Managing Uncertain Spatio-Temporal Data

Thomas Bernecker, Tobias Emrich, Hans-Peter Kriegel, Andreas Zuefle
Institute for Informatics
Ludwig-Maximilians-Universität München
Oettingenstr. 67, 80538 München, Germany
{bernecker,emrich,kriegel,zuefle}@dbs.ifi.lmu.de

Lei Chen, Xiang Lian
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{leichen,xlian}@cse.ust.hk

Nikos Mamoulis
Department of Computer Science
University of Hong Kong
Pokfulam Road, Hong Kong
nikos@cs.hku.hk

ABSTRACT

Many spatial query problems defined on uncertain data are computationally expensive, in particular, if in addition to spatial attributes, a time component is added. Although there exists a wide range of applications dealing with uncertain spatio-temporal data, there is no solution for efficient management of such data available yet. This paper is the first work to propose general models for spatio-temporal uncertain data that have the potential to allow efficient processing on a wide range of queries. The main challenge here is to unfold this potential by developing new algorithms based on these models. In addition, we give examples of interesting spatio-temporal queries on uncertain data.

1. INTRODUCTION

In the past two decades, the problem of modeling and managing uncertain data has received a great deal of interest, due to its manifold applications in spatial, temporal, multimedia and sensor databases. There exists a wide range of work covering spatial uncertainty in the static (snapshot) case, where only one point of time is considered. In contrast the problem of modeling and managing uncertain data with a temporal component has only received very little attention by the community.

Consider the problem of monitoring iceberg activity in the North Atlantic. Ships transiting between Europe and east coast ports of Canada and the US traverse a great circle route that brings them into the vicinity of icebergs carried south by the cold Labrador Current near the Grand Banks. It was here that the R.M.S. Titanic sank in 1912, after it struck an iceberg. This disaster resulted in the loss of 1517 lives and led directly to the founding of the The International Ice Patrol (IIP) in 1914. The mission of the IIP is to monitor iceberg danger near the Grand Banks of Newfoundland and provide the limits of all known ice to the maritime community. The IIP does this by sighting icebergs, primarily through airborne Coast Guard reconnaissance missions. In addition to visual observations from ships and aircraft, the IIP makes use of information from drifting buoys, radar, and side-looking airborne radar (SLAR) as well as model output. The positions and the extent of all currently known icebergs, as well as the calendar date of its last sighting/resighting are stored in a database. In addition, the confidence of the reporting source is stored in the database. For example a visual sighting has a very high confidence, while a garbled radar signal has a lower confidence. Thus, the position of an iceberg at a time t underlies two sources of error:

- the observation measurement error and
- the obsoleteness of the most recent observation

Using a model to describe this uncertainty, we derive at each time t and each uncertain object o a probabilistic density function describing the position of o at t.

2. RELATED WORK

The problem of querying spatio-temporal data has been studied extensively. There exist numerous publications on efficient query evaluation for the case where the attribute values at each time t are known for certain (for a comprehensive coverage, see [1]). From this body of work, our approach is mostly related to spatio-temporal data indexing for predictive querying, for example indexes like [2, 3] and approaches like [4]. Still, these papers neither consider probabilistic query evaluation nor model the data with stochastic processes.

However, in scenarios where data are inherently uncertain, such as sensor databases, answering traditional queries using expected values is inadequate, since the results could be incorrect [5]. In such cases, probabilistic queries that take the full information of the underlying uncertainty into account and that yield results with probabilistic guarantees are required.

One of the first works that deal with uncertainty in trajectories is [6]. This work considers routes that are captured by GPS and assumes that the recorded locations are uncertain. Indexing such data for range queries is considered; the authors use a simple model that sums up the probabilities that the trajectory points are included in the range queries to derive the probabilities of the results. Trajcevski et al. [7], [8] follow a similar approach for the same problem settings. At each point in time the position of an object is modeled as an ellipse. Each trajectory is thus represented by a 3D cylindrical body. Since no assumptions are made about the probability distribution inside the ellipses only binary answers to queries are possible. For example the model can answer if an object is certainly within a query region or could be inside a query region during a time interval but not give a probability to those events.

