REPRESENTING UNCERTAINTY IN SPATIAL DATABASES

Erlend Tøssebro
Dep. of Electrical Engineering and Computer Science
University of Stavanger
NO-4036 Stavanger, Norway
E-mail: erlend.tossebro@uis.no

Mads Nygård
Department of Computer and Information Science
The Norwegian University of Science and Technology
NO-7491 Trondheim, Norway
E-mail: mads@idi.ntnu.no

ABSTRACT

Due to lack of accurate measurements, or rapid changes in time, spatial data are often uncertain. This paper presents a new abstract model for uncertain spatial information. The model is based on the principle that one knows that the uncertain object, regardless of type, must be within a certain area. The model also incorporates probability functions so that it is possible to determine the probabilities that various operations are true. This paper contains mathematical definitions of uncertain points, lines and regions. The paper also contains definitions of some relevant operations on these types. These operations are also evaluated for their usefulness with regard to uncertain data. A corresponding discrete model is already published.

1. INTRODUCTION

Databases which store information about geographic objects are becoming increasingly common in modern society as high-performance computer systems become available and positioning systems become more common and more accurate. However, many forms of spatial data cannot be measured exactly, or they may vary with time in such a manner that one cannot know exactly where the spatial object is at any given time. Examples of these two are given below:

Example 1: A lake is used as a reservoir for a hydroelectric power plant. Because of differences in energy demand and precipitation in the area, the water level, and thus the extent of the lake may vary considerably. Although one could store the exact size of the lake at any time by taking measurements frequently enough, this would be costly both in terms of manpower (taking the measurements) and space. A better solution might be to store the lake in a manner that indicates that it is uncertain. This uncertainty includes both position and the exact shape of the object.

Models for static uncertain regions exist already, and are well documented. See Section 2 for examples. However, other types of spatial data may also be uncertain. The following examples illustrate this for points and lines:

Example 2: If one is tracking a submarine, the sonars may give only an approximate position of the submarine, especially if the submarine is close to the sea floor and irregularities in the sea floor give off false readings.

Example 3: Simulations of the behaviour of an oil reservoir as well as other simulations relating to geological or geographical data may well yield results with some uncertainty. The model presented in this paper can be used to store such uncertain results.

Example 4: There are three different types of lines in geological databases assuming a two-dimensional view, and all of them may be uncertain because they are underground and therefore difficult to measure. The first type is a contact between two different rock types. This is really the boundary of two regions. The second is a fault line, either active or inactive. Because inactive fault lines are not necessarily tied to continental plates or to differences in rock type (there may be the same type on both sides), this is a true uncertain line.

The following example illustrate the need for a system that handles uncertainty in all the spatial data types.

Example 5: Imagine that you have scientists who are driving around making measurements in the Sahara desert to determine the extent of underground water reservoirs. The scientists themselves are uncertain points due to the imprecision of the positioning system that they use. The roads are uncertain lines because the roads in the Sahara desert are more like routes that shift as the sand dunes move than paved roads. The water reservoirs that the scientists are studying are uncertain regions because they are located deep underground and it is therefore not feasible to do more than a few measurements at each site. The scientists therefore lack the necessary information to define them precisely. Such a database would be useful for the scientists mentioned. If they could query such a database while on site using a wireless device, they could coordinate their efforts better.

This example shows that one may need to store uncertain data of all the three types in the same database. Most existing systems handles only one type or two types.

In (Duckham et al. 2001), an ontology of different kinds of uncertainty is defined. The hierarchy of forms of uncertainty, or imperfection, is shown in Figure 1.

Proceedings of the 2008 High Performance
Computing & Simulation Conference ©ECMS
Waleed W. Smari (Ed.)
ISBN: 978-0-9553018-7-2 / ISBN: 978-0-9553018-6-5 (CD)

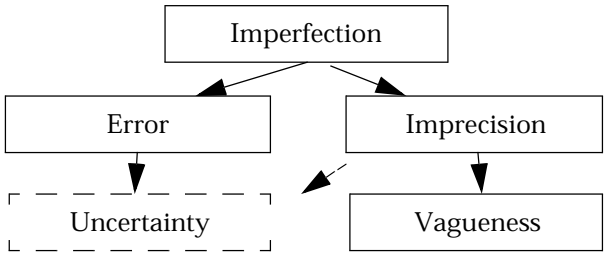

Figure 1: Hierarchy of the Types of Imperfection

Imperfection is considered to be the general form of uncertainty. Error is when measurements do not reflect reality. Imprecision is when measurements are lacking in specificity or are incomplete. (Duckham et al. 2001) considers vagueness to be a subcategory of imprecision. The basic goal of this paper is representing uncertainty in the position or extent of an object, regardless of the source of that uncertainty. In the rest of this paper, uncertainty therefore means either measurement error or imprecision due to incomplete knowledge, but does not cover vagueness. This definition of uncertainty is shown by the dashed box in Figure 1.

This paper will attempt to create a set of data types and operations for uncertain data, building on earlier work in spatial databases and models for vague and uncertain data.

In the rest of this paper, the word "crisp" will be used as the opposite of indeterminate (uncertain or vague).

2. RELATED WORK

There are several different types of models for spatial data. For spatiotemporal data, (Erwig et al. 1999) describes two modelling levels, abstract and discrete. Discrete models for spatiotemporal data can be directly implemented and are based on discrete representations such as vector or raster models. Abstract models are higher-level and usually model spatiotemporal data with point sets. In many abstract models, such as the one described in (Güting et al. 2000), lines and regions are modelled as infinite point sets in the Euclidean plane. This makes the model simpler, and may provide ideas for query operation designs that might be missed if one immediately went to the discrete level.

Abstract models also usually contain some rules to ensure that it is possible to store the data, although discrete models contain a lot more such rules. This paper contains an abstract model. Two distinct discrete models developed by the present authors have already been published in (Tøssebro and Nygård 2002b) and (Tøssebro and Nygård 2003). There is also a partial implementation of the model from (Tøssebro and Nygård 2003).

