Brazilian Journal of Cartography (2016), Nº 68/6, Special Issue GEOINFO 2015: 1087-1096
Brazilian Society of Cartography, Geodesy, Photgrammetry and Remote Sense
ISSN: 1808-0936

GEOINFO
Brazilian Symposium on GeoInformatics
November 29th to December 2nd, 2015 Campos do Jordão, SP, Brazil

A METHOD FOR LOCATION RECOMMENDATION VIA SKYLINE QUERY TOLERANT TO NOISED GEOREFERENCED DATA

Um Método para Recomendação de Localidade via Consulta Skyline Robusta à Imprecisão nas Coordenadas Georreferenciadas

Welder Batista de Oliveira[1], Sávio Salvarino Teles de Oliveira[2], Vagner José do Sacramento Rodrigues[2], Helton Saulo Bezerra dos Santos[3] & Kleber Vieira Cardoso[1]

[1]Universidade Federal de Goiás – UFG
Instituto de Informática – INF
Alameda Palmeiras, Quadra D, Campus Samambaia 131 - CEP 74001-970 – Goiânia – GO – Brazil
weldermat@gmail.com, kleber@inf.ufg.br

[2]GEOMAIS Serviços de Informática LTDA - ME
Rua Leopoldo Bulhões, esquina com a Rua 1014, Quadra 31, Lote 07, Sala 9 Setor Pedro Ludovico
CEP 74820-270 – Goiânia – GO – Brazil
savio.teles@gogeo.io, vagner@gogeo.io

[3]Universidade Federal de Goiás – UFG
Instituto de Matemática – UFG
Alameda Palmeiras, Quadra D, Campus Samambaia 131 - CEP 74001-970 – Goiânia – GO – Brazil
heltonsaulo@gmail.com

Received on December 5, 2015/ Accepted on February 28, 2016
Recebido em 5 de Dezembro, 2015/ Aceito em 28 de Fevereiro, 2016

ABSTRACT

This paper presents a method to perform a location recommendation based on multiple criteria allowing noised coordinates. More specifically, the skyline query is adapted to handle those noises by modeling the errors of georeferenced points with an appropriate probability distribution and modifying the traditional dominance criterion used by that technique. The method is applied to a scenario in which the coordinates are set by a geocoding process in a sample of schools in a specific Brazilian city. It enables one to choose the level of confidence in which a point is removed from the skyline solution (the location recommendation).

Keywords: Skyline, Probabilistic, Modeling Error, Geocoding.

RESUMO

Este artigo apresenta um método para recomendação de localidade baseado em múltiplos critérios na presença de coordenadas imprecisas. Mais especificamente, a consulta skyline é adaptada para lidar com estes ruídos modelando os erros dos pontos georreferenciados com uma distribuição de probabilidade apropriada e modificando o critério de dominância tradicional usado por esta técnica. O método é aplicado ao cenário em que as coordenadas são obtidas por um processo de geocodificação de uma amostra de escolas de uma cidade brasileira. O método permite a escolha do nível de confiança em que o ponto é removido da skyline (a recomendação de localidade).

Keywords: Skyline, Probabilística, Modelagem do Erro, Geocodificação.

1. INTRODUCTION

Database management systems (DBMS) have been increasingly used in recommendation services or applications. Many of these applications are based on multiple, and sometimes conflicting goals, where there may be no single optimal answer. For example, a tourist may be interested in budget hotels with reasonable ratings (say, 3-star) that are close to the city. Traditionally, the DBMS supports the recommendation applications by returning all answers that may meet the user's requirement. This may not useful if the user is overloaded by a large amount of information.

Spatial skyline queries (BORZSONY *et al.*, 2001) have gained attention due to their efficient solution for this issue. These queries retrieve the desirable objects that are no worse than any other object in the database, according to all the criteria under evaluation. In other words, given a set of points, skyline comprises the points that are not dominated by other points. In our example, if there are two hotels, *h1* and *h2*, with the same rating, such that *h1* is both cheaper and nearer to the city than *h2*; then, *h2* would not be would not be presented to the user.

The spatial skyline query is directly impacted by the accuracy of the location provided by the database. Data uncertainty inherently exists in a large number of applications. (DING *et al.*, 2014) due to factors such as limitations of measuring devices, e.g. GPS, and inaccuracy of the geocoding algorithms, when only the address of data (for example, the hotel) is provided. Returning to example, if the hotels *h1* and *h2* had been incorrectly located, thus *h2* could dominate *h1* if *h2* is nearer to city than *h1* because of the location error.

Due to the importance of those applications with uncertain data, this context provides opportunities and demands the creation of automated recommendation services that are tolerant to data inaccuracy. How to perform analysis using inaccuracy locations, especially the skyline analysis, remains an important and challenging problem. In this paper, we present a novel technique to perform skyline queries over inaccurate locations.

