Positioning of Public Service Systems Using Uncertain Data Clustering

Ivica Lukić, Mirko Köhler, Ninoslav Slavek

Faculty of Electrical Engineering, Josip Juraj Strossmayer University of Osijek,
Cara Hadrijana bb, 31000 Osijek, Croatia
ivica.lukic@etfos.hr, mirko.kohler@etfos.hr, ninoslav.slavek@etfos.hr

Abstract: Positioning of public service system is crucial and very challenging task. Proper positioning ensures that the public service system would complete its tasks to the end users. This paper is focused on finding the best location for public service system, to improve its efficiency when using uncertain data clustering. By choosing the best location for the service system the respond time can be minimised, and the given tasks could be performed in a reasonable time. Improved bisector pruning method was proposed for clustering previous data of public service system to find the best location for its application. Presented method can be used for different Public Service Systems, like traffic services, positioning of ambulance vehicles and other mobile objects. Cluster centres are used as best locations for public service systems because, cluster centres minimized total expected distance from tasks that have been set to the service system. On this way, public service system will be improved and can fulfil more tasks during the shortest period of time.

Keywords: clustering; data mining; expected distance; service systems; uncertain data

1 Introduction

Public Service System (PSS) data are saved in databases and this data can contain uncertainty and, therefore mining useful data from such uncertain database is not a simple task [1, 2]. Different factors, as measurement errors, sampling discrepancy, outdated data source contribute to data uncertainty. To cluster data with location uncertainty deploys various methods such as improved bisector pruning [3], MinMax pruning [4] and Voronoi pruning [5] that were used. Clustered objects are mutually similar, near to the cluster centre and similar objects are positioned in the same group. Cluster centres have minimized the total expected distance from all observed objects. Thus PSS should be located in the cluster centres or near them. Clustering methods are used for tracking of moving objects [6, 7], such as mobile devices, traffic services, ambulance vehicles, etc. Proper positioning ensures that public service system is as close as possible to accomplish its tasks and serve the end users. By choosing the best location, service system would minimize the distance to accomplish the previously set tasks and also their time to react properly. Data source can be outdated or contain errors and thus object location can contain uncertainty. Uncertain object location is not represented as a discrete point, but as an uncertainty region. In practice, uncertainty region is represented by a Probability Density Function (PDF). In this paper, PSS object's locations were presented in 2D dimensions and a two-dimensional uncertainty. In real life PDF applications can be specified using Gaussian distributions with the means and variances [6]. For Gaussian distributions, density function is exponentially dropped, meaning that probability density outside certain region equals zero. Thus, each object can be bounded by a finite bounding region. This region is limited by the maximum speed of the object and elapsed screening time. Clustering process must have short execution time because efficiency is very important for the PSS applications. Most of the computational time was lost on expected distances (ED) calculations [8, 9]. Large number of sample points that were used to represent each PDF [10], and a numerical integration is involved to calculate expected distance. Distance had to be calculated for all samples, thus computational cost were higher than in a simple distance calculation [11]. In [4] and [5], the pruning methods are introduced and thus the ED calculations can be avoided. In this study, the bounding regions may represent the pruning objects. Using these pruning methods, some clusters are eliminated as candidate clusters, if the closer cluster for observed object had been found. In [3] the improved bisector pruning method is presented to improve the clustering process. It is compared to the existing methods and it was experimentally proved that it had the best clustering results and shortest execution times in most of the situations.

2 Improved Bisector Pruning

Uncertain objects are data collection $O=\{o_1,...o_n\}$ in m dimensional space R^m. Distance between two objects is always greater than zero:

$$d(o_i,o_j)\geq 0 \tag{1}$$

Probability density function of each object at each point $x \in R^m$ is $f_i(x) > 0 \quad \forall\, x \in R^m$, where for all points inside MBR is:

$$\int_{x\in R^m} f_i(x)dx = 1 \tag{2}$$

Expected distance from object o_i to any point y is calculated using the formula:

$$ED(o_i,y) = \int_{x\in A_i} d(x,y)f_i(x)dx \tag{3}$$

Bounded region A_i is finite region and $f_i(x)=0$ is set for outside that region. Goal of clustering is to find set of clusters points $C=\{c_1,...c_m\}$ and all relations between objects and clusters $h:\{1,...,n\}\rightarrow\{1,...,m\}$, for which total expected distance (TED) from all the objects assigned to cluster centre is minimised, as shown in formula (4). $TED = \sum_{i=1}^{n} ED(o_i, c_{h(i)})$ (4)

