**Proceedings of the 8th International Symposium on**
**Spatial Accuracy Assessment in Natural Resources and Environmental Sciences**
*Shanghai, P. R. China, June 25-27, 2008, pp. 341-347*

# Propagation and Visualization of Uncertainty in NL-Based Spatial Analysis

Danhuai Guo[1] +

[1] Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China

**Abstract.** Natural language (NL) as the most natural expression way of human is an ideal approach in spatial analysis. A growing attention of GIS scientists and AI scientists has been paid to the uncertainty of NL-based spatial queries and spatial analysis which include uncertainty of natural language expression and processing as well as the one of GIS model and spatial data source. The uncertainty of natural language expression that origin from vagueness of natural language itself, limitation of spatial cognitive ability of speakers and incompleteness of natural language expression will probably be brought into natural language processing result. In spatial analysis calculation based on natural language, the calculation result uncertainty is from uncertainty of starting feature, uncertainty of operator and uncertainty of calculating factor. The confidence of spatial analysis result is a mixture of the uncertainty sources listed above. Furthermore the visualization of spatial analysis uncertainty as another hot issue in uncertainty research field helps domain experts and other users realize reliability of spatial analysis results plainly. In this paper focusing on the study of propagation and visualization of uncertainty in NL-based spatial analysis, firstly, we summarize the uncertainty source of NL-based spatial analysis especially in NL-based location determination systems, secondly, we set up an uncertainty propagation model of NL-based spatial analysis and take an example of variant confidence distribution in variant part of linear and regional features in spatial analysis based on natural language, thirdly, we design a plain uncertainty visualization in GIS GUI, finally, an experimental prototype was developed to verify models, analysis result and visualization design.

**Keywords:** spatial analysis, natural language, natural language-based spatial analysis, uncertainty propagation, uncertainty visualization, confidence

## 1. Introduction

GPS receiver can tell us the coordinates of our location, map of electric media or traditional media can help us fix our bearing and site. How can we describe "where am I?" to other persons like policemen or rescuers without the two aides or other location tools in an unfamiliar place? It is the core question to be solved in Natural language (NL)-based spatial analysis that covers the technique of NL interpretation and spatial analysis based on NL.(Fig 1) The NL interpretation is the process of translating human natural language dialogue to the statement that computers can understand and handle. As incomplete, vagueness and ambiguous of natural language that caused by variant age, gender, educational background and other factors, complete NL interpretation is a great challenge [1].

A growing amount of GIS scientists, AI scientists and linguists have paid attention to the uncertainty and its related issues of NL-based spatial queries, spatial analysis and spatio-temporal reasoning. The uncertainty of NL-based spatial analysis mainly arises from vagueness of natural language, limitation of natural language interpretation model, and shortcomings of spatial analysis and spatial reasoning etc. The uncertainty would be accumulated and propagated during the whole process of spatial analysis. The analysis of the error propagation is based on prior knowledge of errors in the sources, procedure or manipulation of the data. The spatial data uncertainty is from the complexity of the real word, the limitation of human spatial recognition, the weakness of computerized machine, or the shortcomings of techniques and methods. And the uncertainty

+ Corresponding author. Tel.: +86-10-64889558
*E-mail address*: guodanhuai@gmail.com

**ISBN:** 1-84626-170-8, **ISBN13:** 978-1-84626-170-1

of NL interpretation origins from vagueness of natural language itself, limitation of spatial cognitive and expressing ability of speakers, and incompleteness of natural language expression and their mixed affect.

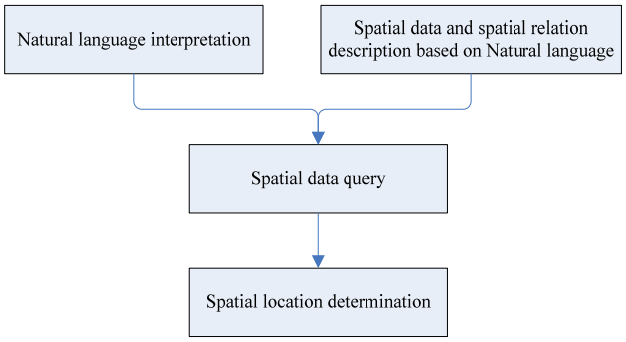


Fig 1 Key techniques of location determination systems based on natural language and their relationship [1]

This paper focuses on the propagation and visualization of uncertainty in NL-based spatial analysis. The following sections are arranged as follow: Section 2 introduces the related work. Section 3 investigates the uncertainty sources and its propagation model. Section 4 compares the main uncertainty visualization method and proposes uncertainty visualization based on confidence. The conclusion is described in Section5.