The remainder of this paper is organized as follows. In Section 2 we briefly give an overview of the use of verifying approaches for recommendation services. The problem of skyline queries over inaccurate locations is formally defined in Section 3. Section 4 presents our approach of recommendation service that are tolerant to inaccuracy of spatial data. Section 5 presents results and discussion of our approach. Finally, we provide some concluding remarks in Section 6. This paper is based on (OLIVEIRA *et al.*, 2015), previously presented at GEOINFO-2015 conference in *Campos do Jordão-SP*, Brazil (http://www.geoinfo.info/).

2. RELATED WORK

Since the introduction of the skyline operator (BORZSONY *et al.*, 2001), skyline query processing has received considerable attention in multidimensional databases. A number of different algorithms for skyline computation have been proposed. (TAN *et al.*, 2001) use auxiliary structures on progressive skyline computation, (KOSSMANN *et al.*, 2002) show a nearest neighbor algorithm for skyline query processing, (PAPADIAS *et al.*, 2003) introduce the branch and bound skyline (BBS) algorithm, (CHOMICKI *et al.*, 2003) present a sort-filter-skyline (SFS) algorithm leveraging pre-sorting lists, and (GODFREY *et al.*, 2005) propose a linear elimination sort for skyline algorithm with attractive average-case asymptotic complexity.

The problem of computing skylines under noisy information was only investigated recently. Heuristic approaches have also been suggested for skyline computation on incomplete or imprecise data in (PEI *et al.*, 2007, KHALEFA *et al.*, 2008, LOFI *et al.*, 2013). In (GROZ & MILO, 2015) is returned the correct skyline with high probability of minimizing the number of comparisons required for computing or verifying a candidate skyline. None of these works, however, consider spatial relationships between data objects.

The concept of spatial skyline query (SSQ) was introduced in (SHARIFZADEH & SHAHABI, 2006), which given a set of data points *P* and a set of query points *Q*, each data point has a number of derived spatial attributes, and each attribute is the point's distance to a query point.

(YOU *et al.*, 2013) proposed the threshold farthest spatial skyline (TFSS) and branch and bound farthest spatial skyline (BBFS) algorithms. The TFSS algorithm uses a standard set of accesses such as sorted access from distributed sources, which uses *R-tree* for accessing node and retrieves data objects in decreasing order of the attribute value. The BBFS algorithm uses minimum Bounding rectangle (MBR) of an *R-tree* for batch pruning. Full space skyline can be supported incrementally by using naive online maintenance module (NMA) in (HUANG *et al.*, 2010).

Several studies have also focused on the skyline query processing with moving object. (PEI *et al.*, 2007) tackle the problem of skyline analysis on uncertain data and (HUANG *et al.*, 2006) introduce the continuous skyline over precise moving data. (ZHANG *et al.*, 2009) present techniques that enable inference of the current and future uncertain locations efficiently. A novel probabilistic skyline model is proposed in (DING *et al.*, 2014) where an uncertain object may take a probability to be in the skyline at a certain point in time.

The preset paper brings something new to the area, namely the possibility to perform skyline query even when the data is not precise. For precision we mean that the values provided by the data are exactly what can be found in reality, i.e. the data describe the reality with fidelity (with no errors from measurements, estimation or other sources). The hypothesis of precision is assumed by all current skyline procedures present in literature. We intend to change this scenario providing a more elastic approach to this kind of query, especially concerning spatial attributes.

The two closest related works in literature were made by (PEI *et al.*, 2007) and (LOFI *et al.*, 2013). The first deal with what they call "uncertain" data. Here, uncertain possess not the same meaning as the term chosen by us - "imprecision". By uncertain the authors mean that more than one record is available for each attribute for the several objects under evaluation. They provided an example with NBA players data. To each player is collected statistics like number of assists and the number of rebounds, both the larger the better. As players have different performances in different games, the values for the attributes for each player is called to be "uncertain". To solve cases like that, (PEI *et al.*, 2007) provide two algorithms. Moreover, the concept of dominance probability is introduced in this paper. On the other hand, (LOFI *et al.*, 2013) use a method based on crowd based data. This way they can perform skyline query for incomplete data. None of them, however, deal with "imprecise" data, i.e., data which contain values with some error.

In our paper, the values of spatial attributes are considered random variables and are modeled with a probability density function. This enables to compute dominance probabilities even for imprecise data and then provide a more noisy tolerant technique.

3. FORMAL PROBLEM DEFINITION

In order to provide a location recommendation tool able to deal with multi-criteria decision analysis, one may face the problem of noised georeferenced data. Furthermore, one important step that such a tool should have is discarding points that are not Pareto efficient. Pareto efficiency is an equivalent expression to skyline query and will be explained later. One problem arises in this context: "how to build a skyline query that minimizes the bad effects caused by imprecise georeferenced data?"