In (Tøssebro and Nygård 2002a), we outlined our current work on uncertainty in spatial and spatiotemporal database. A spatiotemporal extension to an abstract model like the one presented here is presented in (Tøssebro and Nygård 2002c).

One early model for uncertain points and lines is presented in (Dutton 1992). In this model, a point is represented as a central point with a circular deviation and a Gaussian distribution function over this area. A line is represented as a series of such points. The line segments between the points are represented by the union of the straight line segments going between all possible positions of the two points. The paper then shows that such a line will have the greatest variance in the points themselves, and the least variance is in the centre of the lines between the points. This is contrary to what one might expect. The uncertainty is usually smallest in the sample points and greater between them.

(Mark and Csillag 1989) describe a way to model uncertainty in the location of the boundary of a region that uses probabilistic error bands. This means that on each side of the estimated border there is an area with a certain width in which the border can be. Additionally, the probability that a point p is inside the area is a function of the distance from the estimated border to p.

The egg-yolk model described in (Gohn and Gotts 1996) and (Clementini and Di Felice 1996) models an uncertain region with only one face as two regions, one inside the other. The inner region is referred to as the 'yolk' and the outer region as the 'white' in the egg. This representation is then used to find a lot of different topological relations between uncertain regions, each consisting of only one component. A model for broad lines is presented in (Clementini 2005). A broad line in this model is a region that could result from the continuous deformation of a line as well as two broad points that represent its end points. A broad point is represented by the area that it might be in.

Models based on fuzzy sets have been frequently used to model vague regions. Fuzzy sets (Zadeh 1965) are sets in which the membership of any individual point in the set is not either yes or no, but rather a number between 0 and 1. Many of these models, such as the discrete models presented in (Lagacherie et al. 1996) and (Lowell 1994), use rasters to represent the fuzzy sets. The models described in (Schneider 1999) and (Erwig and Schneider 1997) represent another type of fuzzy set model, because these, like the abstract models for crisp objects, use infinite point sets. The most comprehensive model for vague data using fuzzy sets is the one presented in (Schneider 1999), which models all the standard spatial types (points, lines and regions) using fuzzy sets. (Schneider 1999) models a vague region as a fuzzy set where places which are certainly members of the region have values

of 1 and regions which are only partially members have values between 0 and 1. A vague line is a line with a crisp position but uncertain membership. In Schneider's model, this uncertain membership is indicated by a function which gives values between 0 and 1 for each point in the crisp line.

Although these fuzzy models cannot be used as they are to model positional uncertainty, some of the ideas from them may be adopted.

(Wang and Hall 1996) describe a model for fuzzy boundaries between regions in which the fuzzy membership function indicates how sharp the boundary is. A membership of 1.0 indicates a crisp boundary.

(Cheng et al. 1997) describe several ways to extract fuzzy objects from observations. These methods use a combination of fuzzy sets and probability theory. The model that they use is a raster model because fuzzy membership values are stored in each cell. However, they also group the cells into objects according to different criteria.

Another possible model for uncertain regions, regardless of the type of uncertainty, is the vector-based discrete model presented in (Schneider 1996). This model bases itself on two boundaries, like the egg-yolk models.

(Worboys 1998) uses rough sets to define the outer and inner boundaries of possibly imprecise spatial objects. (Worboys 1998) defines resolution objects that are partitions on the underlying space, and shows how to convert objects from one resolution to another. This process may introduce imprecision even if the original representation was precise, because the object may only partially overlap one of the new partition parts. Another approach using rough sets is presented in (Beaubouef and Petry 2001). This paper essentially shows that rough sets can be used to create a more general version of the egg-yolk approach.

There has been an effort to create a comprehensive type system for different kinds of spatial databases. (Güting et al. 2000) describes such a type system for spatiotemporal databases. (Schneider 1999) describes a similar kind of model for vague spatial data.

3. BASIS FOR THE NEW MODEL

The new model presented in this paper takes ideas from several of the models described earlier. The model in (Dutton 1992) is adequate for modelling digitization error, but not adequate for some other applications. One example of this is Example 1 from the introduction. This example cannot be modelled by the one in (Dutton 1992). because the region may have an arbitrary shape. However, the concept that the point is known to be located within a region and has a certain probability distribution can be used in the new model. Another example is that the approach suggested in (Dutton 1992) cannot model uncertainty about the length of a line. As with points, the concept of a line with a probability distribution function is useful for our work.

(Schneider 1999) describes an abstract model for vague spatial data. The region model in that paper may be used as a basis for a model for uncertain regions. An uncertain region may be modelled as a probability function where points which are certainly members have a value of 1 and points for which membership is uncertain have values between 0 and 1. Schneider's model for a vague lines or vague points, however, is not so useful for uncertain data. An uncertain line typically has uncertainty about exactly where it is, which means that a different type of model must be used. However, an uncertain line may also have uncertainty about whether it exists or not. This existence uncertainty may be modelled in the same way as vagueness. The difference between vague and uncertain points is the same as for lines.

An important difference between our new model and Schneider's is that his model uses somewhat different mathematics. While Schneider uses fuzzy sets, our new model uses probability theory. This is both because uncertainty is best modelled by probabilities, and because the probabilities for uncertain points and lines must be modelled by probability density functions. The authors do not know of a similar concept in fuzzy set theory.

Some of the types from (Güting et al. 2000) are used as building blocks for the types presented here. Therefore, a brief description of these types is given now. The basic type for points is A_{points}, which is a finite set of points. The type for a single point is A_{point}. A line in (Güting et al. 2000) (of type A_{line}) is defined as a set of curves forming a graph. A curve is defined by a function from a variable t, which is between 0 and 1, to the X-Y plane. Curves cannot intersect with themselves. The carrier set of this type is called A_{curve}. A region (Güting et al. 2000) is an infinite set of points in the plane with the condition that there may be no singleton points or lines. That is, the region must be a valid result of a regularized set operation. A region may consist of a finite set of disjoint components, or faces. These again can have a finite number of holes. The carrier set of faces is called A_{face}, while the carrier set of regions is A_{region}.