The work of [9] is based on the same model as [7, 8]; the authors assume a database of (historical) uncertain trajectories, each having a 3D cylindrical body. The objective is to identify the nearest neighbors of an uncertain query trajectory, throughout its lifetime; i.e., to partition the lifetime into intervals, each containing a stable most probable nearest trajectory from the database. Each interval is then partitioned recursively according to the second most probable neighbor, etc. Again, this work falls into the category of papers that do not consider location dependencies between consecutive timestamps and do not rely on stochastic models.

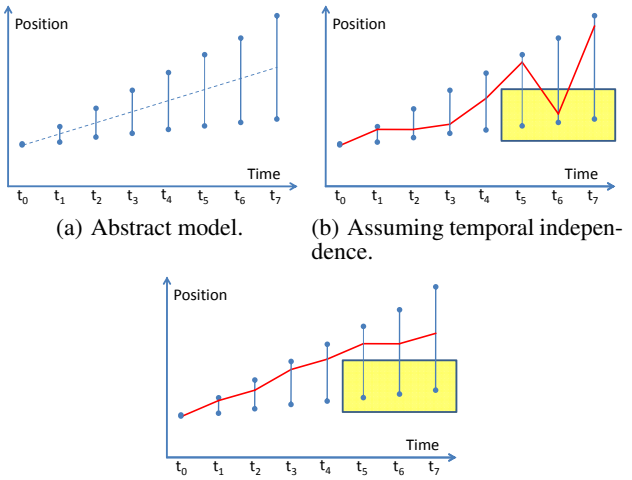
Cheng et al. [10] consider possible world semantics on static data with multidimensional or interval PDFs. A wide range of queries is studied. In [11], the model is extended to support search in databases with uncertain trajectories. Similar to [7], recorded trajectories (e.g., from GPS data) are spatially extended to capture all possible locations that the object may have passed through. Then, indexing is used to prune regions in space and time (together with the corresponding trajectory data) that do not satisfy the queries and possible worlds semantics are used to refine the overlapped areas in order to determine the probabilistic query results. Again, this work ignores inference based on stochastic models.

Approaches like [12] and [13] consider uncertain time series and data streams, respectively. Similar to other work, they also disregard correlations between points in time; that is, the position of an object at time t is assumed to be independent of its previous position at time $t-1$.

Mokhtar and Su [14] describe a model where the uncertainty region of each object is described by a time dependent stochastic process. Objects are given by MBRs which change their location and extent over time following the stochastic process. The paper shows how to answer certain types of window queries based on this model. However, describing the parameters of the uncertainty regions and not the trajectories of the objects through a stochastic process yields wrong results regarding to possible worlds semantics. The reason is that location dependency between consecutive timestamps is ignored by this model.

The work described in [15] focuses on the prediction of uncertain trajectories in street networks. The authors propose the use of time-dependent inhomogeneous Markov processes for each crossing. This so called Trajectory Continuous Time Bayesian Network is constructed by analysing training data. Afterwards, it is used to predict the next movement of an object when arriving at a crossing. The proposed algorithm shows very high accuracy rates of around 80% predicting routes of objects. However the system does not consider data management and efficiency in query evaluation, but only tries to predict the routes of single objects.

To illustrate the problem of previous approaches disregarding temporal dependencies, consider Figure 1(a), where an uncertain object trajectory is modeled. Here, it is assumed that the object



(a) Abstract model.

(b) Assuming temporal independence.

(c) Not assuming temporal independence.

Figure 1: Modeling spatio-temporal data.

moves with an uncertain speed upwards. The speed of o may change over time, but will not drop below some minimal speed greater than zero and will not exceed some maximum speed. Therefore, given the position of an object o at time t_0, the future position of an object can be modeled using the expected speed of o (the dashed line in Figure 1(a)), and lower and upper bounds, or alternatively a variance, depicted by the intervals at each point of time. Since at each point of time, the positions of o are modelled as independent random variables, a trajectory such as depicted in Figure 1(b) has a probability greater than zero. However, this trajectory is actually not possible, since o makes a large leap backwards between times t_5 and t_6, which is not possible given knowledge about the movement of o, which this model should incorporate. A possible trajectory is shown in Figure 1(c). Here, the object moves within its speed limits at each points of time.