Skyline query is a multi-criteria decision making technique. It aims to find a set *S* of all points that are not dominated by any other in the database under consideration. According to (PAPADIAS *et al.*, 2003), under the *min* condition, a point *pi* dominates another point *pj* if and only if the coordinate of *pi* on any axis is not larger than the corresponding coordinate of *pj*. In this case, the desirable points are those with the smaller values. In this paper, only the min condition is being considered when the term dominance is mentioned. Naturally, the results are analogous for the *max* condition.

The Figure 1 shows the data for a classical example: "choose a hotel both cheap and near the beach". These are the typical conflicting criteria faced at making a decision concerning multiple goals. The nearest hotel may not be the cheaper. Note that the points *a*, *c* and *l* are not dominated by any other, i.e., no other is better in both dimensions then those. Therefore, *S = {a, c, l}* is the skyline query for these points.

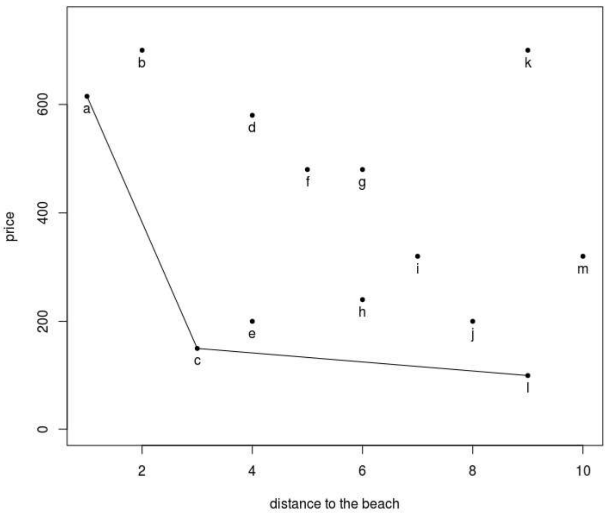

Fig. 1 - Hotels near the beach: a classical skyline query example.

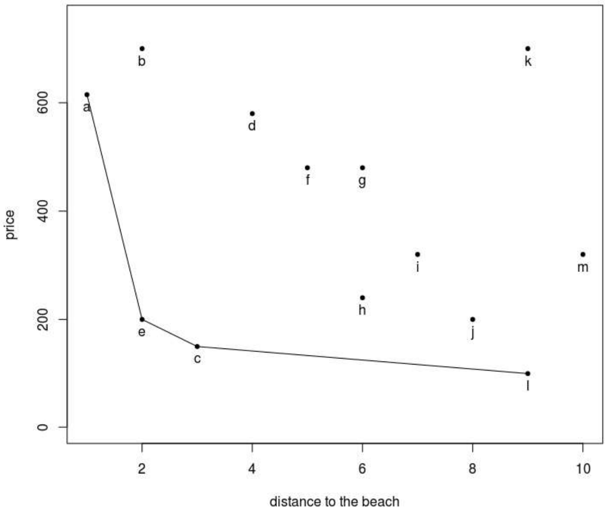

Fig. 2 - Hotels near the beach: with a translation in point e.

However, due to coordinates imprecision, one may not guarantee that point e is in fact dominated by the points in *S*. Figure 2 shows how *S* are changed by a relatively small translation in point *e*. A translation of this size in e may reflect just the imprecision in its georeferenced position. The imprecision is measured by the error given by

$$error = |real - database| \qquad (1)$$

here, *real* is the real position and *database* is the database position. Therefore, a specific point is said to be imprecise if *error > 0.*

Two distinct types of error can occur: I) exclude a point from *S* when it should be in *S*; II) include in *S* a point that should not be there. In this work, we address the type I error by presenting a method to provide a skyline query solution for noised georeferenced data.

4. THE METHOD

In order to solve the problem proposed by this paper, three steps are required:

1. Model the error with a probability density function (PDF).
2. Generate a table of dominance probabilities via Monte Carlo method.
3. Rewrite skyline query by modifying the dominance criterion in order to take a controlled risk of a type I error.

In the following subsections, these steps will be explained in detail.

4.1 Error modeling

The error must be mathematically modeled, this is, it is necessary to find a mathematical function that enables one to accurately predicts the behavior of the error. For instance the probability of the error lies in the interval (200, 300) meters. To find this model, one must search for a PDF. A PDF is a function f that associates a value in the interval (0, 1) to each set A of possible values of a random variable *X*. Some additional properties may be verified as well, but discussing about them is not in the scope of this paper.

With a PDF one can calculate the probability of a value of some random variable lies in a specific range, for example, error be more than 200 and less then 300 meters. In this first step of the method, it is necessary to find a PDF that fits well the curve of the errors. A good way to make a guess of such a function is by the histogram or even the kernel estimation of the curve.