The International Symposium on spatial accuracy assessment in natural resources and environmental sciences (SPATIAL ACCURACY) mainly publishes original research and applied papers on uncertainty in spatial information, analysis, and applications emphasizing natural resources and the environment, etc.

## 2. Related work

The spatial analysis based on natural language and its related issues have attracted attentions of increasing amount of researchers in the past decades. [2] developed a formal model that captured metric details for the description of natural language spatial relations. The metric details were expressed as 9-intersection, a model for topological spatial relations and provided a more precise measure than does topology alone as to whether a geometric configuration. Yao Xiaobai and other researchers investigated spatial query qualitative locations in spatial information systems, and proposed a mechanism and a conceptual framework to handle queries with qualitative locations in geospatial information systems [3, 4]. [5] defined a series of quantitative indices that were related to natural-language spatial relation terms on the basis of a human subject test of natural language descriptions of spatial relations between linear geographic objects, and used these indices to formalize the ambiguous natural-language representation with decision-tree algorithm. [6] comprehensively investigated the uncertainties of spatial data and analysis. [7, 8] presented a methodology for the design and implementation of spatial query language, and described three natural-language spatial relations including 9-intersection model, interior direction relations and topological and interior directions relations. [9] presented the attribute uncertainty in GIS data, with a complete perspective of concepts, sources, nature and applicable tools. [10] pointed out that the GIS uncertainty can be accumulated and propagated during the process of machine-based computing analysis. [1] introduced how to make use of uncertainty to improve the querying accuracy of natural language location determination system. [11] pointed out that natural language was highly ambiguous in its structure, and in the many possible structures only one could be assigned to a sentence.[12] investigated how people describe objects in spatial scenes using natural language, and described a system that uses synthetic vision to “see” such scenes from the person’s point of view, and that understands complex natural language descriptions referring to objects in the scenes.

## 3. Uncertainty propagation in NL-based spatial analysis

As mentioned in the introduction section, the uncertainty of NL-based spatial analysis result is from the uncertainty of natural language, the uncertainty of natural language interpretation, the uncertainty of spatial data, and the uncertainty of spatial analysis. These uncertainties will be accumulated and propagated in the whole process of NL-based spatial analysis. It is necessary to build a model to express the complicated propagating process mathematically.

## 3.1. Uncertainty source

The uncertainty of NL- based spatial expressing is origin from the complexness of real world, the uncertainty of natural language, and spatial recognition ability of speakers. The complexness of real world makes it difficult to describe the real world in a simple and complete mathematic model. All the models are only approximation to real world in variant level. In natural language based spatial objects and spatial relation description, the current methods are of flaw. Most of current research is in the stage of quality expressing of spatial relation [2-4, 7, 13]. The human natural language expression is filled with ambiguousness, vagueness, incompleteness and fuzziness. In the conversations between two people, each sentence can be understood in more than one way. Due that the information is interactive and the information interchange is progressive, the talkers would not feel uncomfortable and confusing. But it is different in the human-machine interface circumstance, the machine receives the input information passively, and cannot judge whether there is ambiguity in the inputting statement. Each inputting sentence is treated as a complete, clear statement. The output hardly accord with the speaker's real meaning. In the spatial expressing based on natural language, same objects may be described variously by various people even by same person in variant circumstance. e.g. A university have several tiles due to dialect and historic reasons. It is found from questionnaire and on-the-spot investigation that the same university has four different titles. The first title is official tile, with 30% using ratio; the second title is title-in-short, with 50% using ratio, greater than the official title. The third title is historic tile and is still used by 15% people; while the forth title is an incorrect title with 5% using ratio.(Fig 2) The natural language-based spatial expression is impacted by the expresser's age, gender, educational background, occupation of transporting tools, spatial cognition ability and the mode at that time. e.g. The distance which the word "near" represents is different by a walker and a driver. Some spatial relation description especially directional one is influenced by a historical factor because of changes of city layout and other reasons. [1]

| *Filed Name* | *Data Type* | *Value* |
|---|---|---|
| Standard Name | string | 昆明理工大学 |
| Geo_Code | integer | 30010 |
| Title(0).Name | String | 昆工 |
| Title(0).Ratio | Single | 0.5 |
| Title(1).Name | String | 昆明理工大学 |
| Title (1). Ratio | Single | 0.3 |
| Title(2).Name | String | 昆明工学院 |
| Title (2). Ratio | Single | 0.15 |
| Title(3).Name | String | 昆大 |
| Title (3). Ratio | Single | 0.05 |
| …… | | |

