MANAGING UNCERTAINTY IN SPATIO-TEMPORAL SERIES

Yania Molina Souto, Ana Maria de C. Moura, Fabio Porto
Laboratório de Computação Científica – LNCC – DEXL Lab – Petrópolis – RJ– Brasil
yaniams@lncc.br, anamoura@lncc.br, fporto@lncc.br

Abstract. *Uncertain time series analysis has recently become an important research topic, particularly when searching for features of natural phenomena using similarity functions. Natural phenomena are often modeled as time series, such as in weather forecast, in which temperature variation is monitored through space and time. In such a context, different models for weather forecast produce variations on predictions that can be interpreted as predictions uncertainty. One important problem is to represent the variations presented in predictions along space and time. In order to address a solution to this problem, this paper defines a new type of series, here named uncertain spatio-temporal series, and proposes a computational strategy to manage uncertainty in probabilistic database. Using this new series some analytical queries can be performed, leading to the discovery of interesting observation patterns.*

1. Introduction

Recently, uncertain data management has received great attention from the scientific community [Dan Suciu, 2011]. Among the main causes that generate data uncertainty, some of them can be mentioned: increasing use of sensors data; multiple sources data integration inconsistencies; current privacy policies of information, where data is disturbed to safeguard the identity of their owners; information transmitted over the network and corrupted in the process; the growing interest for the management of applications involving moving objects where the position is regularly updated; in different application domains data is generated using complex models, after which they are published on the web. Hence, there is neither track of the original data sets nor of the process they have been submitted to.

In applications such as the weather forecast, the temperature in a given region is predicted using measurements obtained from sensors, or by inferring from historical data published in the weather sites for several years. However, when temperature is forecasted using a single model, a time series can be used to analyze temperature changes, since for each timestamp a single value is measured and no uncertainty is taken into consideration.

Differently from single model forecasting, *ensemble* forecast [Dufek, 2015] involves the analysis of results from different models. In such scenario, the variations on model predictions can be interpreted as prediction uncertainty. Modeling uncertainty in ensemble forecasting requires having multiple values for each timestamp, in a particular point in space, departing from traditional time-series representation.

In this case, for each timestamp sets of predictions are obtained with different probabilities of occurrences, and various levels of uncertainty measured by different devices or models. In this context, a new time series model has been defined **[**Abfalg, 2009], the so-called *uncertain time series* (UTS). Consequently, uncertainty can be naturally assessed through a probabilistic or deterministic similarity measure, considering the set of values in each time slot. Some works have been developed in this direction [Abfalg, 2009], [Mi-Yen Yeh, 2009], [Sarangi, 2010], [Dallachiesa, 2012], [Orang, 2014], and some of them are reviewed in Section 2.

Similarity measures in uncertain time series make it possible to analyze the uncertainty of the variable studied with respect to its temporal component. However, it is known that most natural phenomena depend on the spatial and temporal components. To illustrate how the use of other components (such as altitude) can help in the analysis of an event, consider the example shown in Figure 1, where different sensors collect temperature measurements during 5 consecutive days in a specific region.