4. DATA TYPES FOR UNCERTAIN SPATIAL INFORMATION

This section describes a set of data types for modelling uncertain spatial information. The first subsection will describe how to model the basic datatypes such as numbers. The other subsections will describe uncertain

Table 1: Carrier Sets for the Data Types from (Güting et al. 2000)

Individual	Set
A_{point}	A_{points}
A_{curve}	A_{line}
A_{face}	A_{region}

points, lines and regions. All the types will be defined by their carrier sets.

To define the data types that follow, the operation support is needed. This operation comes from fuzzy set theory, but has a slightly wider application here. In this paper, support is defined as follows for any function $f: z \rightarrow \Re$:

$$Support(f) \equiv \{z | f(z) > 0\}$$

The z in this formula is a member of whatever type or set of types the function f accepts as input values. This means that support is defined for all uncertain types, whether they are spatial or not. A more complete definition and discussion of this operation can be found in Section 5.3.

All the uncertain data types defined in this paper rely on probabilities or probability density functions. The properties of these are defined by the following functions:

- Probability Density:
  $ProbDens(P) \equiv (\forall x: P(x) \geq 0) \wedge \int_x P(x) \leq 1$
- Spatial Probability Density:
  $SProbDens(P) \equiv (\forall x \forall y: P(x, y) \geq 0)$
  $\wedge \int_x \int_y P(x, y) \leq 1$
- Probability Function:
  $ProbFunc(P) \equiv \forall x: (P(x) \geq 0 \wedge P(x) \leq 1)$
- Spatial Probability Function:
  $SProbFunc(P) \equiv \forall x \forall y: (P(x, y) \geq 0 \wedge P(x, y) \leq 1)$

4.1 Base Types

An uncertain number can easily be modelled by a probability distribution function. For a real number, this function would have to be defined as a probability density function, whereas for integers, it might be just a collection of probabilities for the number having particular values.

Definition 1: An *uncertain number* is defined as follows.

$$A_{UNumber} \equiv \{NP(x) | ProbDens(NP) \wedge ((PieceCont(NP) \wedge Support(NP) \in A_{Range(number)}) \vee (DiracDelta(NP) \wedge Support(NP) \in A_{number}))\}$$

PieceCont(F) is true if the function F is piecewise continuous. DiracDelta(F) is true if F is a dirac delta function.

Many queries in spatial databases return Boolean values for data without uncertainty. Because a single Boolean value cannot indicate uncertainty, different ways of answering these queries must be found. The most appropriate way to answer such queries for uncertain data is to give the probability that the answer is "True". However, there are some cases in which this probability is difficult to determine. An example of this is the "Cross" operation from Section 5.2. In such cases a third "Boolean" value, Maybe, is used to indicate uncertainty. This last approach was used in (Erwig and Schneider 1997).

These two forms of Boolean values are treated as two different types in this paper. The uncertain Boolean is the version with three values, and the other is called a probability. A third type which is useful for uncertain data is a type which indicates to which degree a statement is true. Some operations may return a degree of truthfulness which cannot be interpreted as a probability:

$$A_{UBool} \equiv \{False, Maybe, True\}$$

$$A_{Prob} \equiv [0, 1]$$

$$A_{Degree} \equiv [0, 1]$$

4.2 Uncertain Points

An uncertain point is a point for which an exact position is not known. However, one usually knows that the point is within a certain area. One may also know in which parts of this area the point is most likely to be. For instance in Example 2, one knows that the submarine is somewhere within the sonar reflection (a region) and by looking at the varying intensities of the reflection one might have an idea of where the submarine is most likely to be. An uncertain point is therefore defined as a probability density function P(x, y) on the plane. The support of this function is the area in which the point may be. To be able to store the function P(x, y) in a computer, it must be piecewise continuous. The probability that the uncertain point exists at all is the double integral of P(x, y) over the plane. To be able to model crisp points, P(x, y) must be allowed to be a dirac delta function.

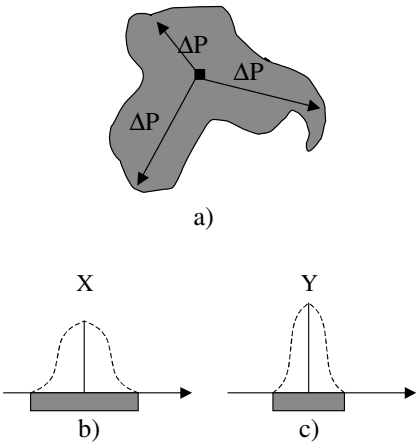

Figure 2: Uncertain Point

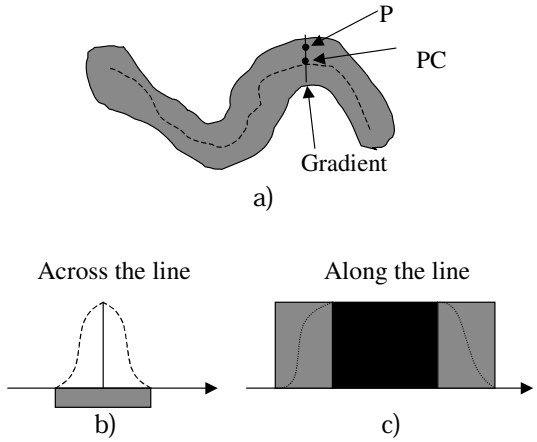

Figure 3: Uncertain Curve

Definition 2: An *uncertain point* is defined as follows.

$$A_{UPoint} \equiv \{PP(x, y) | SProbDens(PP) \wedge ((Support(PP) \in A_{region} \wedge PieceCont(PP)) \vee (Support(PP) \in A_{point} \wedge DiracDelta(PP)))\}$$

A possible uncertain point is shown in Figure 2a. Figures 2b and 2c show views of the X and Y directions. The central spikes indicate the expected value of the points. The thick bar underneath indicates the area of uncertainty.

