Uncertain Location based Range Aggregates in a multi-dimensional space

Ying Zhang [#], Xuemin Lin [#], Yufei Tao [*], Wenjie Zhang [#]
[#] *The University Of New South Wales*
{yingz,lxue,zhangw}@cse.unsw.edu.au

[*] *Chinese University of Hong Kong*
taoyf@cse.cuhk.edu.hk

***Abstract*— Uncertain data are inherent in many applications such as environmental surveillance and quantitative economics research. Recently, considerable research efforts have been put into the field of analysing uncertain data. In this paper, we study the problem of processing the uncertain location based range aggregate in a multi-dimensional space. We first formally introduce the problem, then propose a general *filtering-and-verification* framework to solve the problem. Two filtering techniques, named *STF* and *PCR* respectively, are proposed to signficantly reduce the verification cost.**

I. INTRODUCTION

Uncertain data are inherent in many applications such as environmental surveillance, market analysis, information extraction, moving object management and quantitative economics research. The uncertain data in those applications are generally caused by data randomness and incompleteness, limitation of measuring equipment, delayed data updates, etc. With the rapid development of various optical, infrared and radar sensors and GPS techniques, there is a huge amount of uncertain data collected and accumulated everyday. So how to efficiently analysing large collections of uncertain data becomes a great concern in many areas [1], [2]. An important operation in those applications is the range query. Although the studies of the range query on spatial database has a long history, it is until very recently that the community starts to investigate this problem against the uncertain data [3], [4], [5], [6], [7]. There are many applications for the range query operation against uncertain data. In this paper, we focus on the distance based range aggregates computation where the location of the query point is uncertain while the target data are conventional points (i.e, certain points). In general, an "uncertain location based query", denoted by Q, is a multi-dimensional point whose location might appear at any location x within a region denoted by $Q.region$, subject to a probabilistic density function $pdf(x)$. Then for given set of data points P and query distance γ, we want to retrieve the aggregate information from the data points which are within distance γ to Q with probability at least θ.

There are many applications for the problem we studied in the paper. One application is to estimate the extent of damage a missile attack might cause [8]. As we know, even the most advanced laser-guided missile can not exactly hit the aim point with 100 percent guarantee. So the commander can not simply predicate the effect of the missile attack by issuing a conventional distance based range aggregate query centred at the aim point to count the number of military targets (e.g., buildings, missile wells, mines, parked fighters) being covered. Instead, it is more reasonable to consider the likelihood of being destroyed for each target points. The distribution of the falling point of various missiles has been extensively studied and different probability density functions (PDFs) are proposed, and *bivariate normal* distribution is the simplest one [8].Therefore, the commander can predicate the effect of attack by counting the number of target points which might be destroyed with likelihood at least θ, which may depend on the confidence level of the commander. Moreover, suppose there are different military values for the target points, the evaluation can be based on the *sum* of the values for those target points.

A straightforward solution of this problem is to compute the appearance probability [1] of each points $p \in P$ for Q. Then do the aggregate computation on the points which appear in Q with probability at least θ. Usually the appearance probability computation is expensive because it requires costly numerical evaluation of a complex integral. So the key of the problem is how to efficiently disqualify a point p or validate it as an result based on some pre-computed information. That is, we need to filter as many data points as possible to reduce the number of appearance probability computations.

In the paper, we first propose a *filtering-and-verification* framework to solve the problem based on filtering technique. Then we propose a distance based filtering techniques, named *STF*. The basic idea of the *STF* technique is to bound the appearance probability of the points by applying some well known statistical inequalities where only a few statistics about the uncertain location based query Q is required. The *STF* technique is simple and space efficient (only $d + 2$ float numbers required), and experiments show that it has a decent filtering capacity. We also investigate how to apply existing probabilistically constrained regions (*PCR*) technique [5] to our problem.

The remainder of the paper is organized as follows. In Section II, we formally define the problem. Section III proposed a general *filtering-and-verification* framework and two filtering techniques. Section IV evaluates the proposed techniques with experiments. Then Section V concludes the paper.