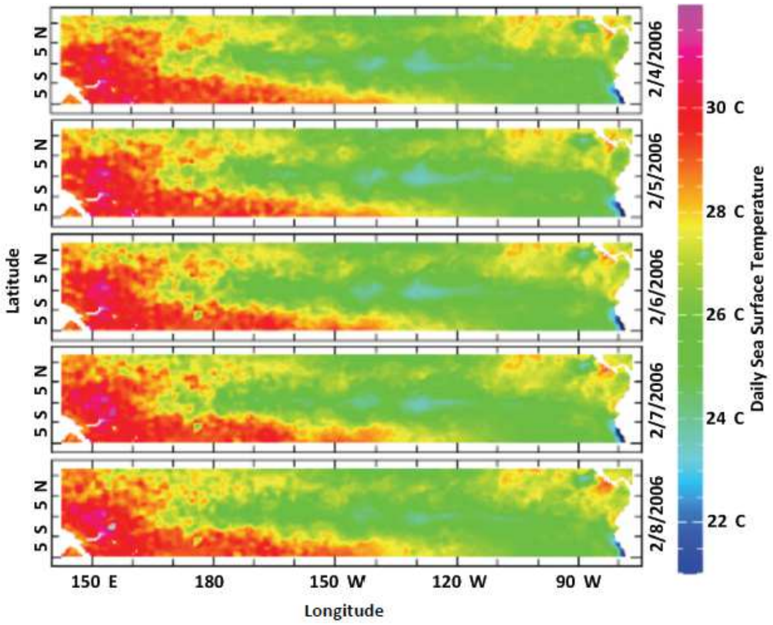


Figure 1: A time series of daily SST between the dates of 2/4/2006 and 2/8/2006 measured by the NOAA AVHRR satellite. Font: [McGuire, 2013]

It is possible to see some areas where temperature changes are visible, whereas in other areas these changes are not so obvious, due to the high volume of information. Furthermore it is known that the altitude is a variable that changes the atmospheric temperature, so that the spatial location is an important aspect to study this phenomenon. Moreover, identifying regions with similar spatial and temporal conditions helps to understand the dynamics of these weather events. In this case the uncertainty not only depends on the temporal component, which hinders analysis by traditional methods or uncertain time-series approaches.

In the literature uncertainty in spatio-temporal series has not been explored yet. This work aims at studying the inclusion of spatial components in the phenomenon analysis, as well as proposing a computational strategy to manage uncertainty in probabilistic database, making use of the UpsilonDB model [Gonçalves and Porto, 2014], which manages the uncertainty in numerical simulation data.

Being able to compute uncertainty helps not only deal with noise in the data, but contributes as well to the efficiency of clustering techniques, classification, outlier detection and probabilistic queries in Big Data problems. One simple reason can be theoretically used to justify this fact: the larger the number of measurements for the same observation, the closer their average to the expected value, which consequently reduces uncertainty. On the other hand, regions of datasets with higher levels of uncertainty can be discarded in the early stages of the investigation.

The rest of this paper is structured as follows: Section 2 describes the main concepts around Uncertain Time Series (UTS); Section 3 extends the UTS theory to define uncertainty in Spatio-Temporal Series; Section 4 describes the UpsilonDB system and shows how to model Uncertain Spatio-Temporal Series using UpsilonDB; and finally, Section 5 concludes the paper with suggestions for future work.

2. Uncertain Time Series: Background

In this section, a formalization for uncertain time-series is presented. Furthermore, techniques to compute the uncertainty of values based on similarity measures are discussed.

Definition 1 (Uncertain Time Series) [Abfalg et al., 2009]: an uncertain time series *X* of length n consists of a sequence <X_1, X_2, X_3, ..., X_n> of *n* elements, where each element X_t contains a set of *s* d-dimensional points (sample observations), i.e. $X_t = \{x_{t,1}, x_{t,2}, ..., x_{t,s}\}$ with $x_{t,i} \epsilon \mathbb{R}^d$. Where *s* is the sample size of *X*. Figure 2 shows an uncertain time series in $\mathbb{R}^2$.

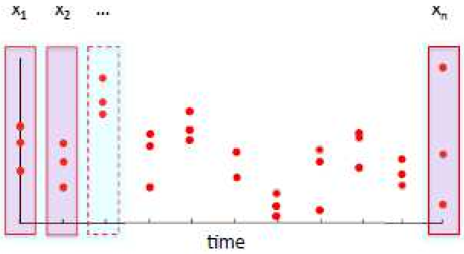


Figure 2: Example of uncertain time series $X = \{X_1, X_2, ... X_n\}$ in $\mathbb{R}^2$. Font: (Dallachiesa, 2012)

Consider two time series X = ((x_1, t_1), (x_2, t_2), ..., (x_i, t_i)) and Y = ((y_1, t_1) , (y_2, t_2), ..., (y_i, t_i)), where each value x_i (or y_i) is a point in the d-dimensional space to the timestamp t_i. It is called *similarity measure* the distance between X and Y in the timestamp i.