Fig 2 An example of a university's titles and corresponding using rate. [1]

The uncertainty of natural language processing is mainly from the flaw of processing model, incompleteness of training sample and weakness of computerized machine. Diversity of natural language determines that there is no complete model to process all language. Most of current research focuses on machine-translation (MT). The machine translating method includes the method based on language rules, the method based on corpus, and multi-engine machine translation [14]. In machine translation research and implementation, language translating knowledge is from statistics result, corpus and example library. The knowledge gathering, knowledge extraction and rule (model) building are inevitably influenced by uncertainty. No matter which method mentioned above is employed, there will be more than one translating result with various possibilities. The object of machine translation is to find the translating result with most possibility. Many ambiguities and uncertainties may be preserved during translation, and thus will be accumulated and propagated in NL-based location determination systems.

Spatial data is conceptual abstraction of real spatial world, and observed and recorded by surveying and mapping instruments. The uncertainty pervades the GIS data with the lifecycle of capture, storage, update, transmission, access, archive, restore, deletion, and purge. It mainly arises from the complexity of the real

word, the limitation of human recognition, the weakness of computerized machine, or the shortcomings of techniques and methods. In detail, they may include instruments, environments, observers, projection's algorithms, slicing and dicing, coordinate system, image resolutions, spectral properties, temporal changes, etc. At the same time, their current limitations might further propagate even enlarge the uncertainty during the process of GIS analysis [9]. The uncertainty of spatial data includes but not limits: lineage, positional accuracy, attributes accuracy, logical consistency, completeness, temporal accuracy, and semantic accuracy[6].

Spatial analysis transforms spatial data to spatial knowledge according to specific spatial parameter. The assumption and premise of current GIS software and other spatial data analysis tools designing is that there is no error in spatial data. And GIS software only processes certain data, but it is contradictious with universality of uncertainty in spatial data. Spatial analysis based on certain theory and model will produce a large amount of result conflict with reality. It is difficult to eliminate uncertainty totally. The feasible way is to extend ability of current GIS and other spatial analysis model to handle uncertainty or to build new model and analysis method based on uncertainty.

## 3.2. Uncertainty propagation model

The analysis of the uncertainty propagation is based on prior knowledge of uncertainty in the sources, in the procedure or in the manipulation of the data. The spatial description often includes spatial entity, spatial relationship of spatial entity and spatio-temporal procedure[15]. And their spatial semantic role can be defined as: Entity, Attribute, Path, Motion, Time, Spatial Relation of Topology, Spatial Relation of Direction, and Spatial Relation of Measure. There are three alternative methods to analyze the propagation of the attribute uncertainty, i.e., Taylor series method, Monte Carlo method and sensitivity analysis. Taylor series method is to approximate the function by a linear function that is locally a good approximation of the function. Monte Carlo method uses an entirely different approach to analyze the propagation of error through the GIS operation. Sensitivity analysis can be of great value in acquiring meta-information when results of uncertainty analysis are difficult to be obtained. [10] As investigation in Section 3.1, the uncertainty of NL-based spatial analysis is a combined result from variant sources. It is hard to make a simple uncertainty propagating model covering all these abundant sources and complicated propagating procedure. This section proposes a model based on confidence to express the NL-based spatial analysis result and its uncertainty, and take NL-based location determination as an example.

In NL-based position determining, a location query statement could be divided to three components, which are Query Feature, Query Operator and Query Rang. In natural language interpretation considering uncertainty, each source statement may be translated to one or more target statements. All the target statements and their various confidences will be kept in the whole procedure, while in current machine translation, only the one with max confidence will be kept and the others will be neglected. Each element of each query statement has one or more translated result with variant confidence.

$$\text{Statement} = S\{\text{QueryFeature}, \text{QueryOperator}, \text{QueryRange}\} \tag{1}$$

$$\text{QueryFeature}=F\{\text{Feature}_1, \text{Feature}_2, ...\text{Feature}_m\} \tag{2}$$

$$\text{QueryOperator}=O\{\text{Operator}_1, \text{Operator}_2, ...\text{Operator}_n\} \tag{3}$$

$$\text{QueryRang}=R\{\text{Rang}_1, \text{Rang}_2, ...\text{Range}_k\} \tag{4}$$

$$\text{Result}=F\times O\times R \tag{5}$$

where, the natural translating result is the Cartesian product of the three element collection. It must be noted that all the elements have their confidence, and the confidence will be brought into the calculating procedures.