The flaw of modeling trajectories which are not actually possible becomes a problem when processing spatio-temporal queries based on this model. For example, consider a spatio-temporal window query, which is to return for an object o, the probability that o intersects the query window q, depicted in Figures 1(b) and 1(c). For any model that ignores the dependency between locations at subsequent points of time, the probability that o is always outside the window is the product of many probabilities and gets very small. Thus, for a large number of points of time inside the query region, the probability that o intersects the query window converges to one. However, if the dependency between locations at subsequent points of time is considered, then the probability that o is outside the window at time t_6 depends on the probability at time t_5. If o is not in the window q at t_5, then it cannot be in q at t_6 either, since the object cannot move backwards. Thus, the probability that o intersects the query window q at any time, is equal to the probability that o intersects the query window at time t_5. This is intuitive because the object cannot move back into the window. One of our aims in this work is to properly model such dependencies, instead of simply treating time as an additional dimension in space.

To summarize, all works so far on querying uncertain spatio-temporal data assume that the location probabilities of an object at two different times are independent. However, time-dependence is the main characteristic of temporal data, which cannot simply be ignored and this is the focus of this work.

3. STOCHASTIC PROCESSES FOR MODELING UNCERTAIN SPATIO-TEMPORAL DATA

In order to manage uncertain spatio-temporal data we should first select a suitable mathematical model for the data. To permit the existence of uncertainty in the data, we suppose that the attributes of an uncertain object X at time t are realizations of a random variable X_t. This consideration suggests modeling the data as a realization of a stochastic process [16] $\{X_t \in S, t \in T\}$, where T is the temporal domain, and S is the spatial domain. Depending on the model, T and S can be continuous or discrete. In the following, we review a few examples of common stochastic processes.

3.1 Stochastic Differential Equations

A general model describing spatio-temporal uncertainty in a continuous temporal and spatial domain without any assumption on the distribution is given by stochastic differential equations, having the form

$$dX_t = a(t, X_t)dt + b(t, X_t) \cdot \xi$$

where a is the drift of X_t, b measures the intensity of the deviation of X_t and ξ is an error of arbitrary distribution.
For more details on stochastic differential equations the interested reader is referred to [17]. At this point, we only need to note that dealing with stochastic differential equations is not computationally viable, since even the evaluation of a path (i.e. a possible world) of an object requires expensive numeric integration. Therefore, different models are required that are more practical and still model most applications well.

3.2 Time-Parameterized Parametric Distributions

The position X of an iceberg is a random variable that can be modeled by a Gaussian process with a drift. Therefore, the expected position $\mu_t(X)$ of an iceberg is a function of t that moves into a certain direction (which is estimated using the previous trace of the iceberg, or by using empiric studies of other icebergs in the same region). In addition, the deviance $\sigma_t(X)$ from the expected position is also a function of t. The initial value of $\sigma_t(X)$ depends of the type of observation. Over time, $\sigma_t(X)$ increases until the position of the iceberg is re-sighted. When the iceberg is re-sighted, $\mu_t(X)$, $\sigma_t(X)$ as well as the expected drift are updated.

In some situations, we can make the assumption that the error at a time (i.e. the deviation from the expected position) follows a certain parametric distribution. The parameters may change over time and are thus modeled by a function over time. For instance, if a normally distributed error is assumed, then the distribution X_t of an object X at time t is is given as

$$X_t \sim N(\mu(t), \sigma(t)),$$

where $N(\cdot, \cdot)$ represents the normal distribution and $\mu(t)$ and $\sigma(t)$ are unary functions. For example, the position of an iceberg can be modeled by a normal distribution, with an expected position $\mu(t)$ moving in the direction of the current, and a variance $\sigma(t)^2$ that slowly increases, until a new observation of the iceberg is made and the variance is set to a small value depending on the quality of the sighting.