In the literature, various approaches to calculate similarity between time series and data sequences have been described. The three most important ones to measure uncertainty in UTS are: *Euclidean distance* (ED) [Faloutsos, 1994], L_p-norms (Lp) and *Dynamic Time Warping* (DTW) [Donald J Berndt, 1994]. These measures have been adapted to compute uncertainty in uncertain time-series. Next section will present the most important approaches to deal with this problem.

2.1 Uncertain similarity measure

Uncertainty can be measured in two ways: probabilistically and deterministically. For this reason similarity measures may be classified as [Orang and Shiri, 2014]:

Probabilistic Similarity Measures (PSM): distance is determined using a probabilistic distribution function (pdf), a combination of some pdfs, or simply calculating a probability for each value in each timestamp i. The *Lp-norm distance* is one of the most analyzed similarity measures. [Abfalg 2009] assumes that for two uncertain time series X and Y, the distance is given by all probable distances between each independent observed values at different timestamps. Distances are calculated using a L_p*-norm distance* with p = 2, which corresponds to the Euclidean distance. These distances are summarized through an empirical distribution function, reflecting the distribution of all possible distance values between the samples of the corresponding uncertain time series. A threshold ϵ is defined in order to decrease the exponential volume calculation. A distance is considered feasible if it is smaller than this threshold. These upper and lower distance bounding can ensure efficiency and decrease computing workload. Applying the same procedure [Abfalg, 2009], used a *DTW-distance, as* defined by Berndt (1994), to compute uncertainty.

Other works define new uncertain similarity measures. Mi-Yen Yeh (2009) calculates the Euclidian distance between uncertain time series X and Y as a pdf with a corresponding mean and variance: $Dist(X,Y) = \sum(x_i - y_i)^2$. Assuming that the distances D_i^2 are independent random variables with mean $E(D_i^2)$ and variance $V(D_i^2)$, by the Central Limit Theorem, the distance *Dist(X,Y)* is a random variable which tends to a normal distribution with a corresponding mean and variance $Dist(X,Y) \propto N(\sum E(D_i^2), \sum Var(D_i^2))$.

Orang (2014) formulates the notion of normalization for UTS, using a model for uncertain correlation as a normal random variable. For this purpose, they assumed that the random variables in the given UTS are independent and identically distributed, and that the only available information is the expected value and variance at each timestamp. Whereas all are independent random variables, the correlation coefficient has been expressed as the sum of two variables of the same type. The methodology to compute the similarity between two series is very similar to that used by [Mi-Yen Yeh, 2009]. According to the Central Limit Theorem, as n increases, *corr(X,Y)* approaches the normal distribution $corr(X,Y) \propto N(E(corr(X,Y)), Var(corr(X,Y)))$.

Deterministic Similarity Measures (DSM): In this case, a real distance is calculated between two UTSs. The most used DSM is *DUST-distance.* This similarity measure was defined in [Sarangi, 2010]. To calculate this distance, it is necessary to know the distribution of the data. Dust is defined as $DUST(X,Y) = \sqrt{\sum dust(x_i, y_i)^2}$ where $dust(x_i, y_i) = \sqrt{-log(\emptyset(|x_i - y_i|)) - \mathrm{k}}$ and $k = -\log(\emptyset(0))$. The term $\emptyset(|x_i - y_i|)$ is calculated using the distribution function specified by the user.

There are other approaches that define uncertain moving averages as filters to reduce noise in the data. Through the calculation of average values, the values of each series are recalculated. Some similarity measures can then be applied to these recalculated new values and depending on the approach, probabilistic or deterministic similarity can be obtained. Examples of uncertain moving averages are UMA and UEMA defined in [Dallachiesa, 2012].

3. Uncertainty in Spatio-Temporal Series

A spatio-temporal series is an extension of a time series, where spatial and temporal components qualify the studied variable. Having discussed, in section 2, uncertainty in time-series, this section extends the notion to spatio-temporal series.