Improved Bisector Pruning inherits principles of Voronoi and Bisector pruning method and it is improvement of the aforementioned method [5, 12]. Bisector pruning is a side product of Voronoi diagrams construction, and bisectors are calculated as a small additional calculation cost after Voronoi diagrams was constructed. Thus, Bisector pruning is combined with the Voronoi cell pruning [12]. In Improved Bisector pruning, Voronoi diagrams are not constructed and bisectors are calculated using formula (7), which is more effective than Voronoi pruning. Bisector is a line segment that is perpendicular to the line segment joining c_p and c_q, and that passes through the mid-point of the line segment. For each pair of clusters c_p and c_q in $C=\{c_1,...c_m\}$ bisector $B_{p/q}$ is calculated using the following formulas:

$$a = -\left(\frac{x_{cp} - x_{cq}}{y_{cp} - y_{cq}}\right) \tag{5}$$

$$b = \frac{x_{cp}^2 - x_{cq}^2 + y_{cp}^2 - y_{cq}^2}{2(y_{cp} - y_{cq})} \tag{6}$$

$$B_{p/q} = a * x + b \tag{7}$$

All bisectors are constructed using representative cluster points (x_{cp}, y_{cp}) and (x_{cq}, y_{cq}). For each cluster pair it was checked that if MBR_i of object o_i completely lies on the same side of bisector $B_{p/q}$ as cluster c_p, and if it does so, cluster c_q is pruned from object o_i. For the opposite situation, if MBR_i of object o_i completely lies on the same side of bisector $B_{p/q}$ as a cluster c_q, then cluster c_p is pruned from object o_i. Pruned cluster is instantly removed from cluster candidates. For 50 clusters there are 50 x 50 bisectors calculations. But for once pruned cluster, remaining bisectors are not constructed and this number is significantly reduced. For the clusters to be pruned, next properties must be satisfied:

$$\begin{aligned} &(y_{bcp} > y_{cp} \text{ and } y_{boi} > y_{oi}) \\ &or\,(y_{bcp} < y_{cp} \text{ and } y_{boi} < y_{oi}) \end{aligned} \tag{8}$$

In Figure 1 are explained principles used in above formula. For the points on the perimeter of MBR_2 it is checked do they lie on the same side of bisector as cluster c_3. To check this statement coordinate x_{c3} of cluster c_3 and coordinate x_{o2} of point on the perimeter of MBR_2 are included in bisector formula, and results y_{bc3} and y_{bo2} are obtained.

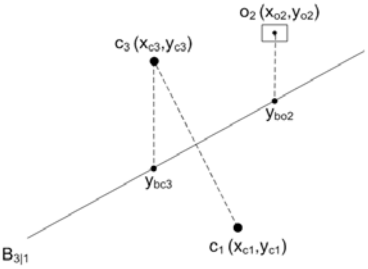

Figure 1
An example of Bisector pruning where projected coordinates are lower than original coordinates

From Figure 1 it is obvious, that object o_2 and cluster c_3 are on the same side of the bisector $B_{3/1}$ according to the formula (8). Obtained result y_{bc3} is lower than original coordinate y_{c3} of cluster c_3, and also y_{bo2} is lower than original coordinate y_{o2} of point on perimeter of MBR_2.

In the opposite situation, as it was shown in Figure 2, obtained result y_{bc3} is higher than y_{c3}, and y_{bo2} is higher than y_{o2}. Condition that was set in formula (8) is satisfied and object o_2 is on the same side of the bisector as cluster c_3. All the aforementioned steps must be repeated for the peak points on MBR_2. If all points satisfy the formula (8), cluster c_1 is pruned from object o_2.