Since one suspects that a particular PDF fits well the curve of the errors, a formal hypothesis test may be applied in order to confirm (or not) the guess. A very well recognized hypothesis test in this sense is the Kolmogorov-Smirnov test. As states (MASSEY JR, 1951),

If F0(x) is the population cumulative distribution, and SN(x) the observed cumulative step-function of a sample (i.e., SN(x) = k/N , where k is the number of observations less than or equal to x), then the sampling distribution of d = maximum |F0(x) − SN(x)| is known, and is *independent of F0(x) if F0(x) is continuous.*

Therefore, the distribution of *d* can be used to perform inference related to the hypothesis that *F0(x)* is the true populational distribution of the errors.

In our work, we employed the software *R* to perform the Kolmogorov-Smirnov test. The hypothesis may be considered false with at least 95% level of confidence if the computed p-value is at most 5%. Usually, a 95% level of confidence is considered good enough in order to discard or confirm the use of a specific probability distribution for most applications.

4.2 Table of dominance probabilities via Monte Carlo method

In this subsection, it will be evaluated the probability of a database point *P* dominates another, say *Q* in order to construct a table of dominance probabilities. This will be done by Monte Carlo method. These two points possess the coordinates *G* and *H* in the dataset, respectively, which may be imprecise (*error* > *0*). Therefore, if one computes the distances between each of those points to a third, say *L*, the one with the minimum distance to *L* in the dataset may be not the one with minimum distance in reality. Let *p* be the probability of *P* be closer than *Q*. In this subsection an estimative for *p* is provided.

The Monte Carlo method aims to estimate a mean $M = E(X)$ by simulations with random numbers, where *X* is a random variable. About its history and applications, one can see (METROPOLIS, 1987). In order to estimate *M* , *n* simulated values of *X* should be performed, say *x1*, *x2*, ..., *xn* . After generating the *n* values for *X*, its average $m = \sum xi / n$ is computed. As *n* increases, the central limit theorem guarantees that m becomes arbitrarily closer to *M*, with the difference going to zero as *n* tends to infinity.

In this paper we define a random variable *Y* and set 1 to it if *P* is closer than *Q* in relation to *L*. Otherwise, the value *Y* is set to *0*. This way, the mean *M* of the *Y* coincides with the probability that *P* is the closest. As the goal is to estimate this probability, the steps (2), (3) and (4) shown above can be performed. However, it is necessary to define how the simulations will be performed.

In the case of estimating the head probability for a given coin, the simulation is simple: just flip the coin and compute the values. To simulate a value for *Y*, one can follow the paths:

1. generate a value *error1* and a value *error2* both from the chosen PDF;
2. generate an angle $\theta 1$ and $\theta 2$ both from an uniform distribution with parameters $(0, 2\pi)$, since it is assumed that the error is equally probable in any direction;
3. add to *G* a vector of length *error1* in the direction $\theta 1$, resulting in the point *G'*. Similarly, and to *H* a vector of length *error2* in the direction of $\theta 2$, resulting in the point *H'* ;
4. compute the distances $d1 = |G' - L|$ and $d2 = |H' - L|$;
5. if $d1 < d2$ then set 1 to *Y* , otherwise set 0;
6. repeat *n* times steps 1 to 5 with a large value of *n* (for instance, *n* = *10, 000*);
7. calculate $m = \Sigma yi / n$ dominates *Q*.

In order to provide a table of dominance probabilities as shown in Table 1, the algorithm showed above must be applied for several pairs of points *P* and *Q* placed at different distances from a reference place *L*. In this table, *pij* means the probability of *P* be closer to *L* than *Q*, such that in the database the distance of *P* to *L* is *100i* and from *Q* to *L* is *100j*. One may construct a more complete table by considering multiples of 1 meter instead of multiples of 100 meters like in this table example. Nevertheless, probabilities for intermediate or even fractional values may be estimated by interpolation.

Table 1: Structure of a dominance table

near / far	200	300	400	500
100	p12	p13	p14	p15
200	p22	p23	p24	p25
300	p32	p33	p34	p35
400	p42	p43	p44	p45
500	p52	p53	p54	p55

With the Table 1, the dominance criterion of skyline query can be modified. Now, it is possible to talk about probability of dominance. Instead consider the set *S* of skyline query points, it is possible to construct the set *Wp* of points which are not dominated with level of confidence *p* by any other point. The new criterion is the following:

"a point x dominates a point y concerning a dimension i with level of confidence p if

$$Prob(xi < yi\,) > p". \qquad (2)$$

Therefore, there is a chance of a type I error of $(1 - p)$.

