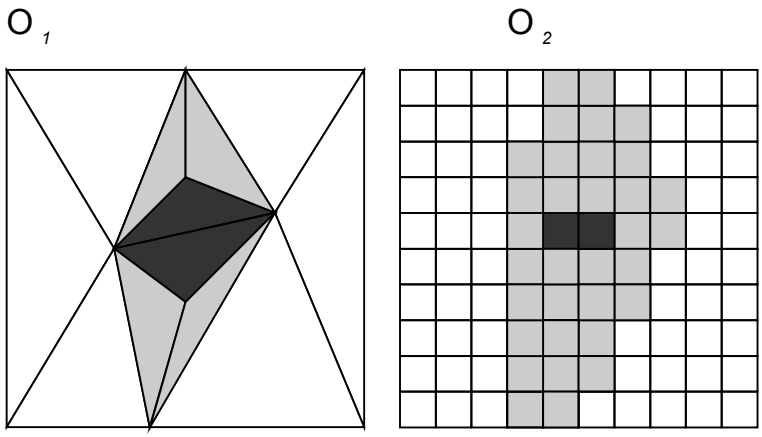

Figure 12: Transport of an observation to a different frame

In order that a set of observations of an object in different frames may be amalgamated, they may need to be compatible. For example, if the observations all purport to be spatial observations of the same underlying spatial region, then it cannot be the case that in one frame a particular location is definitely part of the region while in another frame the same location is definitely not part of the region. The notion of observation compatibility is discussed in (Worboys 1998) for the case of spatial frames. The most important point is that several compatible observations of the same phenomenon may be amalgamated so as to improve precision. Figure 13 shows an example of this for an amalgamation of two compatible spatial observations (shown on the left) of a region into a region shown on the right. It can be noted that the amalgamated region is a more precise observation than either of its constituents, having a larger definite area and smaller possible area. Further formal details may be found in (Worboys 1998).

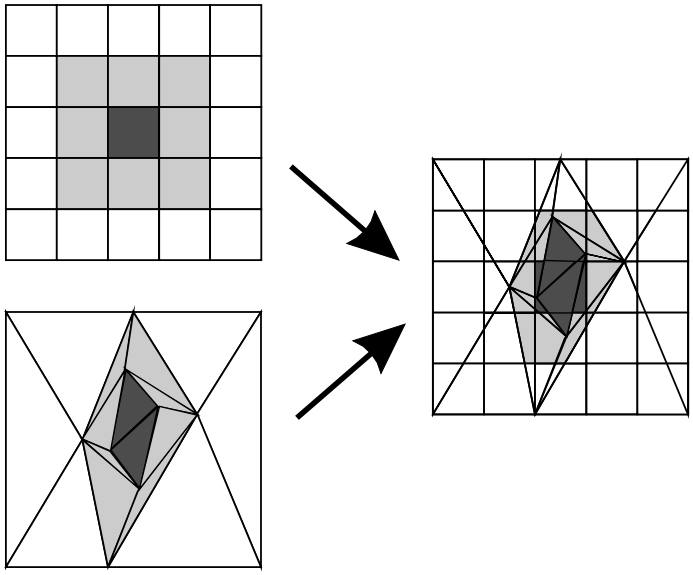

Figure 13: Amalgamation of two compatible spatial observations

6. CONCLUSIONS

The phenomena that this paper has discussed have their extents in geographic space. There has been discussion of the nature of geographic space (Egenhofer and Mark 1995, Montello 1992), and it can at least be agreed that its components include geometric (metric and topological space) and temporal dimensions, as well as domains of attributes. This paper has provided a formal framework in which we can reason about imprecision introduced by observations of the spatial and attribute components of geographic phenomena. In the context of geographic phenomena, the paper has formalized the notions of imprecision, indiscernibility, frame of discernment and imprecise observation. We have also provided the beginnings of a treatment in which observations of the same or different phenomena may be integrated, and the imprecision of the integrated product discussed. Further work is needed on:

- Incorporation of temporal contexts.
- Elaboration of semantic schemata to move beyond simple taxonomic hierarchies.
- Elaboration of spatial schemata to include examples of topological generalization, and a theory of imprecise topological reasoning.
- A fuller theory of observation amalgamation.