After analyzing Figure 1, it is possible to observe that the approaches discussed in section 2 are not sufficient for the uncertainty computation for this type of problem. The main cause is that the uncertainty is related to the density values in space and time.

Based on the above ideas it is possible to outline a new model of uncertain series, as follows:

Definition 2 (Spatio-Temporal Series - STS): a spatio-temporal series G of length n consists of a sequence of values v={v_1, v_2,…, v_n,}, a spatial coordinate s(x,y,z), such that G.s indicates that a series G is in a position *s* of space and each value v_i is at timestamp i.

Definition 3 (Uncertain Spatio-Temporal Series - USTS): USTS is a spatio-temporal series such that for each time instant *t* and spatial position (x,y,z), multiple values $v_{i,j}$ of *v* exist. Thus v={($v_{1,1}$, $v_{1,2}$,.., $v_{1,m}$), ($v_{2,1}$, $v_{2,2}$,.., $v_{2,m}$),…,($v_{m,1}$, $v_{m,2}$,.., $v_{m,n}$)}.

Definition 4: Uncertain Spatio-Temporal Series Dataset: is a set *D* of uncertain spatio-temporal series (USTS), i.e., *D*={st_1, st_2,…, st_k}

The calculation of the uncertainty in these kinds of series helps reducing the load on the query processing spatio-temporal database, since it can be grouped by similarity. Other data exploration processes, such as clustering and classification with high volumes of data, can also take benefit from this approach.

4. Managing Uncertainty with UpsilonDB

UpsilonDB [Gonçalves and Porto 2014] is a system designed to manage the uncertainty in numerical simulation data. The system adopts the U-relation Model [Lyublena, 2008] in which relations may represent facts that are uncertain, due to possible alternative interpretations. The different alternatives (or hypothesis) are represented by a random variable that assumes a probability associated to each alternative.

Typically, numerical simulations are based on imprecise mathematical models. Moreover, a numerical simulation execution receives as input a set of parameter values that specifies the initial simulation state and border conditions. The chosen parameter values approximate the model to reality, introducing a new uncertainty factor on the model predictions. UpsilonDB computes the uncertainty of predictions considering the model and the parameter uncertainties, and stores simulation output and predictions using the U-Relational model.

4.1 Methodology for encoding hypothesis

To understand how hypotheses are encoded, an example from the original article regarding UpsilonDB [Gonçalves and Porto 2014] is analyzed. When studying the free fall of an object, the phenomenon φ can be expressed by three models or different laws. In this case, there are three possible hypotheses:

H_1 ⟶ Law of free fall

H_2 ⟶ Stokes´ law

H_3 ⟶ Velocity-squared law

$$H_1 \begin{cases} a(t) = -g^2 \\ v(t) = -\mathrm{gt} + \mathrm{v}_0 \\ s(t) = -\left(\frac{g}{2}\right) t^2 + \mathrm{v}_0\mathrm{t} + \mathrm{s}_0 \end{cases}$$

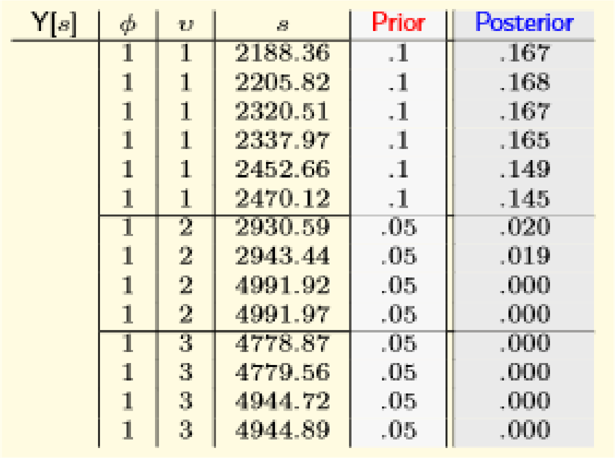
The fact of having three possible explanations for the same phenomenon suggests that there is some uncertainty associated with each hypothesis. To explain this level of uncertainty, UpsilonDB extracts the scheme of functional dependencies (FD) from the mathematical structure of each model. Thus, for H1 the following FDs are extracted:

$$FDs= \{\varphi \longrightarrow g, v_{0,} s_0 ; \quad g, \mu \longrightarrow a; \quad g, v_0, t, \mu \longrightarrow v; \quad g, v_0, s_0, t, \mu \longrightarrow s\}$$