For several criteria, say d, and assuming independence between the errors of each of these dimensions, the new dominance criterion may be rewritten like: "a point x dominates y with level of confidence p if the product of the probabilities $Prob(xi < yi)$ is greater than p", i.e.,

$$Prob(x1 < y1) \ldots Prob(xd < yd) > p \qquad (3)$$

Thus, the type I error has been controlled as it was the goal of this paper. This new version of skyline query is designed to be more tolerant to geo-referenced imprecise data.

4.3 Birnbaum-Saunders distribution

In this subsection, we discuss the PDF that is used to model the geo-referenced error in the example presented in section 5, more specifically the data displayed by the Figure 3. (Birnbaum and Saunders 1969) introduced a family of Birnbaum-Saunders (BS) distributions motivated by problems of vibration in commercial aircraft that caused fatigue in materials. Although, in principle, its origin is for modeling equipment lifetimes subjected to dynamic loads, the BS distribution has been used for various other purposes, such as finance, quality control, medicine, and atmospheric contaminants. This distribution has two parameters, one of shape a and another of scale b, with b being also the median of the distribution. In addition, the BS distribution is asymmetric with positive skewness and unimodality. If a random variable T follows a BS distribution, denoted by T ~ BS(a; b), then is cumulative distribution function is given by

$$FT(t) = P(T \leq t) = \Phi(\ 1/a\ [\sqrt{}(t/b) - \sqrt{}(b/t)]\ ), \quad (4)$$

where $t > 0$, $\Phi(\cdot)$ is the standard normal cumulative distribution function. The BS model holds proportionality and reciprocal properties given by $kT \sim BS(a,\ k\ b)$, with $k > 0$, and $1/T \sim BS(a,\ 1/b)$, respectively. Also, when a tends to zero, the BS distribution tends to a normal model with mean b and variance τ, where $\tau \rightarrow 0$ when $a \rightarrow 0$.

5. RESULTS AND DISCUSSION

The method presented in this paper is exemplified with a data collected from a database of geocoded addresses in Goiânia-GO, Brazil. More specifically, it is a sample of 32 schools in that city. As it is common in geocoding process, the coordinates presents a significant error. To compute this error, it is necessary to have the real positions of each points – the schools in this case. For these 32 schools, the right location has been collected by taking its coordinates from Google Maps application. The errors error i were calculated using the expression

$$\text{error}_i = |real_i - database_i|, \qquad (5)$$

for $i = 1, ..., 32$. Those errors in meters are shown in y-axis of Figure 2. The cut line represents the value 350 of y-axis. Thus, one can see that most of the errors is smaller than 350 meters.

The first step is to find a PDF to model the errors. Figure 4 shows the histogram and Figure 5 presents the kernel density estimation with bandwidth equals to 91.5 for the errors verified in school data. The curve suggest a highly heavy tail distribution for modeling the errors. Despite most of the errors are relatively near to zero, there are a reasonable probability for some stay far beyond the mean or median of the data. Therefore, symmetric options like Normal distribution are not suited for the required model. Thus, the search must rely on asymmetric heavy tail PDFs. Bellow some statistics about the errors are presented.

min *= 4.0*
first quartile *= 40*
mean *= 290.1*
median *= 135.0*
third quartile *= 312.5*
max *= 2000.0*
standard deviation *= 433.9*
pearson's second skewness coefficient *= 1.07*

As can been seen, the skewness is greater than 0, indicating that most values of the distribution is concentrated left to the mean and that there is a heavy tail to the right. In this context, several heavy tail and asymmetric PDFs have been evaluated for modeling the error, until one in particular have been proved to achieve this purpose - the Birnbaum- Saunders (BS) distribution. BS possess the desirable requirements exposed before. In order to confirm (or not) the suspect that the error may be "well" modeled follows a BS, the Kolmogorov-Smirnov test was performed. The subsection 4.3 provides more information about this PDF.

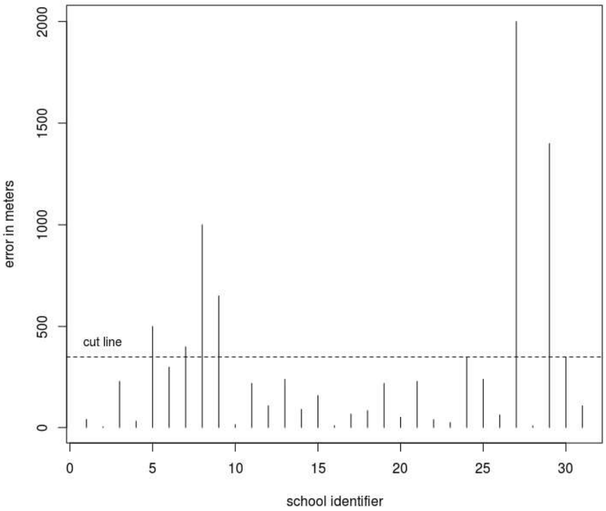

Fig. 3 - Location error for the 32 schools.