II. PROBLEM DEFINITION

A data point $p \in P$ or query instance $q \in Q$ referred in this paper, by default, is in a d-dimensional numerical space. The distance between two points x and y is denoted by $|x-y|$. An object o in the paper has arbitrary shape which might enclose a set of data points[2], $|o_1 - o_2|_{min}$ denotes the $min(\{|x_i - y_j|\})$ for $\forall x_i \in o_1$ and $\forall y_j \in o_2$; Similar definition goes to $|o_1 - o_2|_{max}$. Note that the *Euclidean distance* is employed as the distance metric in the paper. Nevertheless, our technique can be easily extended to other distance metrics as long as

[1]For presentation simplicity, we say a point p appears with respect to query point q if the distance between query point q and p is not larger than γ

[2]object o is corresponding to the MBR of an entry in R tree in the paper

the triangle inequality holds. For presentation simplicity, we use "uncertain query" to represent "uncertain location based query". Following is the definition of uncertain query Q on both *continuous* and *discrete* case.

Definition 1 (Uncertain Query Q (continuous)): Uncertain query Q is described by a probabilistic density function $Q.pdf$. Let $Q.region$ present the region where Q might appear, then $\int_{x \in Q.region} Q.pdf(x)dx = 1$;

Definition 2 (Uncertain Query Q (discrete)): The uncertain query Q consists of a set of instances $\{q_1, q_2, \ldots, q_n\}$ where q_i appears with probability P_{q_i} and $\sum_{q \in Q} P_q = 1$;

For a point p, we use $P_{app}(Q, p, \gamma)$ to represent the probability of point p located within distance of γ towards uncertain query Q. It is called the appearance probability of p regarding uncertain query Q and query distance γ for presentation simplicity. Following is the formal definition of the appearance probability of p under the *continuous* and *discrete* cases respectively. For the *continuous* case,

$$P_{app}(Q, p, \gamma) = \int_{x \in Q.region \ \wedge \ |x-p| \leq \gamma} Q.pdf(x)dx \qquad (1)$$

As to the discrete case,

$$P_{app}(Q, p, \gamma) = \sum_{q \in Q} P_q, \quad where \ |q - p| \leq \gamma. \qquad (2)$$

Specially, we have $P_{app}(Q, p, \gamma) = 0$ for any $\gamma < 0$; Note that when there is no ambiguity, we use $P_{app}(p, \gamma)$ to replace $P_{app}(Q, p, \gamma)$. And Q and pdf are employed to represent $Q.region$ and $Q.pdf$ respectively. It is immediate that $P_{app}(p, \gamma)$ is a monotonic function with respect to distance γ.

Problem Statement.
In this paper we investigate the problem of uncertain location based range aggregate query on spatial data; it is formally defined below.

Definition 3 (Uncertain Range Aggregate Query): Given a set of points P, an uncertain query Q, query distance γ and probabilistic threshold θ, we want to compute the aggregate information against points $p \in Q_{\theta,\gamma}(P)$, where $Q_{\theta,\gamma}(P)$ denotes the set of points $p \in P$ and $P_{app}(p, \gamma) \geq \theta$.
In the paper, we employ the *count* as the aggregate operation. That is, we want to efficiently compute $|Q_{\theta,\gamma}(P)|$. Nevertheless, our technique can be easily extended to other aggregates (e.g., *sum*, *avg*, *max* and *min*).

III. *Filtering-and-Verification* ALGORITHM

In this section, we first introduce a general framework for the *filtering-and-verification* Algorithm based on filtering techniques. Then we proposes a simple statistical filtering technique. We also investigate how to apply the *PCR* technique [5] to tackle our problem.

A. A framework for filtering-and-verification Algorithm

For the given uncertain query Q, probability threshold θ and distance γ, the naive way is to compute the $P_{app}(p, \gamma)$ for each data point $p \in P$ based on Equation 1 , then count the number of points with $P_{app}(p, \gamma) \geq \theta$. Clearly, it is inefficient as we need to visit every point $p \in P$ and the integral computation is expensive. To reduce the number of verifications, it is desirable to apply filtering technique to prune or validate the data points. Suppose P is organized by aggregate R tree R_P and a filter F on Q is available, Algorithm 1 describes how to apply a filter for the aggregate query processing in a *branch-and-bound* fashion. Note that the filter should support the intermediate entry so that a group of data points can be filtered at same time.