The model described here can model Example 2 because it allows the point to be inside an arbitrarily shaped region, and not just a circle like (Dutton 1992). It also enables a point to be modelled where its existence is not certain.

One problem with this model is how to determine the probability density function so that the double integral of it over the universe becomes 1 if the point is certain to exist.

Definition 3: The *uncertain points* set type is defined as follows.

$$A_{UPoints} \equiv \left\{ UP \subseteq A_{UPoint} \middle| Finite(UP) \right\}$$

The Finite function returns True if the set contains a finite number of elements and False otherwise.

4.3 Uncertain Lines

The line type as defined in (Güting et al. 2000) is a set of curves where each member is a simple curve. The first step in developing a model for an uncertain line is therefore to create a model for an uncertain curve. An uncertain curve is a curve for which the exact shape, position or length is not known, but it is known in which area the curve must be. An example of an uncertain curve is shown in Figure 3a. It may also be known where in this area the curve is most likely to be. The dashed line in Figure 3a exemplifies this.

When seen along a line crossing it, a crisp curve would look like a point, or a set of points in the case of multiple crossings. When seen along the same line, an uncertain curve should be a probability density function indicating where the curve is most likely to cross. Such a function is shown in Figure 3b. This function may apply along the line marked “Gradient” in Figure 3a. Formally this line and its probability density function may be defined as follows:

$$A_{gradient} \equiv \{(gc, fg) | gc \in A_{curve} \wedge \forall(p \in gc):(fg{:}p \rightarrow \Re) \wedge ProbDens(fg)\}$$

When seen along its length, the uncertain curve has a probability of existing at each point. In Figure 3c, one common example of such a probability function is shown. In this example, there is uncertainty about the length of the line. This means that the line is certain to exist in the middle, and the probability of the line existing becomes lower the closer one comes to the ends.

One way of modelling this probability is that the uncertain line has a central line with a probability function associated with it. This probability function should not have areas in the middle where it is 0, because a curve with such a function is really two curves and not one, and should therefore be modelled as two curves. Such an illegal function is shown in Figure 4.

If there is uncertainty as to the number of curves, this may be modelled by a function which is less than 1 in a period between two places in which it is 1. This is shown in Figure 5.

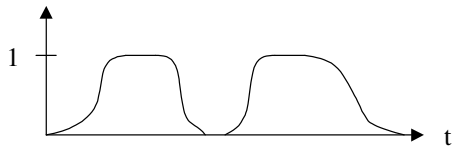

Figure 4: Illegal Probability Function for Uncertain Curve

This property of a function may be expressed mathematically as follows:

$$NoDip(f) \equiv (\forall x \forall y \forall z:((f(x) > 0) \wedge (f(z) > 0) \wedge x < y < z) \rightarrow f(y) > 0)$$

An uncertain line may be defined as a central line and a set of gradient lines. This set of gradient lines models the uncertainty in position and shape of the line. For each point of the central line there should be one and only one gradient line crossing it. The expected values of the probability functions of all the gradients should be somewhere on the central line. This is ensured by the following three conditions:

- The gradients do not share points or parts:

$$NoCross(G \subseteq A_{gradient}) \equiv (\forall(g_1, g_2 \in G):(g_1 \cap g_2 \neq \varnothing) \rightarrow (g_1 = g_2))$$

- For each point p on the curve, there is a gradient. The expected value of this gradient is p:

$$ExpectedCurve(ec \in A_{Curve}, G \subseteq A_{gradient}) \equiv (\forall(p \in ec)\exists(g \in G):E(g.fg) = p)$$

- The expected value of all the gradient lines are on the central line:

$$CurveExpected(ec \in A_{Curve}, G \subseteq A_{gradient}) \equiv (\forall(g \in G)\exists(p \in ec):E(g.fg) = p)$$

For all of these functions, E(x) is the expected value for a probability density function.

To ensure that the type is implementable, the probability density values of points that are close to one another should have similar values. To ensure this, we use the condition that all iso-lines of probability must be continuous. This means that for all possible probability density values, the set of points formed from the points along all the gradient lines that have this probability density should form either a continuous line along the central curve or a set of continuous cycles. The set of points from all the gradient lines that have a given probability density value is returned by the *ISet* function, which is defined as follows:

$$ISet(i > 0, G \subseteq A_{gradient}) \equiv \{x|\exists(g \in G):(x \in g.gc \wedge g.fg(x) = i)\}$$

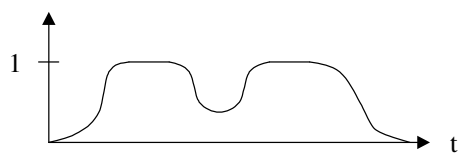

Figure 5: Probability Function Indicating Uncertainty about the Number of Curves

To ensure that the iso-lines are continuous cycles, the following condition is used:

$$ContIso(ec \in A_{Curve}, G \subseteq A_{gradient}) \equiv (\forall i:((Iset(i, G) \subseteq Points(ec)) \vee \exists(C \subseteq A_{cycle}):Finite(C) \wedge points(C) = Iset(i, G)))$$

The function points(C) returns a set containing all the points which are parts of at least one cycle in the set of cycles.

To compute the area in which the uncertain line may be, one can take the union of all the gradient lines. The following condition ensures that the union of all the gradient lines forms a crisp face:

$$FormFace(G \subseteq A_{gradient}) \equiv (\{x|\exists(g \in G):x \in g.gc\} \in A_{Face})$$

Definition 4: An *uncertain curve* is defined as follows.

$$A_{UCurve} \equiv \{(ec, fe, G)|ec \in A_{curve} \wedge G \subseteq A_{gradient} \wedge \forall(p \in ec):(fe:p \rightarrow A_{Prob}) \wedge NoDip(fe) \wedge NoCross(G) \wedge ExpectedCurve(ec, G) \wedge CurveExpected(ec, G) \wedge ContIso(ec, G) \wedge FormFace(G)\}$$

This type definition is quite complex, and is the most complex type of the three main ones. The reason for this is that a point is a probability density function, a region is a probability function where each point has a probability of being in the region. A curve, however, is a little of both, as shown in Figure 3.