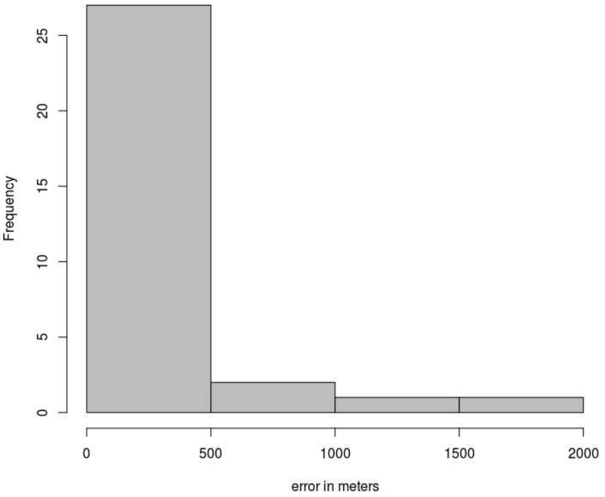

Fig. 4 - Histogram of errors.

The parameters *a* and *b* was estimated by the Maximum Likelihood Estimation method (MLE). The parameters estimation in this model is discussed, among others, by (Lemonte *et al.* 2007). Here we use a *R* implementation of MLE for this probability distribution. The estimated values were 1.88 for *a* and 100.2 for *b*.

To decide whether the *BS(1.88, 100.2)* fits well the data, an *R* implementation of the Kolmogorov-Smirnov test was used. The *p-value* returned by the test was 0.8745, which it is too high for refuting the hypothesis that the error does not follows a BS distribution. Therefore, the *BS(1.88, 100.2)* is considered a satisfactory model for the location error of the schools. Following the steps reported in subsection 4.2, Table 2 was constructed.

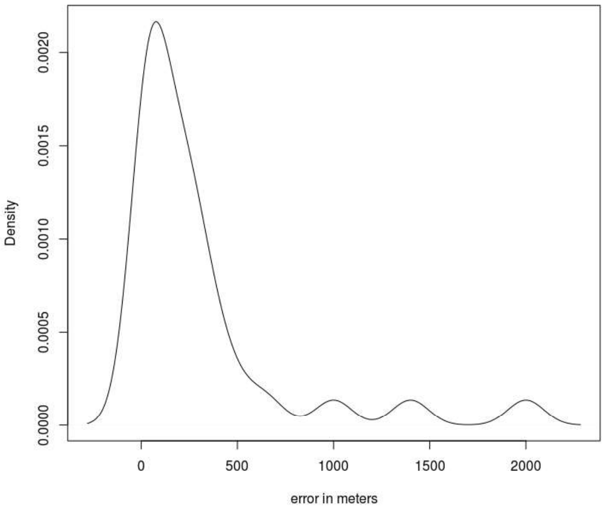

Fig. 5 - Density kernel estimation for the 32 schools.

Table 2: A sample of the dominance table generated from the school data

near/far	200	300	400	500	600	700
100	0.62	0.71	**0.76**	0.81	0.84	0.87
200		0.68	0.75	**0.80**	0.83	0.87
300			0.69	0.77	**0.82**	0.86
400				0.70	0.77	**0.83**
500					0.70	0.78
600						0.71

Using the table, the dominance criterion was changed. For example, if *P* is 200 meters from a reference place *L* and *Q* is 500 meters from *L*, then there is a probability of 80% that *P* is in fact closer to *L* than *Q*. Thus, there is 1 chance in 5 to get a type I error if one considers the point *P* closer to *L* than point *Q*. Therefore, for a cut probability *p* = 0.8, *P* dominates *Q*. If more than one dimension was considered, then the product of the probabilities would be used like explained in subsection 4.2. Note that when the difference between *P* and *Q* is 300 meters (bold values), the probability of dominance is approximately 80%. Thus, one looking to write a Skyline solution for these data could adopt this difference in meters as a rule of thumb to get a

80% confidence level in the evaluation of the dominance criterion.

In Figure 6 is shown a scenario where the distances of each school to a hypothetical person's residence is computed and displayed in y-axis. In x-axis is brought the percentage of teachers that have no mastery degree in each school. The problem is to find a school which minimize values in both dimensions, i.e. be cheap and with a high percentage of master teachers.