Algorithm 1: ***Filtering-and-Verification*(R_P, Q, F, γ, θ)**

Input : R_P : an aggregate R tree on data set P,
Q : uncertain query, F : Filter,
γ : query distance, θ : Probabilistic threshold
Output : $|Q_{\theta,\gamma}(P)|$

```
1  Stack := ∅; S := 0; C := ∅;
2  insert root of R_P into Stack;
3  while Stack ≠ ∅ do                          /* filtering */
4      Remove top entry e from Stack ;
5      Load entry e from disk ;
6      for each child entry e_i of e do
7          status := F.check( e_i );
8          switch status do
9              case pruned do
10                 break ;
11             case validated do
12                 S := S + |e_i|; break;
13             case unknown do
14                 if e_i is data entry then
15                     C := C ∪ e_i;
16                 else
17                     put e_i into Stack;
18                 break;
19 for data entry e in C do                    /* verification */
20     if P_app(Q, e, γ) ≥ θ then
21         S := S + 1 ;
22 return S
```

An immediate filtering technique is based on the distance between the entry and uncertain query. Clearly, for any θ we can safely prune an entry with $|Q - e|_{min} > \gamma$ or validate it if $|Q - e|_{max} \leq \gamma$. We refer this as maximal/minimal distance based filtering technique, named *MMD*. *MMD* technique is time efficient as only $O(d)$ time is required to compute the minimal and maximal distance between $Q.MBR$ and $e.MBR$, where $Q.MBR$ is the minimal bounding rectangle of Q. However, the *MMD* technique does not make use of θ, which inherently limits its filtering capacity. In the sequel, we introduce two filtering techniques which can enhance the filtering capacity with some pre-computed information.

B. Statistical Filter

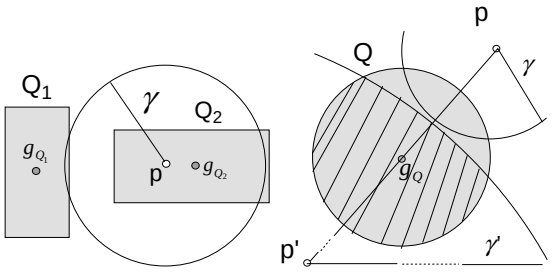


Fig. 1. Motivation Example Fig. 2. Proof of Upper bound

In this subsection, we propose a statistical filtering technique, named *STF*. As shown in Figure 1, for given $\theta = 0.5$ we can not prune p for uncertain query Q_1 based on *MMD* technique although intuitively p should be pruned. Similarly, we can not validate p for Q_2 either. This motivate us to develop a new filtering technique which is as simple as *MMD*, but can exploit θ to enhance the filtering capacity.

We first introduce three statistics and a lemma which are employed by *STF* technique.

Definition 4 (Geometric Centroid (g_Q)): Informally, geometric centroid is the "average" of all points of an object. Let g_Q denote the geometric centroid of uncertain query Q, we have $g_Q = \int_{x \in Q} x \times pdf(x)dx$ and $g_Q = \sum_{q \in Q} q \times P_q$ for *continuous* and *discrete* case respectively.

Base on the g_Q, we have two definitions, named η_Q and σ_Q respectively, which describe the variance of the distribution of uncertain query Q. η_Q represents the weighted average distance to g_Q with $\eta_Q = \int_{x \in Q} |x - g_Q| \times pdf(x)dx$ and $\sum_{q \in Q} |q - g_Q| \times P_q$ for *continuous* and *discrete* case respectively. Similarly, σ_Q denotes the variance of Q with $\sigma_Q = \int_{x \in Q} |x - g_Q|^2 \times pdf(x)dx$ and $\sum_{q \in Q} |q - g_Q|^2 \times P_q$ for *continuous* and *discrete* case respectively.

The *Cantelli's inequality*[9] described by Lemma 1 is employed in our statistical filtering technique and it is one-sided version of the *Chebyshev inequality*.