Note that this definition of a curve does not allow a curve that is partially crisp and partially uncertain. This is because there would be a point where the uncertain area ends and the crisp area begins where the probability density function of the gradients rises until it becomes infinite. In this place, some of the iso-lines would not be cycles as they will end right next to the point where the line becomes crisp.

This problem can be solved by defining such a curve as a line with several curves, some uncertain and some crisp. The crisp curve is defined by having all its gradient lines have length 0 and their probability functions being dirac delta functions.

The probability densities along gradient lines that are near to each other are dependent on each other in such a fashion that the line must be continuous. The extent of this dependence depends on the line in question, but all of the gradient lines must obey the following principle:

Let us say that along gradient line A the line passes through point p. For any line B in the neighbourhood of A, the following holds:

$$\left(\lim_{B \to A} E(B_{|}(A = p)) = p\right) \wedge \left(\lim_{B \to A} V(B_{|}(A = p)) = 0\right)$$

In this formula, $E(X)$ is the expected value of X and $V(X)$ is the variance of X.

Both points and regions are defined as functions over the plane. To make it simpler to define operations which are common to all uncertain spatial types, a view of the uncertain curve as a function over the plane is therefore also given:

Computational definition. An uncertain curve may be defined as a function over the plane:

$$C.f(x, y) = gl.fg(x, y) \cdot C.fc(cp)$$

In this function, gl is the member of C.G on which the point (x,y) lies and cp is the point at which gl crosses C.ec.

The line type is a set of curves for the same reasons as given for points.

Definition 5: The *uncertain line* is defined as a set of uncertain curves.

$$A_{ULine} \equiv \left\{ UC \subseteq A_{UCurve} | Finite(UC) \wedge \forall(ac \in UC)\forall(bc \in UC): (ac \neq bc \rightarrow \neg Cross(ac, bc)) \right\}$$

The requirement that two curves should not cross is there to ensure the uniqueness of the representation. If two curves that cross are added to the same set, they must be divided so that all four get the crossing as their end points. The Cross operator is defined in Section 5.2.

One problem with this model for lines is that it involves fairly complex mathematics, such as finding gradients of a function. Also, some operations, such as testing whether two lines cross each other, are much more complex in this model than in models for crisp or vague lines. This complexity exists because the uncertain curve is neither a simple probability density like for the uncertain point nor a simple probability function for each point like in an uncertain face.

4.4 Uncertain Regions

An uncertain region is a set of uncertain faces. An uncertain face is one where the location of the boundary or even the existence of the face itself is uncertain. This may be modelled as a probability function P(x,y) which gives the probability that the point (x,y) belongs to the face. Support(P) must be a valid crisp face. Additionally, an alpha-cut operation must yield a valid crisp region for all input values between 0 and 1. The alpha-cut function is defined as follows:

$$\alpha cut(f, i) = \{z | f(z) > i\}$$

A more complete definition may be found in Section 5.3. Note that the Support operation is the same as an alpha-cut with i=0.

Definition 6: An *uncertain face* is defined as follows.

$$A_{UFace} = \{FP(x, y)| SProbFunc(FP) \wedge Support(FP) \in A_{Face} \wedge \forall(i \in [0, 1]):\alpha cut(FP, i) \in A_{Region} \wedge PieceCont(FP)\}$$

This definition is used because it is very general, and gives the capability of modelling uncertain regions in which the exact number of faces is unknown. This is possible because the uncertain face can have a core which contains multiple crisp faces. It also allows holes which are not certain to exist (such as the submerged islands in Example 1) because there may be an area with a function value less than one inside an area with function value one. An uncertain face is known to exist if at least one point has probability one of being a member of the face

Figure 6a shows an example of an uncertain face where the black area is the area in which the face is certain to exist and the grey area is the area of uncertainty. Figures 6b and 6c show views of the probability distribution along the X and Y axis..

For regions, a dependence condition similar to the one described for gradient lines in Section 4.3 holds for the individual points that the uncertain face contains. If one

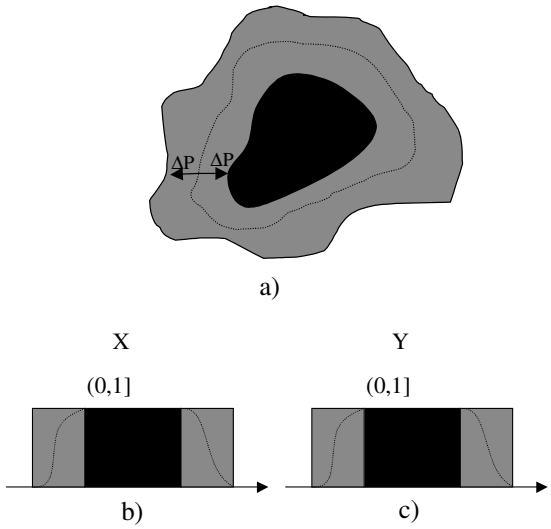

Figure 6: Uncertain Face

knows that point p is in the face F, the following condition holds for all points q in the neighbourhood of p:

$$\lim_{q \to p} P((q \in F)|(p \in F)) = 1$$

In this formula, $P(X)$ is the probability that X is in the region.

A similar condition also holds if it is known that p is not in the face:

$$\lim_{q \to p} P((q \in F)|(p \notin F)) = 0$$

These condition only holds for the continuous parts of the probability function of the face and not across any discontinuities.