After iterating all cluster pairs, we will find that the most of the clusters were pruned, and only for few remaining clusters ED calculation will be needed. In Voronoi pruning, if all clusters except one cluster are not pruned, ED must be calculated for all the clusters. Thus, Voronoi diagram is combined with Bisector pruning, to prune remaining clusters and thus avoid all the unnecessary ED calculations. Besides, calculations that are using formulas (1-5) are faster than Voronoi diagrams construction, and for the each object must be

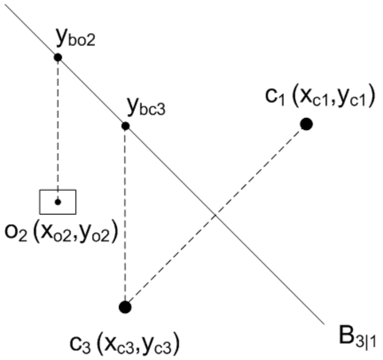

Figure 2
Example of Bisector pruning where projected coordinates are higher than original coordinates

checked does its MBR_i completely lies inside the Voronoi cell. Improved Bisector pruning is described by the following algorithm:

```
for all distinct $c_p, c_q \in C$ do
    if $MBR_i$ on the same side of bisector $B_{p/q}$ as cluster $c_p$ then
    Remove $c_q$ from $CC_i$ /*candidate clusters*/ Remove $c_q$ from iteration loop (for)
    else if $MBR_i$ on the same side of bisector $B_{p/q}$ as cluster $c_q$ then
        Remove $c_p$ from $CC_i$ /*candidate clusters*/
        Remove $c_p$ from iteration loop (for)
    for all the remaining candidate clusters calculated $ED$
```

2.1 Combination with the SDSA Method

Improved Bisector pruning method is combined with SDSA (Segmentation of Data Set Area) to shorten execution time of clustering process [13]. In SDSA method data set area was divided into small segments. Segments are parts of a total data set area, as it was shown in Figure 3. All segments have the same size, and they are rectangular. Observed were only objects in one segment and they pairs with clusters from that and neighbouring segments. Thus, number of object clusters observations was decreased. Experiments have showed that SDSA method combined with other pruning methods speeds up clustering process, depending on the number of clusters and objects [13]. Improvement of clustering process is reverse proportional to the size of segments. If segments are smaller than the clustering process is more effective. By decreasing the size of segments, the number of observed object clusters pairs is decreased, as shown in Figure 3.

An entire data set area C is shown in Figure 3a, and in Figure 3b data set area was divided into four small clusters segments SSDSA. Each segment was divided into four smaller segments. In Figure 3c are showed 16 small segments SSDSA and that was the final number of segments observed in this paper. Enlarged areas with object set inside them are shown in the Figure 3d, and Figure 3e. For 1600 objects and 64 clusters, there are 102400 object cluster pair calculations. If the SDSA method were used, the data set area was divided into 16 smaller segments of SSDSA, and clusters area C was divided into 4 small areas CSDSA. The average number of objects in one segment is 100, and the average number of clusters is four. Each segment was observed separately. Total number of observed objects was 100, number of observed clusters was 16 and the number of segments was also 16. Thus, total number of object cluster pair calculations was 25600, what represented the four times less calculations to proceed with. In this case, there was no need to check for all the clusters, but only clusters which are near one another and those who surrounded the area. All remaining clusters are pruned for all the objects that were contained inside the area. In this case, SDSA pruning decreases the total number of calculations by four times. When decreasing of the total number of calculations was done, it was proportional to decrease of the areas with clusters. However, segmentation had size limits, which are dependent on the number of clusters and their positioning. Segments have to be surrounded by clusters, and if the number of clusters is high, then segments can be very small and speed up the clustering process.