- Integration of semantic, spatial and temporal schemata into a seamless theory of imprecision of geographic information.
- Resulting theory of reasoning with vague geographic information.

ACKNOWLEDGEMENTS

The author thanks John Stell and Peter Fletcher for very helpful discussions and ideas relating to this paper.

REFERENCES

1. F. Baader, M.A. Jeusfeld, W. Nutt. Intelligent access to heterogeneous information sources. Report on the Fourth Workshop on 'Knowledge Representation Meets Databases', *Association for Computing Machinery SIGMOD Record*, **26**(4):44-48, 1997. (Full Proceedings obtainable at `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-8/`).

2. C. Batini, M.Lenzerini and S.B. Navathe. A comparative analyis of methodologies for database schema integration. *Association for Computing Machinery Computing Surveys*, **18**(4):323-364, 1986.

3. P.P.-S. Chen. The entity-relationship model - toward a unified view of data. *Association of Computing Machinery Transactions on Database Systems,* **1**(1):9-36, 1976.

4. A.G. Cohn and N.M. Gotts. The 'egg-yolk' representation of regions with indeterminate boundaries. In P. Burrough, and A. Frank, (eds.) *Geographic Objects with Indeterminate Boundaries*, Taylor and Francis, London, pp. 171-187, 1996.

5. H. Couclelis. People manipulate objects but cultivate fields: beyond the raster-vector debate in GIS. In Frank, A.U., Campari, I. and Formentini, U. (eds.), *Theories of Spatio-Temporal Reasoning in Geographic Space, Lecture Notes in Computer Science 639.* Springer-Verlag, Berlin, Germany, pp. 65-77, 1992.

6. H. Couclelis. Towards an operation typology of geographic entities with ill-defined boundaries. In P. Burrough and A. Frank, (eds.) *Geographic Objects with Indeterminate Boundaries*, Taylor and Francis, London, pp. 45-55, 1996.

7. T.J. Davis, and C.P. Keller. Modelling uncertainty in natural resource analysis using fuzzy sets and Monte Carlo simulation: slope stability prediction. *Int. Jour. of GIS*, 11(5): 409-434, 1997.

8. G. Dettori and E. Puppo, How generalization interacts with the topological and metric structure of maps, Proceedings of 7th International Symposium on Spatial Data Handling, Delft, Netherlands, Taylor and Francis, pp 9A.27-9A.38, 1996.

9. M.J. Egenhofer,and D.M. Mark. Naive geography. In Frank, A.U. and Kuhn, W. (eds.), *Spatial Information Theory, Proceedings of COSIT'95, Lecture Notes in Computer Science 988.* Springer-Verlag, Berlin, Germany, pp. 1-15, 1995.

10. M. Erwig, and M. Schneider. Vague regions. In Proceedings of the 5th Int. Symp. in Spatial Databases (SSD'97), Lecture Notes in Computer Science 1262, Springer, Berlin, pp. 298-320, 1997.

11. P. Fisher. The pixel: a snare and a delusion. *Int. Jour. Remote Sensing* **18**(3): 679-685, 1997.

12. L. de Floriani, P. Marzano, and E. Puppo. Hierarchical terrain models: survey and formalization. In *Proceedings SAC'94, Phoenix, AR, USA,* pp. 323-327, 1994.

13. M.F. Goodchild. Data models and data quality: problems and prospects. In M.F. Goodchild, B.O. Parks, and L.T. Steyaert, (eds.), *Visualization in Geographical Information Systems.* John Wiley, New York, pp. 141-149, 1993.

14. M.F. Goodchild and J. Proctor. Scale in a digital geographic world. *Geographical and Environmental Modelling* **1**(1): 5-23, 1997.