Definition 7: The *uncertain region* is defined as a set of uncertain faces.

$$A_{URegion} \equiv \left\{ UF \subseteq A_{UFace} | Finite(UF) \wedge \forall(af \in UF)\forall(bf \in UF):(af \neq bf \rightarrow Disjoint(af, bf)) \right\}$$

Disjoint for uncertain types is defined as follows:

$$Disjoint(A, B) \equiv Union(Support(A), Support(B)) = \varnothing$$

This type of model for faces and regions has the advantage that such faces and regions are well documented for the vague case using fuzzy sets in (Schneider 1999) and (Erwig and Schneider 1997). It is also very general, capable of modelling any kind of uncertainty. Error-band based models can only model uncertainty about the position and shape, not the number of components or holes such as in Example 1 above.

5. OPERATIONS ON UNCERTAIN DATA

An important part of a set of data types is a general definition of the operations that can be applied to them. Some of the operations from (Güting et al. 2000) as well as a few new ones are described here. For a more complete overview of operations for uncertain spatial data, see (Tøssebro 2002). The operations are divided into three categories, those that are applied to data with no uncertainty, but which cannot be determined with certainty for uncertain data, those that can be applied to both kinds of data, and new operations for uncertain data.

Table 2: Type Designators[a]

Letter	Type
Po	Point
C	Curve
F	Face
S	Spatial (Point, Curve or Face)
Ss	Spatial Set (Points, Line or Region)
N	Number
T	Any non-spatial or spatial type
B	Boolean
Pr	Probability
D	Degree
CX	Crisp X
CI	Crisp interval

a. All these stand for uncertain data types except for CX and CI

In this section, the letter name of the variable describes its type as given in Table 2. A signature of the type $S \times S \rightarrow S$ means that both the inputs must be of the same type, and the output is of the same type as the input. In a signature of the type $S \times F \rightarrow B$, S may denote a type other than F or B.

In the semantics for the operations, the letter R is used for the result, $I1$ for the first input and $I2$ for the second input.

The Core and Support operations from fuzzy set theory will be used for operations on uncertain data. These have slightly different semantics than in the vague case

Table 3: Operations for which a Positive Answer is Impossible for Uncertain Data

Operation	Signature	Semantics
Equal	$S \times S \rightarrow B$	Maybe: $(core(I1) \cap support(I2) = core(I1)) \wedge$ $(core(I2) \cap support(I1) = core(I2))$ False otherwise
Touch	$F \times F \rightarrow B$	Maybe: $(core(I1) \cap core(I2) = \emptyset) \wedge (support(I1) \cap support(I2) \neq \emptyset)$ False otherwise

because of the differences between uncertainty and vagueness. Core is defined as follows.

- *Core($T \rightarrow CT$):* For a region, this operation returns the crisp set containing all the points or values having membership 1 in $I1$. For a complete definition, see Section 5.3.
- For *Support*, see Section 4 and Section 5.3.

5.1 Operations on Crisp Data which cannot be Determined with Certainty for Uncertain Data

The operations described in this subsection are listed in Table 3. They are much less useful for uncertain data because they cannot be determined with certainty. However, one may determine whether the operation is certainly false or not. The formula for determining this is given in the table.

Equal: One cannot determine equality between two uncertain objects. Even if the two objects have exactly the same type and probability function, they are not necessarily equal, because the real objects may be different even if the uncertain representations are "equal". For instance, if two regions both have the representation given in Figure 7, one of them can be bordered by line A and the other by line B. The regions produced by A and B are clearly not equal, but they can both correspond to the same uncertain region. The only way one can know that two uncertain objects are equal is if they are in fact the same object, with the same object identity or primary key value. The *Resemble* operation from Section 5.2 may be used to test approximate equality.

Touch: In the uncertain case this operation determines the possibility that two faces have a common border. Even if the supports overlap and the cores do not, one cannot be sure whether they actually have common borders or the borders just cross each other. Therefore, the operation cannot return "True" when there is uncertainty, unless the fact that the two faces have a common border is explicitly stored.

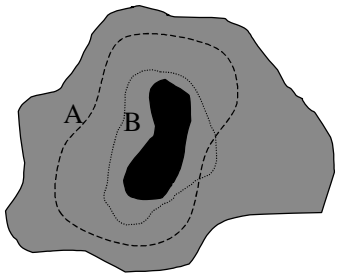

Figure 7: An Uncertain Region and two Possible "Real Regions"

5.2 Operations which may be used on both Crisp Data and Uncertain Data

The operations in this section are divided into five categories, depending on the types of their input and output: set operations, operations applicable to all uncertain spatial data types, operations for uncertain regions, operations for uncertain lines and projections. There are no operations that are only applicable to uncertain points and cannot also be applied to other types as well.

Set Operations

The set operations for the points and line data types are the same as in the crisp case. Using a set operation on the individual points and curves does not make sense. The point data type is not a set. Performing a set operation on a curve will most likely produce an illegal value. An uncertain region is in essence an infinite point set where the individual points have a certain probability of being

Table 4: Operations Applicable to All Uncertain Spatial Data

Operation	Signature	Semantics
Intersection	$Ss \times Ss \rightarrow Ss$	Points, Line: $I1 \cap I2$ Region: $R(x, y) = I1(x, y) \cdot I2(x, y)$
Resemble	$S \times S \rightarrow A_{Degree}$	$(area(min(I1, I2)))/(area(max(I1, I2)))$

members. Each set operation should therefore return a set that for each point gives the probability of the operation being true. This is easiest to do by combining the probability functions of the two input sets. Probability theory has been used to arrive at the formula given in Table 4. The events that a point belongs to regions *I*1 and *I*2 are considered to be independent.

The intersection operator returns a result of the lowest dimension of the two inputs. An intersection between a point and a line does not make sense in the uncertain case because the probability that the two are at exactly the same place is 0. An intersection between either a point or a line and a region uses the same semantics as the intersection of two regions. The output type is point or line.

Other Operations applicable to all Uncertain Spatial Data Types
Only one such operation is defined here. For a complete list, see (Tøssebro 2002). Its semantics is defined in Table 4.