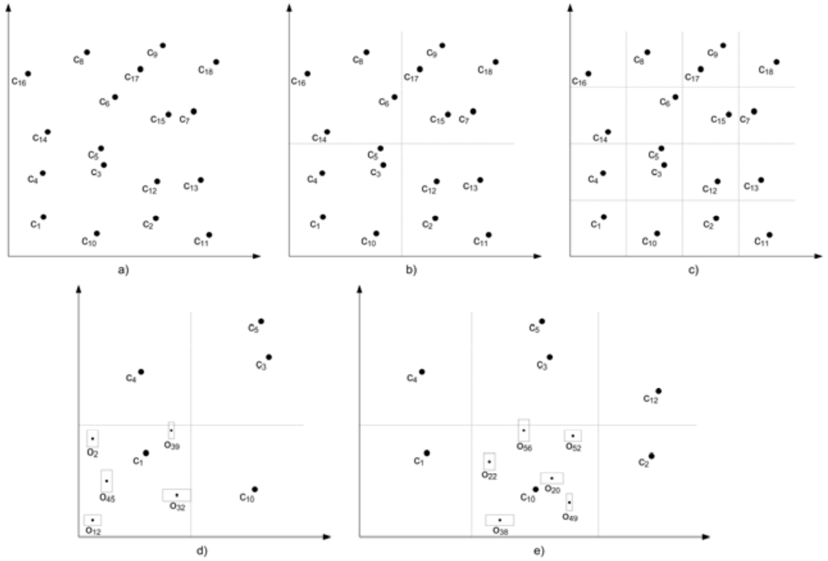

Figure 3
Data set area process of segmentation

3 Experiments

For clustering the existing PSS data, improved Bisector pruning method was used. In these experiments, as an emergency response public service vehicles in the city of Osijek were used. During the previous two months all the data were collected and processed to find the best location for positioning of the vehicles. Without clustering, vehicles were located in one place and needed more time to arrive on the remote locations. Arrival time is critical in some situations and human lives can depend on it. By clustering the existing data about PSS we could find cluster centres as the best locations that minimize total distance from executing possible tasks. Vehicles should be located in cluster centres and wait to be called upon for the fulfilment of the next task. All the new tasks were assigned to the waiting vehicles in the nearest cluster centre. It was found that vehicles from the cluster centre could accomplish the set task much faster than vehicle located in remote peripheral areas. In this paper, experiments were conducted with aims to prove that clustering existing service system data improves reaction time of PSS. In the set experiment, distance from cluster centres to tasks that have to been accomplished was measured and compared to the distance from existing central places to the set tasks. All experiments data were implemented in MATLAB 7.0 and carried out on a PC with an Intel Core(TM) 2 Duo Mobile CPU at 2.00 GHz, and 1.75 GB of main memory (RAM) All locations were presented in geo – coordinates system, where φ represents latitude, λ longitude and *r* radius of the Earth. Clustering methods are designed for clustering the Cartesian coordinates and conversion from spherical coordinates that had to be made. From Figure 4 is visible that conversion from spherical to Cartesian coordinates can be done using the following formulas:

$$x = r\,cos\,\varphi\,cos\,\lambda \tag{8}$$

$$y = r\,cos\,\varphi\,sin\,\lambda \tag{9}$$

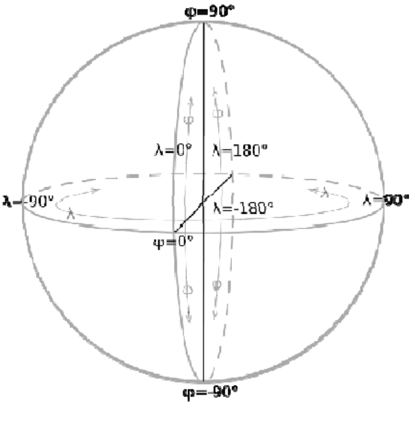

Figure 4

Spherical geo - coordinates representation.

3.1 Experiments Done When Using Three Clusters

In this experiment, all existing data about emergency responses in the city of Osijek were clustered in total of three clusters. Depending on the number of tasks that needed to be accomplished and which surround the cluster area, in each cluster centre should be located the exact number of vehicles. Three cluster centres are presented as red dots in Figure 5.