A way to precisely identify H1's formulation is needed, i.e., a data representation of a scientific hypothesis. This is achieved by introducing hypothesis ***id*** υ as a special attribute in the FD. On the other hand, for the set of FDs a global key is calculated; this key represents the set of parameters describing the phenomenon φ.

To reflect the uncertainty present in the parameters describing a given phenomenon, the **id φ** is introduced, and hence, the same phenomenon can have different values for input parameters. Through FDs, the hypotheses **id** and the phenomenon φ may also generate the database schema in UpsilonDB automatically.

A set of simulations is executed to calculate the uncertainty associated with each parameter. The initial data uncertainty is calculated by the frequency of the initial values. Subsequent uncertainty values are obtained through the propagation of the initial uncertainty into the predictive data as shown in Figure 4.



Y[s]	φ	υ	s	Prior	Posterior
	1	1	2188.36	.1	.167
	1	1	2205.82	.1	.168
	1	1	2320.51	.1	.167
	1	1	2337.97	.1	.165
	1	1	2452.66	.1	.149
	1	1	2470.12	.1	.145
	1	2	2930.59	.05	.020
	1	2	2943.44	.05	.019
	1	2	4991.92	.05	.000
	1	2	4991.97	.05	.000
	1	3	4778.87	.05	.000
	1	3	4779.56	.05	.000
	1	3	4944.72	.05	.000
	1	3	4944.89	.05	.000

Figure 3: Table generated with UpsilonDB [Gonçalves and Porto 2014].

Finally, a probabilistic database is obtained, together with all the competing predictions, as possible alternatives, which are mutually inconsistent.

4.2 Modeling Uncertain Spatio-Temporal Series in UpsilonDB

Spatio-temporal series represent a subset of numerical simulations, those in which predictive variables depend only on their spatial-time position and constants, provided as parameters.

Thus, in line with UpsilonDB model, let us consider ϕ (phi) as the phenomenon a spatio-temporal series represents, and Y the set of alternative spatio-temporal series for a given phenomenon. Consider yet, *S* as the set of n-dimensional points in space and *t* a set of timestamp values. Furthermore, let us consider X a variable, whose values in time are taken as a series. Finally, let us assume a variable P=[0,1] representing the uncertainty on each spatio-temporal value of X. Then, in line with UpsilonDB uncertainty introduction approach, we can adapt the definition of Uncertain Spatio-Temporal Series in *Definition 3* by Y_1(Upsilon, S, t, X, P).

Now the question is how can one produce Y_1 having USTS ? The procedure in Figure 3 discusses a method for producing UpsilonDB relations from USTS.

In Figure 3, *Y*_1 represents an uncertain relation for the spatial-time series representing the phenomenon phi=1. The probability distribution for each X value corresponds to its frequency in a spatial-time position. By adopting the U-Relational model, analytical queries may now be submitted to UpsilonDB exploring the uncertainty spatio-temporal characteristics present in the model, such as: (i) giving all the uncertain values collected for a specific phenomenon ϕ, which is the most predicted value considering the position s at time t?; (ii) group all the observations that have similar behavior at a certain position s in a giving period of time; (iii) Select the regions and time periods where the uncertainty prediction is below a specific threshold.