Resemble: This operator determines how much two uncertain spatial objects resemble one another. It may be used to replace equal for uncertain objects. For crisp regions it is used to determine similarity in shape. The function min returns the minimum probability or probability density value of *I*1 and *I*2. Max returns the maximum.

Table 5: Operations for uncertain regions and lines

Operation	Signature	Semantics
Intersect	$S \times F \rightarrow A_{Prob}$	$Existence(I1 \cap I2)$
Cross	$C \times C \rightarrow B$	See text

Operations for Uncertain Regions
Intersect: This operator determines the probability that *I*1 and *I*2 intersect. This is the "overlap" criterion used in many spatial searches. The semantics of this operation is given in Table 5.

Operations on Uncertain Curves
Only the *Cross* operation is defined here. For a complete list, see (Tøssebro 2002). The semantics of the *Cross* operation is given in Table 5.

Cross: Determining whether or not two uncertain curves cross each other is far more complex than for crisp curves because one does not know quite where the curves are. If $support(I1) \cap support(I2) = \varnothing$, the two curves cannot cross. Otherwise they may cross. Computing the exact or even approximate probability that they cross is complex. In many cases a "Maybe" answer is sufficient. To know for sure, the following conditions must be checked:

- Both curves must exist in the entire area in which they may cross. The following formula test whether curve *I*1 exists in the entire area. The test is analogous for *I*2.

  $\forall(g \in I1.G):(g \cap ca \neq \varnothing) \rightarrow (I1.fe(g \cap I1.ec) = 1)$

  In this formula, $ca = support(I1) \cap support(I2)$.
- Let
  $ba = boundary(ca) \cap boundary(support(I1))$
  and
  $bb = boundary(ca) \cap boundary(support(I2))$.
  Both *ba* and *bb* must consist of at least two disjoint crisp curves. This condition prevents line *A* in Figure 8 from getting "Yes" to the question *Crosses*(*A*, *L*).
- Each component of *ba* and *bb* must cross line *ec* of the other line an odd number of times. This applies to both lines. This prevents lines such as line *B* in Figure 8 from getting "Yes" to the question *Crosses*(*A*, *L*).

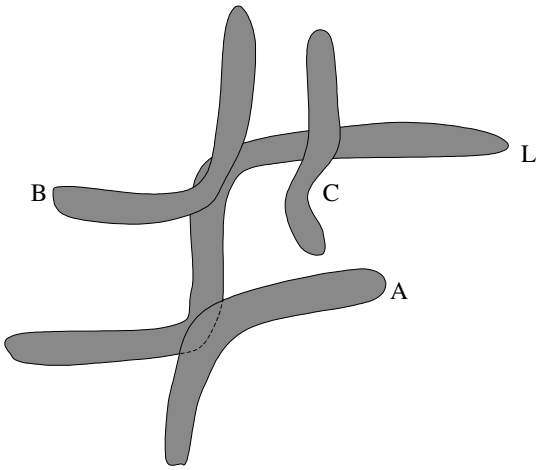

Figure 8: Two Curves which May Cross, and a Curve which Certainly Crosses, the Curve L

5.3 New Operations for Uncertain Data

These operations are new for uncertain data because they determine different aspects of that uncertainty. A list of these operations and their semantics is given in Table 6.

Alpha_Cut: This operation is described for fuzzy sets in (Zhan 1998). It returns a crisp set which contains all the points which have a membership value or probability density value above I2 in I1. The support operation can be seen as a special case of the alpha cut operation with I2=0.

Core: For regions, this operation returns the set of values which the object must contain. For lines it returns the

central line. It is defined for lines because some other operations, such as Equals, need this definition. For points and numbers, the Core does not exist.

Support: This operation returns the set of values which the object might possibly contain or be at.

6. DISCUSSION

The type system described is uniform in that all the types are defined in roughly the same way. For all the types there is a "region" indicating where the object might be. In this sense all the uncertain types are based on the crisp region. For all the types there is also a probability function indicating where the object is most likely to be. This means that many of the same operations may be run on all three types with little alteration needed.

Because the model presented here models points, lines and regions in a uniform way, it is better suited to a database like Example 5 than many previous models. Our model for uncertain regions is similar to Schneider's model from (Schneider 1999), but in our work that method is applied for all the data types. The models from (Mark and Csillag 1989), (Gohn and Gotts 1996) and (Wang and Hall 1996) only handle regions while (Dutton 1992) only handles points and lines. The models in (Lagacherie et al. 1996), (Cheng et al. 1997) and (Worboys 1998) are essentially rasters and therefore are poor at representing lines. The types presented here are better integrated than a system consisting of a point model, a line model and a region model from different authors chosen because they are good models for the individual types. We are currently working on the issue of the completeness and computational closeness of this model. We have already managed to convert from line to region (*enclosed_by* operation) and from region to line (*border* operation). This work will be published elsewhere.

Additionally, some of the operations from (Güting et al. 2000) are evaluated for use in the uncertain case. Some of the operations cannot give a "certainly true" result for uncertain data, but many can be used for both crisp and uncertain data. Some new operations (and operations from other sources) are also introduced to deal with the uncertainty. For uncertain Boolean values, two methods are used. The simplest is a three-value logic, which has been used earlier. However, due to the definitions of the types, many functions may instead return the likelihood of the answer being true.

One potential problem with our model is that the mathematical complexity will make it a challenge to implement, and that simpler models might be better in certain cases. This is particularly true for the uncertain line model. The reason for the complexity of the uncertain line is that the probability distribution for an uncertain line is neither a probability density function like for points nor a probability distribution function like for regions, but something in between. This and other issues related to implementation of our model and similar models are discussed in (Tøssebro and Nygård 2002b).

Part of the mathematical complexity discussed in the previous paragraph may be easily removed from the model at the cost of a decrease in expressiveness. The probability functions may be arbitrarily simple or complex. The simplest variant is the function with equal probability over the entire point or line, and with probabilities 1, 0.5 and 0 indicating the core, uncertain boundary and the outside of a region. The advantage of a simple function is that it is easier to store and faster to compute.