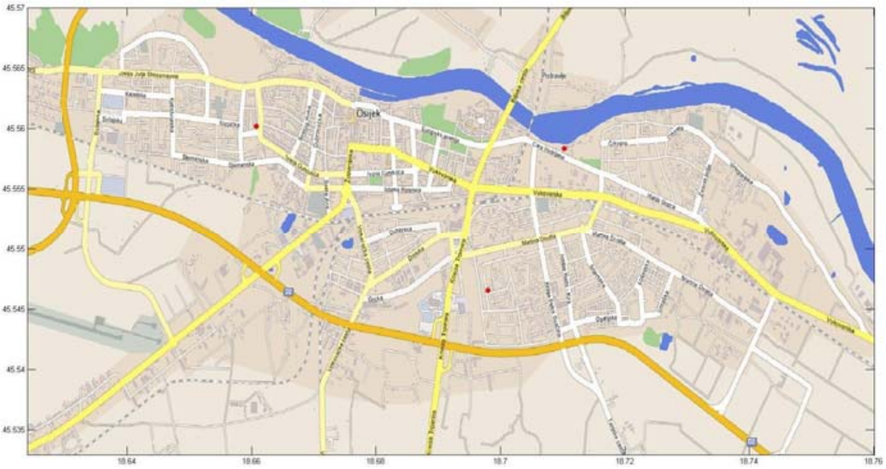

Figure 5

Public service system consisting of three cluster centres

In Table 1 is presented the number of tasks for each cluster centre. The most tasks are located in cluster centre No. 3, and therefore emergency response system should locate more vehicles in the surrounding area. Fewer vehicles were located in the remaining two cluster centres, according to the number of tasks for each cluster centre.

Table 1

Number of tasks settled in each of the cluster centres

Cluster centre	*Number of tasks*
Centre No. 1	362
Centre No. 2	531
Centre No. 3	744

Efficiency and reaction time of the vehicles were improved by positioning the vehicles effectively, when considering the PSS. When positioning the vehicles in the cluster centres, it can be ensured that total foreseeable distance from PSS that connects its tasks and users who use the service, is minimised. Total expected distance is reduced up to 40%, when compared to the situation without clustering been considered.

3.2 Experiments Done When Using Four Clusters

In this experiment all existing data about emergency responses in the city of Osijek were clustered in four clusters. Depending on the number of tasks that needed to be accomplished, and which surround the cluster area, in each cluster centre should be located the exact number of vehicles. Cluster centres are showed as red dots in Figure 6.

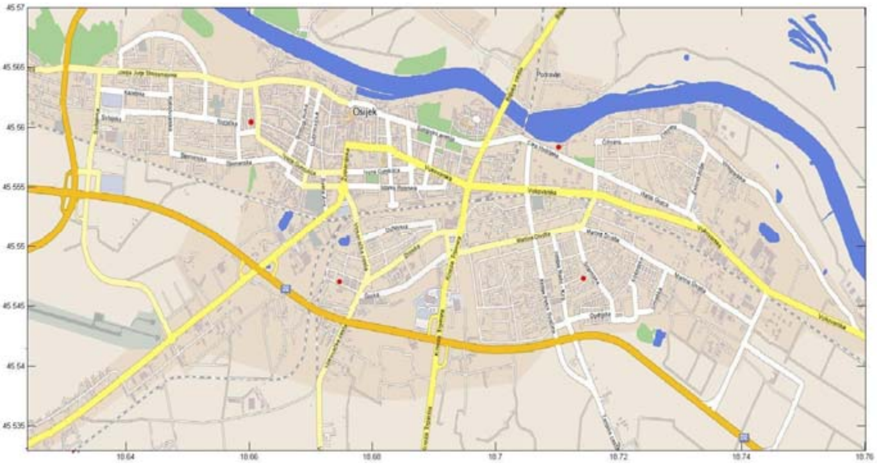

Figure 6

Public service system consisting of four cluster centres

In Table 2 is presented the number of tasks for each cluster centre. The most tasks are located in cluster centre No. 4, and therefore emergency response system should locate more vehicles in the surrounding area. Fewer vehicles were located in the remaining cluster centres, according to the number of tasks for each cluster centre. Total expected distance is reduced up to 40%, when compared to the situation without clustering been considered.