The spatial calculation is consisted of start calculating feature, calculating operator and calculating value. In natural language-based spatial analysis, they are correspondingly QueryFeature, QueryOperator and QueryRang. The confidence of calculating result is the confidence Cartesian product of QueryFeature,

QueryOperator and QueryRang. QueryFeature includes uncertainty of spatial entity position, attribute and PAT(Position and Thematic) [6]. QueryOperator uncertainty mainly origins from uncertainty of quality spatial relation description, and QueryRang uncertainty is from uncertainty of quantity metric spatial relation description. All the uncertainty can be represented by confidence. Current GIS software employs buffering analysis to determine locations. Buffering analysis make a buffering area around the QueryFeature with specific width and direction. The metric buffering area shape of point is donut or circle, the one of line is a belt, and the one of polygon is an annular zone. The shape of directional buffering area is fan or annulus sector. The uncertainty of buffering analysis is from the position uncertainty and buffering width uncertainty [10, 16, 17]. The uncertainty can be measured by Standard Deviations Area defined as formula (6).[6]

$$A_{D\_Error} = Area(D) - \min\{Area(X), Area(\tilde{X})\} \quad (6)$$

Where Area(D) is buffering deviation area; $Area(\tilde{X})$ is the "true" buffering area; and Area(X) is buffering area calculated.

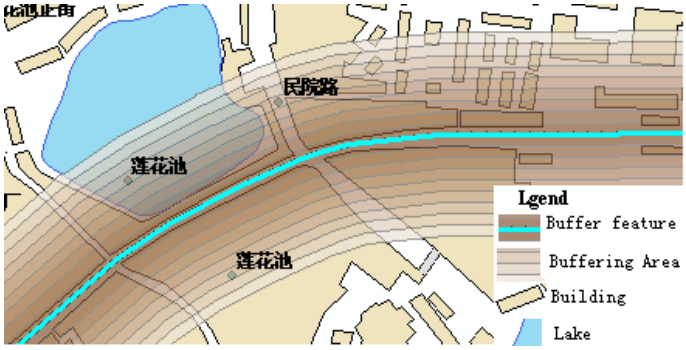


Fig 3 The buffering result include some inaccessible places

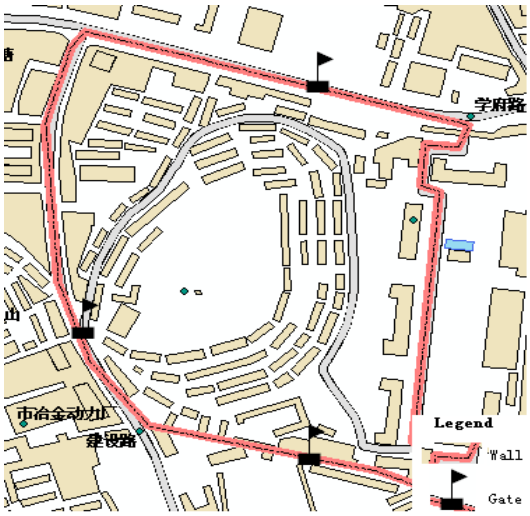


Fig 4 The university have 3 gates. The confidence of buffering result from gate and wall is different

In natural language based spatial analysis, the confidence of analysis result should take location accessibility into consideration. i.e. The query result maybe includes some inaccessible location (Fig 3). The confidence of inaccessible location should be set to smaller value or zero. To line and polygon feature, the confidence of buffering area of different part is different. e.g In the statement of "I am 100 meters away from Kunming University of Science and Technology", it is obvious the places which is 100m away from the gates have more confidence than those from the wall (Fig 4).

## 4. Uncertainty visualization

The goal of uncertainty visualization of natural language-based spatial analysis is to view the variant confidence of multi candidate query result, and to help the users understand and select the query result. The spatial data and spatial analysis uncertainty visualization method includes Error ellipse method, Arrow-based approach, Grey value-based approach, Color-based approach, Symbol-based approach, 3D-Based approach,

Animation method and Sound method (quoted in [6]). Ellipse method is suited to be used to represent location uncertainty of point feature in 2-D space. Arrow-based approach is efficient to describe uncertainty change of grid image location. Grey value-based method is suited to express specific type uncertainty data, whose uncertainty cannot be measured by definite measurement. Color-based approach employs color to express multi attribute uncertainty in one cell. Symbol-based approach represents uncertainty distribution by various symbols. 3D-based approach sets uncertainty to Z-value of 2-D map, the common 3D-based method includes contour graph and perspective graph. Animation method visualizes uncertainty in fast and continuous way. Sound method is an auxiliary means to "visualize" uncertainty.