The advantage of complex functions is increased expressiveness. In some cases, the geologists making the measurements may make good educated guesses as to the probability function. Thus it would be an advantage to be able to store these. The advantages and disadvantages of models of different complexity are discussed further in (Tøssebro and Nygård 2002b) and (Tøssebro and Nygård 2003).

Table 6: New operations for selected uncertain data types

Operation	Signature	Semantics
Alpha_Cut	$S \times [0,1] \to \{CPo\}$ $N \times [0,1] \to \{CN\}$	Spatial types: $R = \{P \in CPo \mid A(P) > B\}$ Number: $R = \{N \in CN \mid A(N) > B\}$
Core	$S \to \{CPo\}$ $N \to CN$	Point, Number: $\emptyset$ Line: $R = \{P \in CPo \mid P \in A.ec \wedge A.fc(P) = 1\}$ Region: $R = \{P \in CPo \mid A(P) = 1\}$
Support	$S \to \{CPo\}$ $N \to CI$	Spatial types: $R = \{P \in CPo \mid A(P) > 0\}$ Number: $R = \{N \in CN \mid A(N) > 0\}$

REFERENCES

T. Beaubouef and F. Petry. 2001. Vagueness in Spatial Data: Rough Set and Egg-Yolk Approaches. In *Proc. 14th Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE)*, 367-373

E. Clementini. 2005. A model for uncertain lines. In *Journal of Visual Languages and Computing*, 16, 271-288

E. Clementini and P. Di Felice. 1996. An Algebraic Model for Spatial Objects with Indeterminate Boundaries. In *Geographic Objects with Indeterminate Boundaries*, GISDATA series vol. 2, Taylor & Francis, 155-169.

A. G. Cohn and N. M. Gotts. 1996. The 'Egg-Yolk' Representation of Regions with Indeterminate Boundaries. In *Geographic Objects with Indeterminate Boundaries*, GISDATA series vol. 2, Taylor & Francis, 171-187.

T. Cheng, M. Molenaar, T. Bouloucos. 1997. Identification of Fuzzy Objects from Field Observation Data. In S. C. Hirtle and A. U. Frank (eds.) *Spatial Information Theory: A Theoretical Foundation for GIS*, LNCS vol. 1329, Springer-Verlag, 241-259.

M. Duckham, K. Mason, J. Stell, M. Worboys. 2001. A formal approach to imperfection in geographic information. In *Computers, Environment and Urban Systems*, 25, 89-103.

G. Dutton. 1992. Handling Positional Uncertainty in Spatial Databases. *SDH'92*, vol. 2, 460-469.

M. Erwig, R. H. Güting, M. Schneider, M. Vazirgiannis. 1999. Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases. In *GeoInformatica* 3, no. 3, 269-296.

M. Erwig and M. Schneider. 1997. Vague Regions. In *Proc. 5th Symp. on Advances in Spatial Databases (SSD)*, LNCS 1262, 298-320.

R. H. Güting, M. F. Böhlen, M. Erwig, C. S. Jensen, N. A. Lorentzos, M. Schneider, M. Vazirgiannis. 2000. A Foundation for Representing and Querying Moving Objects. In *ACM Transactions on Database Systems* 25, no. 1.

P. Lagacherie, P. Andrieux and R. Bouzigues. 1996. Fuzziness and Uncertainty of Soil Boundaries: From Reality to Coding in GIS: In *Geographic Objects with Indeterminate Boundaries*, GISDATA series vol. 2, Taylor & Francis, 155-169.

K. Lowell. 1994. An Uncertainty-Based Spatial Representation for Natural Resources Phenomena. *SDH'94* vol. 2, 933-944.

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

M. Schneider. 1996. Modelling Spatial Objects with Undeterminate Boundaries using the Realm/ROSE Approach. In *Geographic Objects with Indeterminate Boundaries*, GISDATA series vol. 2, Taylor & Francis, 155-169.

M. Schneider. 1999. Uncertainty Management for Spatial Data in Databases: Fuzzy Spatial Data Types. In *Proc. 6th Int. Symp. on Advances in Spatial Databases (SSD)*, LNCS 1651, Springer Verlag, 330-351.

E. Tøssebro and M. Nygård. 2002a. Abstract and Discrete Models for Uncertain Spatiotemporal Data. In Proc. 14th Int. Conf. on Scientific and Statistical Databases (SSDBM), 240.

E. Tøssebro and M. Nygård. 2002b. An Advanced Discrete Model for Uncertain Spatial Data. In *Proc. 3rd Int. Conference on Web-Age Information Management (WAIM02)*, 37-51.

E. Tøssebro and M. Nygård. 2002c. Uncertainty in Spatiotemporal Databases. In *Proc. 2nd Biennial Int. Conference on Advances in Information Systems (ADVIS)*, 43-53.

E. Tøssebro. 2002. Representing Uncertainty in Spatial and Spatiotemporal Databases. Dr. Ing. Thesis at IDI, NTNU. IDI Report 2002:07.

E. Tøssebro and M. Nygård. 2003. A Medium Complexity Discrete Model for Uncertain Spatial Data. *To be published in Proc. 7th Int. Database Engineering and Applications Symposium (IDEAS),* 376-384.

F. Wang and G. B: Hall. 1996. Fuzzy representation of geographical boundaries in GIS. In *Int. Journal of Geographical Information Systems* 10, no. 5, 573-590.

M. F. Worboys. 1998. Imprecision in Finite Resolution Spatial Data. In *GeoInformatica* 2, no.3, 257-279.

L. A. Zadeh. 1965. Fuzzy sets. *Information and Control* 8, 338-353.

F. B. Zhan. 1998. Approximate analysis of binary topological relations between geographic regions with indeterminate boundaries. *Soft Computing* 2, 28-34.