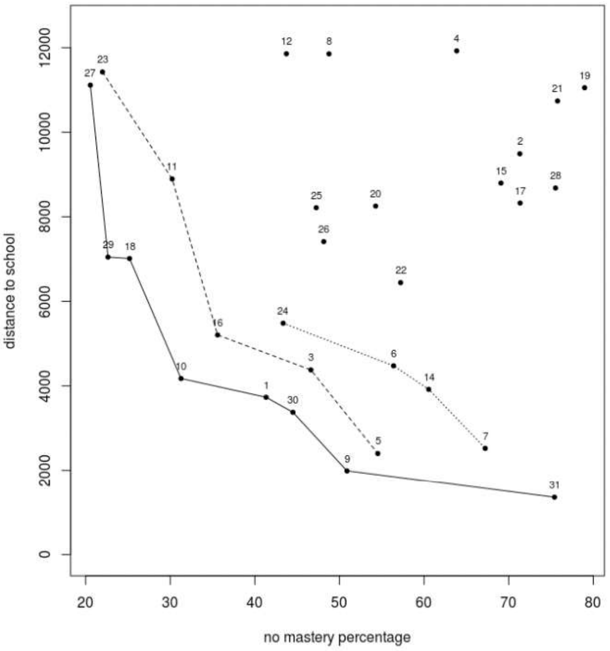

Fig. 6 - Skyline, second and third classes of dominant points in the school example.

In that case, it is assumed that just the distance to school is subjected to imprecision. Thus, the equation (5) suffices for our purposes. The skyline query is shown in the full line which contain the schools number: 27, 29, 10, 9, 31. Removing these points, a second class of Pareto efficient points emerge. Analogously, the third class of Pareto efficient is constructed. Since the positional data are imprecise, points in second and third class may, in fact, compose the *real* solution. Table 2 provide probabilities of dominance concerning the distances between the points and help one to make a more conservative decision about the schools who should not be discarded from the skyline solution.

Adopting the confidence level of 80% in order to a point be considered dominant over another, Table 2 shows that one point must be at least 300 meters lesser than another to be considered dominant. The Table 3 shows the distances from home of each school and class each one belong. Next, we will show which of the second classes points must lie in our version of Skyline, which we will call *S80* to differentiate it of *S* – the traditional Skyline query.

Note that schools 23 and 11 are both dominated by 29 by a distance superior to 300 meters. Then they should not stay in *S80*. For the same reason, 16 is not in *S80*, since it is dominated by 10 by more than 300 meters. The point number 5 are also dominated by point 9 by more than 300 meters. The same is true for point 3 which is dominated by point 30 by more than established threshold. Thus, no other point, instead those already in *S* must also be *S80* (*S80* = *S*). However, adopting 500 meters which correspond to a 90% confidence level, school 5 would be in the solution proposed by this paper.

Table 3: Distance from home of each school

Class	School	Distance
1	27	11113
	29	7047
	18	7013
	10	4174
	1	3730
	30	3376
	9	1989
	31	1355
2	23	11425
	11	8895
	16	5204
	3	4378
	5	2400
	24	5482
	6	4475
	14	3819
	7	2522

The example exposed in this section shows how the method can be used. Particularly, the first step presents some issues to be implemented efficiently, since finding a PDF for a random variable (corresponding to the error in our case) is not a trivial task. However, an alternative procedure could be employed in a future work: estimate the empirical distribution function. This approach would enable the method to be applied without the need to look for an ideal and known PDF. Nevertheless, one drawback in doing so is the lack of power that this kind of fitted has relating the a parametric approach like the ones get by finding the PDF. Also, in a future work, it is possible to implement in a programming language, a version of skyline query with the tolerance of errors proposed in this paper.

6. CONCLUSION

The present paper presented a contribution to the area of multi-criteria decision making providing a method to perform skyline queries in the presence of noised georeferenced data. Following the proposed method, it is possible to change the dominance criterion in skyline query, turning this technique more tolerant to location error. Imprecision of this type may be a common feature, for instance with coordinates obtained by a geocoding process. Therefore, the skyline query is generalized in this work from a deterministic to a probabilistic approach. Depending on the level of confidence one desire to have in order to eliminate a candidate from the Skyline query, the final solution changes. This way, the users of the proposed version of the Skyline query can adjust their confidence level when building their solutions based on this paper's contribution.

Despite the cited contribution, the proposed method only can be implemented in cases where the errors can be modeled by some known probability distribution. In general, this step may be hard to achieve. However, in a future work one can use the empirical distribution in order to automate this step. An advantage of this last approach would be simplicity and the declared automation. A drawback is the impossibility of choosing a parametric function like those provided by the known PDFs, which are able to provide more power in avoiding type I errors in the statistical hypothesis test (the new criterion created for dominance).

Another suggestion for a future work is to implement an end-to-end location recommendation technique, combining a first step with the noised tolerant version of skyline query exposed in this paper with a second step related to rank the points with some optimization procedure which also handle noised data, like, for example, that proposed by (QIN, 2013). The goal of our work has been achieved with the exposure of the method and also with its validation for a real example.

REFERENCES

BIRNBAUM, Z; SAUNDERS, S. C. A new family of life distributions. **Journal of applied probability**, pp. 319–327. 1969.