Lemma 1: Let X be a random variable with expected value μ and finite variance σ^2. Then for any real number $k > 0$, $Pr(X - \mu \geq k\sigma) \leq \frac{1}{1+k^2}$.

Then Theorem 1 indicates that we can further enhance the filtering capacity based on some simple statistics of Q.

Theorem 1: For the uncertain query Q and distance γ, suppose the geometric mean g_Q, weighted average distance η_Q and variance σ_Q for Q are available. Then for a given point p, we have

1) If $\gamma > \mu_1$, $P_{app}(p, \gamma) \geq 1 - \frac{1}{1+\frac{(\gamma-\mu_1)^2}{\sigma_1^2}}$, where $\mu_1 = |g_Q - p| + \eta_Q$ and $\sigma_1^2 = \sigma_Q - \eta_Q^2 + 4\eta_Q \times |g_Q - p|$.
2) If $\gamma < |g_Q - p| - \eta_Q - \epsilon$, $P_{app}(p, \gamma) \leq \frac{1}{1+\frac{(\gamma'-\mu_2)^2}{\sigma_2^2}}$, where $\mu_2 = \Delta + \eta_Q$, $\sigma_2^2 = \sigma_Q - \eta_Q^2 + 4\eta_Q \times \Delta$, $\Delta = \gamma + \gamma' + \epsilon - |p - g_Q|$ and $\gamma' > 0$. The ϵ represents an infinitely small positive value.

Proof: As uncertain query Q can be regarded as a random variable which takes $x \in Q$ with probability $pdf(x)$, we construct another random variable Y as follows: for $\forall x \in Q$, there is a $y \in Y$ such that $y = |x - p|$ and $Y.pdf(y) = Q.pdf(x)$. Then we have $P_{app}(p, \gamma) = Pr(Y \leq \gamma)$ according to the Equation 1. Based on triangle inequality, we have $|x-p| \leq |x-g_Q|+|x-g_Q|$ and $|x-p| \geq |\,|x-g_Q|-|p-g_Q|\,|$ for any $x \in Q$. Let μ and σ denote the *expection* and *standard deviation* of random variable Y respectively, then

$$\begin{aligned} \mu &= \int_{y \in Y} y \times Y.pdf(y)dy = \int_{x \in Q} |x - p| \times pdf(x)dx \\ &\leq |g_Q - p| + \eta_Q = \mu_1 \end{aligned}$$

Similarly, we have $\mu \geq |g_Q - p| - \eta_Q$.

$$\begin{aligned} \sigma^2 &= E(Y^2) - E^2(Y) \\ &\leq \int_{x \in Q} (|g_Q - p| + |x - g_Q|)^2 pdf(x)dx \\ &\quad -(|g_Q - p| - \eta_Q)^2 \\ &= \sigma_Q - \eta_Q^2 + 4\eta_Q \times |g_Q - p| = \sigma_1^2 \end{aligned}$$

Then based on lemma 1 let $k = \frac{\gamma-\mu}{\sigma}$, if $\gamma > \mu_1$ we have

$$\begin{aligned} Pr(Y \geq \gamma) &= Pr(Y - \mu \geq k\sigma) \leq \frac{1}{1 + (\frac{\gamma-\mu}{\sigma})^2} \\ &\leq \frac{1}{1 + (\frac{\gamma-\mu_1}{\sigma_1^2})^2} \end{aligned}$$

Then it is immediate that

$$Pr(Y \leq \gamma) \geq 1 - Pr(Y \geq \gamma) \geq 1 - \frac{1}{1 + \frac{(\gamma-\mu_1)^2}{\sigma_1^2}} \quad (3)$$

As to the upper bound, as illustrated in Figure 2 let p' be the dummy point on the line $\overline{pg_Q}$ with $|p' - p| = \gamma + \gamma' + \epsilon$. Let $\Delta = |p' - g_Q|$, then we have

$$\Delta = \gamma + \gamma' + \epsilon - |p - g_Q| \quad (4)$$

According to Inequality 3 and Equation 4, when $\gamma < |p - g_Q| - \eta_Q - \epsilon$ we have $P_{app}(p', \gamma') \geq 1 - \frac{1}{1+\frac{(\gamma'-\mu_2)^2}{\sigma_2^2}}$, since for any $x \in Q$ and $|x - p'| \leq \gamma'$ (stripped area in Figure 2), $|x - p| > \gamma$. It implies that $P_{app}(p, \gamma) \leq 1 - P_{app}(p', \gamma') \leq \frac{1}{1+\frac{(\gamma'-\mu_2)^2}{\sigma_2^2}}$.