3.3 Discrete Markov Chain Model

In the (first-order) Markov chain model, a discrete temporal and spatial domain is assumed. Therefore, let $S = \{s_1, ..., s_{|S|}\}$ be a finite set of locations and let $T = \mathbb{N}_0$ be the time domain. In addition, the Markov chain model uses the assumption that the location of an uncertain object O at time $t + 1$ only depends on the location of O at time t.

DEFINITION 1. *A stochastic process $O = \{O_t, t \in T\}$ is called a Markov chain if and only if $\forall t \in \mathbb{N}_0 \forall j, i, i_{t-1}, ...i_0 \in S$:*

$$P(O_{t+1} = j | O_t = i, O_{t-1} = i_{t-1}, ..., O_0 = i_0) = P(O_{t+1} = j | O_t = i)$$

The conditional probability

$$P_{i,j}(t) := P(O_{t+1} = j | O_t = i)$$

is the (single-step) *transition probability* of location i to location j at time t.

DEFINITION 2. *A Markov Chain is homogeneous iff the transition probabilities are independent of t, i.e. $P_{i,j}(t) = P_{i,j}$.*

An advantage of this model is that the transitions between locations over time can be performed using matrix multiplications, for which there exist very efficient solutions. The locations of icebergs for example can be modeled by discretizing the spatial domain (e.g. using a grid), and defining the transition probabilities depending on the current of the sea. Generalizations of the Markov chain models are

- the discrete Markov process, in which time is modeled continuously and
- the continuous Markov process, in which both time and space are modeled continuously.

3.4 Continuous Markov Chain Model

In order to reduce the complexity of the model, we can partition the state space into a finite set. This is not even necessary for many random variables that are defined on a finite space. For example, consider a customer database of an insurance company where for each customer, the number of damage events is stored. Obviously, the number of damage events is a discrete ($\in \mathbb{N}$) variable. It may be interesting to perform a prediction of this database in the future. Therefore, for each future point of time $t + \delta t$, the number of damage events has to be estimated. This estimate depends on the previous number of damage events of a customer: The higher the previous frequency of damage events, this higher the probability that another damage event will occur.

For each object o and each state s, the *residence time* $R_{o,s}$ describes the time that o remains in state s. Clearly, $R_{o,s}$ is a random variable. Once $R_{o,s}$ has passed, o switches into another (or possibly the same) state, the successor state $S_{o,s}$ (note that the above example describes a special case, where the successor state of state n is always $n + 1$.). The successor state also follows a probabilistic distribution function.

The challenge for the community is to develop efficient query processing algorithms on top of these models. In particular, the task is to find polynomial algorithms for interesting spatio-temporal queries, and to enhance these by indexing and pruning techniques. A small sample of interesting spatio-temporal queries on uncertain data is given in the following.

4. PROBABILISTIC SPATIO-TEMPORAL QUERIES

Our goal is to efficiently evaluate probabilistic spatio-temporal queries on uncertain spatio-temporal objects; i.e., queries about

objects that are probably located in a given spatial region during a given range in time. Within the scope of this paper, we assume a set of uncertain spatio-temporal objects $\mathcal{D}$, i.e. objects associated with uncertain object trajectories $o(t)$, and focus on spatio-temporal queries specified by the following parameters:

- A spatial region $S^{\Box} \subseteq S$, i.e. a set of (not necessarily connected) locations in space, and
- a set $T^{\Box} \subseteq T$ of (not necessarily subsequent) points in time.