BORZSONY, S.; KOSSMANN, D.; STOCKER, K. The skyline operator. In Data Engineering, 2001. **Proceedings. 17th International Conference on**, pages 421–430. IEEE. 2001.

CHOMICKI, J.; GODFREY, P.; GRYZ, J.; LIANG, D. Skyline with presorting. In: **ICDE,** volume 3, pp. 717–719. 2003.

DING, X.; JIN, H.; XU, H.; SONG, W. Probabilistic skyline queries over uncertain moving objects. **Computing and Informatics**, 32(5):987–1012. 2014.

GODFREY, P.; SHIPLEY, R.; GRYZ, J. Maximal vector computation in large data sets. In: **Proceedings of the 31st international conference on Very large data bases**, pp. 229–240. 2005.

GROZ, B; MILO, T. (2015). Skyline queries with noisy comparisons. In: **Proceedings of the 34th ACM Symposium on Principles of Database Systems**, pp 185–198. ACM. 2015.

HUANG, Z.; LU, H.; OOI, B. C.; TUNG, A. K. (2006). Continuous skyline queries for moving objects. **Knowledge and Data Engineering, IEEE Transactions on**, 18(12):1645–1658. 2006.

HUANG, Z.; SUN, S.; WANG, W. Efficient mining of skyline objects in subspaces over data streams. **Knowledge and information systems,** 22(2):159–183. 2010.

KHALEFA, M. E.; MOKBEL, M. F.; LEVANDOSKI, J. J. Skyline query processing for incomplete data. In: **Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on**, pp. 556–565. IEEE. 2008.

KOSSMANN, D.; RAMSAK, F.; ROST, S. Shooting stars in the sky: An online algorithm for skyline queries. In: **Proceedings of the 28th international conference on Very Large Data Bases**, pp. 275–286. 2002.

LEE, M.-W.; SON, W.; AHN, H.-K.; HWANG, S.-W. (2011). Spatial skyline queries: exact and approximation algorithms. **GeoInformatica,** 15(4):665–697. 2011.

LEMONTE, A.; CRIBARI-NETO, F.; VASCONCELLOS, K. Improved statistical inference for the two-parameter Birnbaum-Saunders distribution. **Computational Statistics and Data Analysis**, 51:4656–4681. 2007.

LOFI, C.; EL MAARRY, K.; BALKE, W. T. Skyline queries in crowd-enabled databases. In: **Proceedings of the 16th International Conference on Extending Database Technology**, pages 465–476. ACM. 2013.

MASSEY JR, F. J. The Kolmogorov-Smirnov test for goodness of fit. **Journal of the American statistical Association**, 46(253):68–78. 1951.

METROPOLIS, N. The beginning of the Monte Carlo method. **Los Alamos Science**, 15(584):125–130. 1987.

OLIVEIRA, W. B.; OLIVEIRA, S. S. T.; RODRIGUES, V. J. S.; SANTOS, H. S. B.; & CARDOSO, K. V. A method for location recommendation via Skyline query tolerant to noised geo-referenced data. In: **GeoInfo 2015**.

PAPADIAS, D.; TAO, Y.; FU, G.; SEEGER, B. (2003). An optimal and progressive algorithm for skyline queries. In: **Proceedings of the 2003 ACM SIGMOD international conference on Management of data**, pp. 467–478. ACM. 2003.

PEI, J.; JIANG, B.; LIN, X.; YUAN, Y. Probabilistic skylines on uncertain data. In: **Proceedings of the 33rd international conference on Very large data bases**, pp. 15–26. 2007.

QIN, Z. **Uncertain random goal programming.** In: http://www.orsc.edu.cn/online/130323.pdf. pp. 1-9. 2013.

SHARIFZADEH, M. & SHAHABI, C. The spatial skyline queries. In: **Proceedings of the 32nd international conference on Very large data bases**, pages 751–762. 2006.

SON, W.; HWANG, S.-W.; AHN, H.-K. Mssq: Manhattan spatial skyline queries. **Information Systems**, 40:67–83. 2014

SON, W.; LEE, M.-W.; AHN, H.-K.; HWANG, S.-W. (2009). Spatial skyline queries: an efficient geometric algorithm. In: **Advances in Spatial and Temporal Databases**, Springer: pp. 247–264.

TAN, K.-L.; ENG, P.-K.; OOI, B. C. Efficient progressive skyline computation. In: **VLDB**, vol. 1, pp.301–310. 2001.

YOU, G.-W.; LEE, M.-W.; IM, H.; HWANG, S.-W. (2013). The farthest spatial skyline queries. **Information Systems**, 38(3):286–301. 2013.

ZHANG, M.; CHEN, S.; JENSEN, C. S.; OOI, B. C.; ZHANG, Z. Effectively indexing uncertain moving objects for predictive queries. **Proceedings of the VLDB Endowment**, 2(1):1198–1209. 2009.