15. G. Grätzer, *General Lattice Theory*. Academic Press, New York, 1978.

16. R.H. Güting, and M. Schneider. Realms: A foundation for spatial data types in database systems. In D. Abel, B.C. Ooi, (eds), *Advances in Spatial Databases, Proceedings of SSD'93, Singapore, Lecture Notes in Computer Science 692.* Springer-Verlag, Berlin, Germany, pp. 14-35, 1993.

17. D.M. Mark, and F. Csillag. The nature of boundaries on 'area-class' maps. *Cartographica* **26**: 65-77, 1989.

18. D.R. Montello. The geometry of environmental knowledge. In Frank, A.U., Campari, I. and Formentini, U. (eds.), *Theory and Methods of Spatio-Temporal Reasoning in Geographic Space, Lecture Notes in Computer Science 639.* Springer-Verlag, Berlin, Germany, pp. 136-152, 1992.

19. J.C. Müller, R. Weibel, J.P. Lagrange, F. Salgé, Generalization - state of the art and issues, in *GIS and Generalization: Methodology and Practice* Müller, J.C., Lagrange, J.P. and Weibel, R. (eds.), Taylor and Francis, pp. 3-17, 1995.

20. Z. Pawlak. Rough sets. *Int. Journal of Inf. and Comp. Sci*., **11**(5): 341-356, 1982.

21. Z. Pawlak. *Rough Sets - Theoretical Aspects of Resoning about Data*, Kluwer, 1991.

22. Z. Pawlak. Hard and soft sets. In Alagar, V.S., Bergler, S. and Dong, F.Q. (eds.), *Proceedings of RSSC'94, The Third International Conference on Rough Sets and Soft Computing*, San Jose State University, San Jose, CA, USA, 1993.

23. J. Peckham and F. Maryanski. Semantic data models. *Association for Computing Machinery Computing Surveys*, **20**:153-189, 1988.

24. E. Puppo and G. Dettori. Towards a formal model for multiresolution spatial maps. In *Proceedings of SSD'95*, Lecture Notes in Computer Science, **951**, Berlin: Springer, pp. 152-169, 1995.

25. P. Rigaux and M. Scholl. Multi-scale partitions: Applications to spatial and statistical databases. In *Proceedings of SSD'95*, Lecture Notes in Computer Science, **951**, Berlin: Springer, pp. 170-183, 1995.

26. H. Samet. *The Design and Analysis of Spatial Data Structures.* Addison-Wesley, Reading, MA, USA, 1989.

27. G. Shafer. *Mathematical Theory of Evidence*. Princeton University Press, 1976.

28. A.P. Sheth and J. Larsen. Federated database systems for managing distributed, heterogeneous and autonomous databases. *Association for Computing Machinery Computing Surveys*, **22**:183-236, 1990.

29. R. Slowinski and D. Vanderpooten. Similarity relation as a basis for rough approximations. Research Report 53-95, Institute of Computer Science, Warsaw Institute of Technology, `ftp://ftp.ii.pw.edu.pl/pub/Reports`, 1995.

30. R. Slowinski and D. Vanderpooten. A generalized definition of rough approximations. Research Report 5-96, Institute of Computer Science, Warsaw Institute of Technology, `ftp://ftp.ii.pw.edu.pl/pub/Reports`, 1996.

31. J.M. Smith and D.C.P. Smith. Database abstractions: aggregation and generalization. *Association for Computing Machinery Transactions on Database Systems*, **2**(2):105-133, 1977.

32. F. Wang, and G. Brent Hall. Fuzzy representation of geographical boundaries in GIS. *International Journal of GIS* 10(5): 573-590, 1996.

33. M.F. Worboys. Object-oriented approaches to geo-referenced information. *International Journal of GIS* **8**(4): 385-99, 1994.

34. M.F. Worboys. Imprecision in finite resolution spatial data. Accepted for publication in *GeoInformatica*, Kluwer, 2(3), 1998.

35. M.F. Worboys, H.M. Hearnshaw, and D.J. Maguire. Object-oriented data modelling for spatial databases. *International Journal of Geographical Information Systems* **4**: 369-83, 1990.