In the remainder, we use $Q^{\Box} = S^{\Box} \times T^{\Box}$ to denote the query ranges in the space and time domain. The most intuitive definition of a probabilistic spatio-temporal (PST) query is given below:

DEFINITION 3 (PST (EXISTS) QUERY). *Given a query region $S^{\Box}$ in space and a query region $T^{\Box}$ in time, a* probabilistic spatio-temporal exists query *(PST∃Q), retrieves for each object $o \in \mathcal{D}$ the probability $P(o(t) = s) \in [0, 1]$ that o is located in $S^{\Box}$ at some time $t \in T^{\Box}$.*

This query type has been studied before (e.g. in [7, 8]), albeit over data models that disregard dependencies between locations at consecutive time stamps, as we have discussed in Section 2.

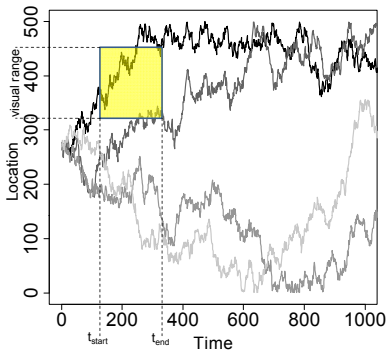


Figure 2: Example of a spatio-temporal query.

For our motivating application described in the introduction, an exemplary query could be to return all icebergs that have been in visual range of a certain signal tower in December 2010 with a probability of at least 30%.

The challenge is to return a correct answer that complies with the possible world semantics. That is, for each object O, return the fraction of possible worlds for which at any time $t \in \{t_{start}, ..., t_{end}\}$ it holds that $O_t \in Q^{\Box}$. An example is given in Figure 2, where four possible worlds (trajectories) of an object O are depicted. Any world for which the corresponding path intersects the spatio-temporal query window satisfies the query predicate. The challenge is to compute this probability (i.e. the fraction of worlds) efficiently, that is without enumeration of all (exponentially many) possible worlds.

In addition, we study the following two interesting probabilistic query variants. Note that the second variant has not been considered in the past:

DEFINITION 4 (PST FOR-ALL QUERY). *A* probabilistic spatio-temporal for-all query *(PST∀Q) retrieves for each object $o \in \mathcal{D}$ the probability $P(o(t) = s) \in [0, 1]$ that o remains in $S^{\Box}$ for* all *times $t \in T^{\Box}$.*

DEFINITION 5 (PST k-TIMES QUERY). *A* probabilistic spatio-temporal k-times query *(PSTkQ) retrieves for each object $o \in \mathcal{D}$ and each parameter $1 \leq k \leq |T^{\Box}|$ the probability that o is located in $S^{\Box}$ at* exactly *k times $t \in T^{\Box}$.*

PST∀Q and PSTkQ are important complements to the PS∃TQ. For example, these queries can progressively determine candidates that remain in a certain region for a while. For example, for a given region somewhere in the north Atlantic we want to retrieve all icebergs that have non-zero probability remaining in this region for a specified period of time, e.g. to be able to make some measurements over a certain time period. Further examples where such queries are useful are for location-based-service (LBS) applications, e.g. a service provider could be interested in customers that remain at a certain region for a while, such that they can receive advertisements relevant to the location.

Additionally to the above mentioned query types, many more queries can be thought of, for example: Given a spatial region $S^{\Box} \subseteq S$ of (not necessarily connected) locations in space and a set $T^{\Box} \subseteq T$ of (not necessarily subsequent) points in time, return all uncertain objects O such that:

- **kNN Query:** With a probability of at least τ, O is a k-nearest neighbor of a specific location $s \in S^{\Box}$ for at least (exactly) n points of time $t \in T^{\Box}$.
- **Top-k-Query:** O is among the k objects having the highest probability to be at a location $s \in S^{\Box}$ at any time $t \in T^{\Box}$.

5. CONCLUSION

In this paper, we studied the problem of probabilistic query evaluation over uncertain spatio-temporal data. We consider uncertain trajectories, for which some points are sampled via observations, while the remaining points are instantiated by a stochastic process. To the best our knowledge, this is the first paper that studies such queries over uncertain moving object data, which are modeled by stochastic processes, specifically Markov chains. This approach has two major advantages over previous work:

- It allows answering queries in accordance with the possible worlds model, and
- dependencies between object locations at consecutive points in time are taken into account.