Table 2

Number of tasks settled in each of the cluster centres

Cluster centre	*Number of tasks*
Centre No. 1	295
Centre No. 2	317
Centre No. 3	334
Centre No. 4	691

3.3 Experiments Done When Using Five Clusters

In this experiment all existing data about emergency responses in the city of Osijek were clustered in five clusters. Depending on the number of tasks that needed to be accomplished, and which surround the cluster area, in each cluster centre should be located the exact number of vehicles. Cluster centres are showed as red dots in Figure 7.

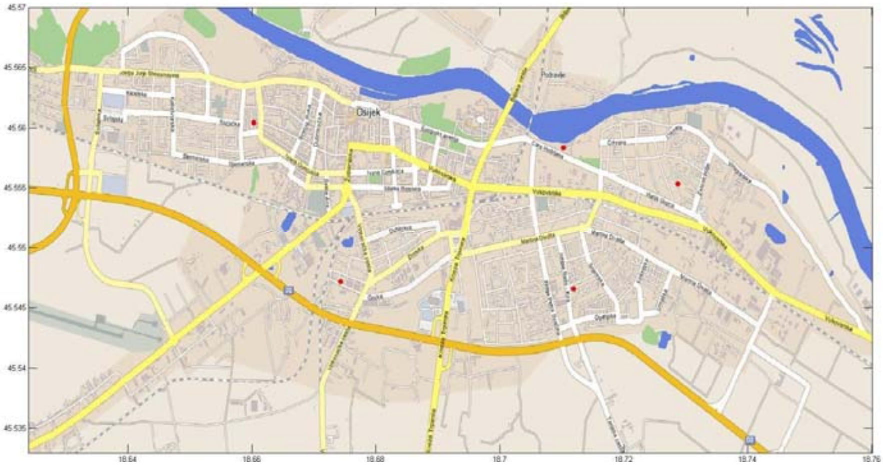

Figure 7

Public service system consisting of five cluster centres

In Table 3 is presented the number of tasks for each cluster centre. The most tasks are located in cluster centres Nos. 2 and 5 and therefore, emergency response system should locate more vehicles in the surrounding area. Fewer vehicles were located in the remaining cluster centres, according to the number of tasks for each cluster centre. Total expected distance is reduced up to 50%, when compared to the situation without clustering been considered.

Table 3

Number of tasks settled in each of the cluster centres

Cluster centre	***Number of tasks***
Centre No.1	227
Centre No. 2	482
Centre No. 3	375
Centre No. 4	86
Centre No. 5	467

4 Dedication to the Future Work

During our future work, we will create a database for storage of the geographical coordinates of all addresses in Osijek. Streets were divided into sections that represent one uncertain object with a minimum bounding region. Uncertain object is sampled, and to each sample is assigned probability for that object located in that very sample. The number of samples is proportional to the number of addresses in one city block, because every address is represented by one sample.

The probability that the object is located in certain sample represents a number of tasks that should be performed by the public service system when that observed sample is being considered. Database will contain all the tasks that service system should perform in each sample. For each pattern, the number of tasks that have to be done can be found in the database, and the value of the probability density function can be calculated. Based on the current data from the database we can predict future tasks of the public service system. The aforementioned prediction can significantly reduce the respond time and save funds for the public service system maintenance. By performing the public service data system, we can determine cluster centres that represent regions with the greatest concentration of tasks, for the operations of the public service system. Using predictive models, we can enable the public service system to deploy their workers in cluster centres and to efficiently accomplish the needs of their customers. Tasks of public service system are dependent on various unexpected events that could change the requests upon them. In our future work we will conduct experiments to predict tasks for the most unexpected events. For the purposes of the experiments we will make the script that shall contain the geographic coordinates of all addresses in Osijek. Experiments and simulations will be conducted for the normal operation of public service systems, and also experiments for the unexpected events where the number and site locations of tasks were significantly changed. Using predictive and existing data about similar events in the past, public service system will know where tasks might happened and properly deploy their employees to the target locations. The proposed model will reduce the costs of the public service system, enhanced the number of tasks that the can be done in a given period of time, and will also increase the reliability of the public service system.