■

Following extension is immediate based on the rationale of Theorem 1.

Extension 1**.** Suppose o is an object with arbitrary shape, we can simply use $|o - g_Q|_{min}$ and $|o - g_Q|_{max}$ to replace $|p - g_Q|$ in Theorem 1 for lower and upper probabilistic bounds computation respectively.

Based on the Extension 1, we can compute the upper and lower bound for $P_{app}(e, \gamma)$ to prune or validate entries in Algorithm 1. Since g_Q, η_Q and σ_Q are pre-computed, the only dominate cost in filtering phase is distance computation between e and p which only costs $O(d)$ time.

Following the similar rationale, another statistical filter can be proposed based the popular statistical inequality *Markov's inequality*. We omit this part due to the space limitation.

C. PCR based Filter

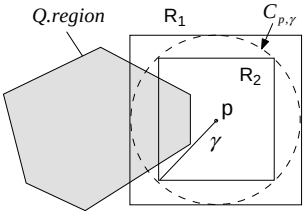


Fig. 3. Transform query

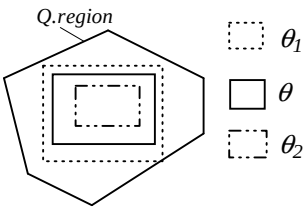


Fig. 4. Choose PCRs

Although Tao *et al.*[5], [6] do not address the problem studied in this paper, the *PCR* technique can be employed

as filter in Algorithm 1. In Figure 3, let $C_{p,\gamma}$ represent the circle(sphere) centred at p with radius γ. Then we can regard the uncertain query Q as an uncertain object, while $C_{p,\gamma}$ serves as a query. Because *PCR* technique only works for rectangle query, as suggested in [6], we can use R_1 and R_2 in Figure 3 to represent $C_{p,\gamma}$ for pruning and validation purpose respectively. Similar transformation can be done for intermediate entries as well. Suppose a finite set of PCRs are pre-computed. For given θ which is not selected for pre-computation, we can carefully choose two closest existing PCRs for pruning and validation as illustrated in Figure 4. Since the θ is fixed during the query, we can choose these two PCRs for all data points before processing of the query. Then pruning/validating rules in [5] can be applied for pruning and validate which are very time efficient - only $O(d)$ time required for each entry test. In order to further improve the performance of the filter, more sophisticate approach from [6] is applied in our implementation. And the worst filtering time for each entry is $O(m + d \log m)$ where m is the number of $PCRs$.

IV. Experiment

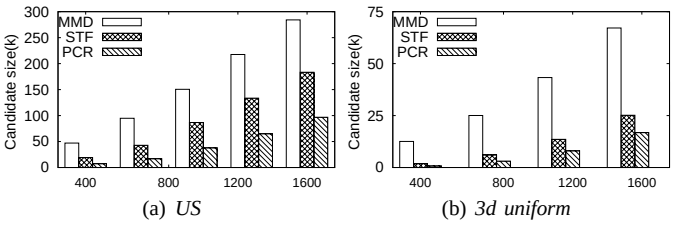
We present results of a comprehensive performance study to evaluate the efficiency of proposed techniques in the paper. Following the frame work of Algorithm 1, three different filtering techniques (*MMD*, *STF* and *PCR*) have been implemented and evaluated. All algorithms are implemented in C++ and compiled by GNU GCC. Experiments are conducted on PCs with Intel P4 2.8GZ CPU and 2G memory under Debian Linux.