We foresee that the combination of newly developed pruning techniques, stochastic calculations and index structures can lead to efficient solutions to the mentioned query types. Additionally we believe that many more applications will arise from this basic idea, so the above list is on our opinion just "the tip of the iceberg".

6. ACKNOWLEDGEMENTS

This work was supported by a grant from the Germany/Hong Kong Joint Research Scheme sponsored by the Research Grants Council of Hong Kong (Reference No. G HK030/09) and the Germany Academic Exchange Service of Germany (Proj. ID 50149322)

7. REFERENCES

[1] R. H. Güting and M. Schneider, *Moving Objects Databases*. Morgan Kaufmann, 2005.

[2] S. Saltenis, C. S. Jensen, S. T. Leutenegger, and M. A. Lopez, "Indexing the positions of continuously moving objects," in *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX*, 2000.

[3] C. S. Jensen, D. Lin, and B. C. Ooi, "Query and update efficient b+-tree based indexing of moving objects," in *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), Toronto, Canada*, 2004.

[4] Y. Tao, C. Faloutsos, D. Papadias, and B. L. 0002, "Prediction and indexing of moving objects with unknown motion patterns," in *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Paris, France*, 2004.

[5] G. Beskales, M. Soliman, and I. Ilyas, "Efficient search for the top-k probable nearest neighbors in uncertain databases," *PVLDB*, vol. 1, pp. 326–339, 2008.

[6] D. Pfoser and C. S. Jensen, "Capturing the uncertainty of moving-object representations," in *Proceedings of the 6th International Symposium on Large Spatial Databases (SSD), Hong-Kong*, 1999.

[7] G. Trajcevski, O. Wolfson, F. Zhang, and S. Chamberlain, "The geometry of uncertainty in moving objects databases," in *Proceedings of the 8th International Conference on Extending Database Technology (EDBT), Prague, Czech Republic*, 2002.

[8] G. Trajcevski, O. Wolfson, K. Hinrichs, and S. Chamberlain, "Managing uncertainty in moving objects databases," *ACM Trans. Database Syst.*, vol. 29, no. 3, pp. 463–507, 2004.

[9] G. Trajcevski, R. Tamassia, H. Ding, P. Scheuermann, and I. F. Cruz, "Continuous probabilistic nearest-neighbor queries for uncertain trajectories," in *Proceedings of the 13th International Conference on Extending Database Technology (EDBT), Saint-Petersburg, Russia*, 2009.

[10] R. Cheng, D. V. Kalashnikov, and S. Prabhakar, "Evaluating probabilistic queries over imprecise data," in *Proceedings of the ACM International Conference on Management of Data (SIGMOD), San Diego, CA*, 2003.

[11] ——, "Querying imprecise data in moving object environments," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1112–1127, 2004.

[12] J. AÃ§falg, H.-P. Kriegel, P. Kröger, and M. Renz, "Probabilistic similarity search for uncertain time series," in *Proceedings of the 21st International Conference on Scientific and Statistical Database Management (SSDBM), New Orleans, LA*, 2009.

[13] M.-Y. Yeh, K.-L. Wu, P. S. Yu, and M. Chen, "Proud: A probabilistic approach to processing similarity queries over uncertain data streams," in *Proceedings of the 13th International Conference on Extending Database Technology (EDBT), Saint-Petersburg, Russia*, 2009.

[14] H. Mokhtar and J. Su, "Universal trajectory queries for moving object databases," in *Mobile Data Management*, 2004.

[15] S. Qiao, C. Tang, H. Jin, T. Long, S. Dai, Y. Ku, and M. Chau, "Putmode: prediction of uncertain trajectories in moving objects databases," *Appl. Intell.*, vol. 33, no. 3, pp. 370–386, 2010.

[16] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*. Academic Pr Inc, 1975, vol. 2.

[17] I. Gikhman and A. Skorokhod, *The Theory of Stochastic Processes*. Berlin: Springer, 1974, 75, 79, vol. 1, 2, 3.