Conclusion

This paper discussed an important role for choosing the best location for public service system. For this purpose, uncertain data clustering methods were used. Improved Bisector pruning method was used for clustering previous aggregated data of public service system to find the best location for their operational service. Cluster centres are used as best location for public service systems, because such centres minimized the total expected distance for the tasks that had been set to fulfil the service system operations. Several experiments were conducted. In the first experiment, public service system was divided into three clusters; in the second experiment in four clusters and finally, in the last experiment, in five clusters. Experiment with four clusters was used for calculation of distance reduction, because clusters in this experiment are nearest to the existing public service system locations set. It was experimentally proved that positioning of public service system can improve effectiveness and reaction time of such system, when comparisons were drawn to the existing system. Proper positioning ensures that public service system is as close as possible to accomplish its tasks and serve the end users. By choosing the best location, service system would minimize the distance to accomplish the previously set tasks and also their time to react properly. Our analyses have shown that the operating distances were reduced up to 50%. On this way, public service system will improve and can accomplish more tasks during the same period of time. In our future work, we will concentrate to further improve our model with additional parameters. We will try to apply our model to the other public service systems to improve existing systems and reduce its costs and also the service response time.

Acknowledgements

This work was supported by research project grant No. 165-1652017-2016, issued by the Ministry of Science, Education and Sports of the Republic of Croatia.

References

[1] D. Nilesh and D. Suciu, *“Efficient query evaluation on probabilistic databases”*. In Proc. of VLDB Conference, pages 864–875, 2004

[2] R. Cheng, X. Xia, S. Prabhakar, R. Shah, and J. Vitter, *“Efficient indexing methods for probabilistic threshold queries over uncertain data”*. In Proc. of VLDB Conference, 2004

[3] Lukić, M. Köhler, N. Slavek, “*Improved Bisector Pruning for Uncertain Data Mining*”, Proceedings of the 34th International Conference on Information Technology Interfaces, ITI 2012., pages 355-360

[4] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, et al. “Efficient clustering of uncertain data”. In ICDM, pages 436–445, 2006

[5] B. Kao, S. D. Lee, D. W. Cheung, W. S. Ho, K. F. Chan, *“Clustering Uncertain Data using Voronoi Diagrams”*. Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on Date:15-19 Dec. 2008. On pages: 333 – 342

[6] O. Wolfson, P. Sistla, S. Chamberlain, and Y. Yesha, *“Updating and querying databases that track mobile units”*. Distributed and Parallel Databases, 7(3), 1999.

[7] R. Cheng, D. Kalashnikov, and S. Prabhakar, *“Querying imprecise data in moving object environments”*. IEEE TKDE, 16(9):1112–1127, 2004.

[8] J. MacQueen, *“Some methods for classification and analysis of multivariate observations”*. In Proc. 5th Berkeley Symposium on Math. Stat. and Prob., pages 281–297, 1967.

[9] M. Ichino and H. Yaguchi, *“Generalized minkowski metrics for mixed feature type data analysis”*. IEEE TSMC, 24(4):698V–708, 1994.

[10] L. Xiao, E. Hung, *“An Efficient Distance Calculation Method for Uncertain Objects”*. Computational Intelligence and Data Mining, CIDM 2007., pages 10–17, 2007.

[11] M. Chau, R. Cheng, B. Kao, and J. Ng, *"Uncertain data mining: An example in clustering location data"*. In PAKDD, pages 199–204, Singapore, 9–12 Apr. 2006. Springer.

[12] B. Kao, S. D. Lee, F.K.F. Lee, D. W. Cheung, W. S. Ho, *"Clustering Uncertain Data using Voronoi Diagrams and R-Tree Index"*. Knowledge and Data Engineering, IEEE Transactions, Sept. 2010. On pages: 1219-1233

[13] Lukić, M. Köhler, N. Slavek, "The Segmentation of Data Set Area Method in clustering of Uncertain Data", Proceedings of the jubilee 35th International ICT Convention – MIPRO 2012., pages 420-425