1. Build a certain USTS relation :

- USTS (phi,upsilon, S, t , X), with key (phi,upsilon, S, t)

2. Repair the key so that we have a U-Relation:

- Create Table *Y*_1 as select U.upsilon,U.S, U.t, U.X, U.ro from (repair key upsilon in (select upsilon, S, t, X, count(*) as ro from USTS where phi=1 group by upsilon, S,t) weight by ro) as U

Figure 3. Algorithm for producing UpsilonDB relations

In the meteorological domain, for example, queries such as the ones mentioned above are very important for the analysis of results, since they allow scientists to predict situations where time or spatial coordinates can be fixed. The spatio-temporal dynamic analysis is very complex and, for this reason, it requires that at least one component should be fixed. Consider, for example, some phenomena studied with respect to time. In this situation, if space components are fixed, it would be possible, for example, from the snowfall measurements of any given region, to predict snowfall measurements for other regions with similar spatial conditions in certain seasons. Moreover, for phenomena that depend on the spatial components, if the temporal component is fixed, it would be possible to predict that the temperature varies when the sea rises or falls with the tide. The development of strategies for this type of analysis is vital for prediction of experiment results.

5 Conclusion

Due to the increasing volume of data produced and processed by the most diversified types of sensors, the need to manage uncertain data gained special attention from the scientific community [Dan Suciu, 2011]. This paper gives a step forward in this direction. It formalizes the uncertain spatio-temporal series to address problems where spatial and temporal location influence experimental results, such as the ones observed in natural phenomena. Additionally, this work proposes a computational strategy to manage uncertainty in probabilistic database with the structure of this uncertain series, through which a set of probable values for each spatial and temporal component can be analyzed, and its uncertainty degree can be quantified, by submitting analytical queries to this database. The next step of this work, currently in development, is to implement the algorithm defined in Figure 3 and process analytical queries applied to real phenomena.

References

Antova L., Jansen T., Koch C., Olteanu D., Fast and Simple Relational Processing of Uncertain Data, ICDE, 2008.

Aßfalg J., Kriegel H., Kroger P., Renz M.. 2009. Probabilistic Similarity Search for Uncertain Time Series. *SSDBM.* 2009.

Berndt D. J, Clifford J.. 1994. Using Dynamic Time Warping to Find Patterns in Time Series. *KDD´94.* 1994, pp. 359-369.

Dallachiesa M., Nushi B., Mirylenka K., Palpanas T.. 2012. Uncertain TimeSeries: Return to the Basics. *VLDB.* 2012, Vol. 5, 11.

Dufek, A. S., 2015. *Aplicação da Computação Evolutiva na Previsão Quantitativa de Chuva por Conjunto. PhD Thesis, LNCC*, Petrópolis, Rio de Janeiro, 2015.

Faloutsos, C.. 1994. Fast Subsequence Matching in Time-Series Databases. *ACM.* 1994, 0-89791-839-5/94/0005.

Gonçalves, B., Porto, F., 2014, Upsilon-DB: Managing Scientific Hypothesis as uncertain data, PVLDB, 7 (11), 956-962, 2014

Magnani M., Montesi D. 2005. Uncertainty in data integration: current approaches and open problems. 2005, doi:10.1.1.95.9931.

Orang M., Shiri N. 2014. *An Experimental Evaluation of Similarity Measures for Uncertain Time Series.* Porto, Portugal : ACM, 2014. IDEAS’14. 10.1145/2628194.2628207.

Sarangi S R., Murthy K. 2010. DUST: A Generalized Notion of Similarity between Uncertain Time Series. *ACM.* KDD´10, 2010.

Sentz K., Ferson S. 2002. Combination of Evidence in Dempster-Shafer Theory. *SANDIA REPORT.* 2002, April 2002.

Suciu D., Olteanu D., Ré C., Koch C. 2011. *Probabilistic Databases.* 2011. Vol. 3. doi:10.2200/S00362ED1V01Y201105DTM016.

Wang Y., Li X., Li X., Wang Y. 2013. A survey of queries over uncertain data. *Springer.* April, 2013, 10.1007/s10115-013-0638-6.

Yeh M., Wu K., Yu P. S.. 2009. PROUD: A Probabilistic Approach to Processing Similarity Queries over Uncertain Data Streams. *ACM.* 2009.

Yi B., Faloutsos C.. 2000. Fast Time Sequence Indexing for Arbitrary Lp Norms. *Proceedings of the 26th VLDB Conference.* 2000.