The spatial dataset, *US*, is employed as target dataset which contains 1m 2-dimensional points representing locations in the United State[3]. All of the dimensions are normalised to domain $[0, 10000]$. To evaluate the performance of the algorithms , we also generate synthetic dataset *Uniform* with 3 dimension, in which points are uniformly distributed. The domain size is $[0, 10000]$ for each dimensions. All of the datasets are organized by aggregate R trees with pagesize 4096 bytes. A workload consists of 200 uncertain queries in our experiment. And the uncertain region of the uncertain queries in our experiment are circles or spheres with radius q_r which varies from 200 to 1000. The centres of the queries are randomly generated within the domain and *Normal* distribution is employed to describe the PDF of the uncertain queries. Moreover, in order to avoid favouring particular θ value, we randomly choose the probabilistic threshold between 0 and 1 for each uncertain query.

We measure the performance of the techniques by means of IO cost and candidate size during the computation. The IO cost is the number of pages visited from R_P. While candidate size is the number of data points which need exact probabilistic computation.

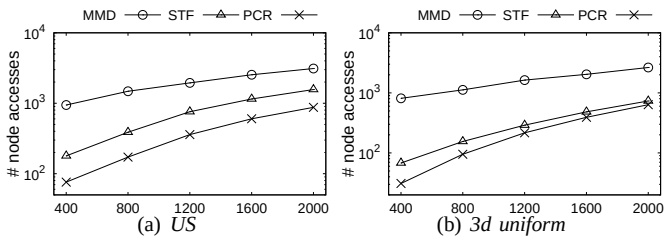
In the experiments, we evaluate the impact of query distance γ on the performance of the filtering techniques in terms of candidate size and IO cost against *US* and 3d *Uniform* datasets. Figure 5 reports the candidate size of *MMD*, *STF* and *PCR* when query distance γ grows from 400 to 2000. Clearly, the large γ results in more candidate data points for verification. It is interesting that with only a few statistics, the *STF* can achieve a great saving on the candidate size compared with the *MMD*. With more space, *PCR* can further reduce the candidate size.

[3]Available at http://www.census.gov/geo/www/tiger/



(a) *US* (b) *3d uniform*

Fig. 5. Candidate Size vs γ



(a) *US* (b) *3d uniform*

Fig. 6. # node accesses vs γ

We evaluate the IO cost of the techniques and report the results in Figure 6. As expected, *PCR* still ranks first on both datasets.

V. Conclusions

In this paper, we formally define the problem of uncertain location based range aggregates in a multi-dimensional space; it covers a wide spectrum of applications. To efficiently process such a query, we propose a general *filtering-and-verification* framework and two filtering technique, named *STF* and *PCR* respectively, such that the expensive computation cost for verification can be significantly reduced. As demonstrated in the experiment, *STF* filtering technique can achieve a decent filtering capacity based on a few pre-computed statistics about the uncertain location based query. Moreover, it is very fast and space efficient due to its simplicity. And *PCR* technique is quite efficient when more space is available.

Acknowledgement. The work was supported by ARC Grant (DP0881035 and DP0666428) and Google Research Award. And the third author was supported by Grant CUHK 4161/07 from HKRGC.

References

[1] N. N. Dalvi and D. Suciu, "Management of probabilistic data: foundations and challenges," in *PODS*, 2007.
[2] M. A. Soliman, I. F. Ilyas, and K. C.-C. Chang, "Top-k query processing in uncertain databases," in *ICDE*, 2007.
[3] R. Cheng, D. V. Kalashnikov, and S. Prabhakar, "Evaluating probabilistic queries over imprecise data." in *SIGMOD 2003*.
[4] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. S. Vitter, "Effcient indexing methods for probabilistic threshold queries over uncertain data." in *VLDB 2004*.
[5] Y. Tao, R. Cheng, X. Xiao, W. K. Ngai, B. Kao, and S. Prabhakar, "Indexing multi-dimensional uncertain data with arbitrary probability density functions." in *VLDB*, 2005.
[6] Y. Tao, X. Xiao, and R. Cheng, "Range search on multidimensional uncertain data," *ACM Trans. Database Syst.*, vol. 32, no. 3, 2007.
[7] J. Chen and R. Cheng, "Efficient evaluation of imprecise location-dependent queries," in *ICDE*, 2007.
[8] G. M. Siouris, *Missile Guidance and Control Systems*, 2004.
[9] R. Meester, *A Natural Introduction to Probability Theory*, 2004.