In NL-based location determination system, it is essential to select a plain and could-be-combined in GIS GUI method to visualize uncertainty. We propose to use contour graph to view confidence.(Fig 5) And in the GIS GUI, use annotation and description text to aid the users to make options.(Fig 6)

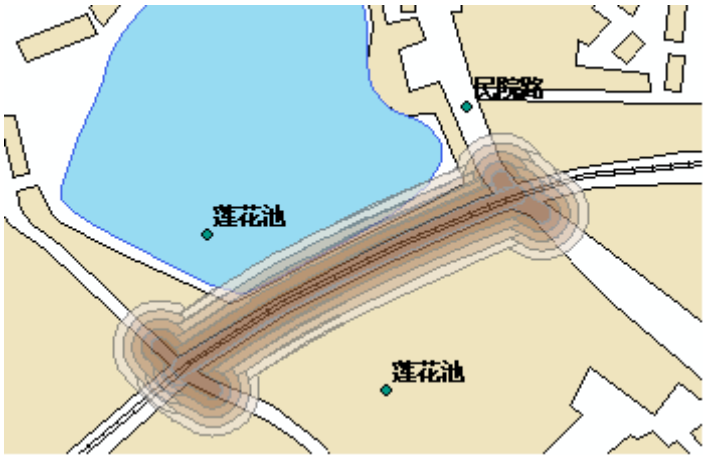


Fig 5 Uncertainty visualization based confidence contours graph

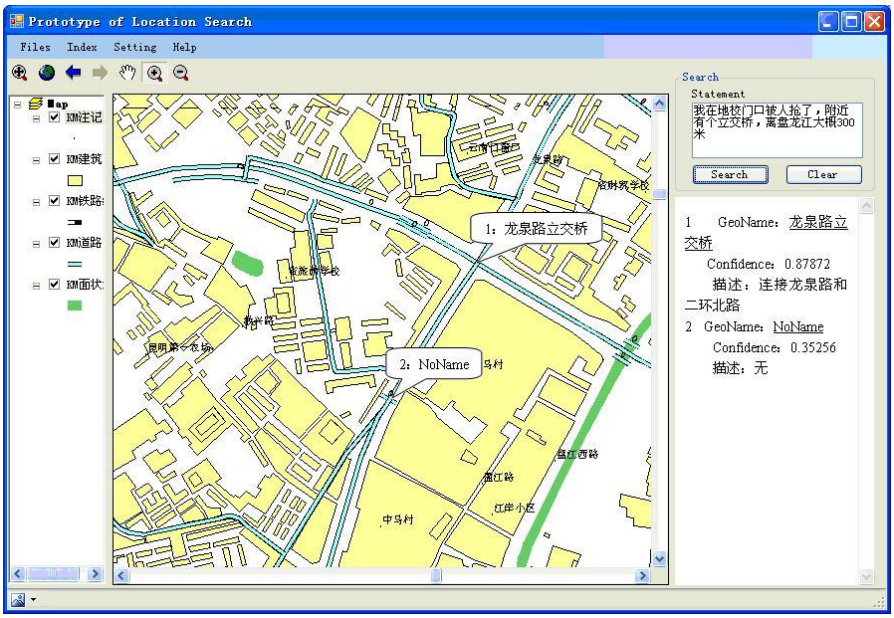


Fig 6 Uncertainty visualization combined with GIS GUI

## 5. Conclusion

This paper starts with the investigation of NL-based spatial analysis uncertainty source and propagating mechanism, proposes uncertainty propagation model and confidence-based uncertainty visualization method. The method combined with GIS GUI make the analysis result is plain and convenient for user option

## 6. Acknowledge

Thanks to Professor Weihong Cui for help in this paper.

## 7. References

[1] G. Danhuai and C. Weihong, "Quickly location determination based on geographic keywords of natural language," Nanjing, China, 2007: 67540.

[2] M. J. Egenhofer, A. Rashid, and B. M. Shariff, "Metric details for natural-language spatial relations," *ACM Transactions on Information Systems,* vol. 16, 1998: 295-321.

[3] X. Yao and J.-C. Thill, "Spatial queries with qualitative locations in spatial information systems," *Computers, Environment and Urban Systems,* vol. 30, 2006: 485-502.

[4] X. Yao and B. Jiang, "Visualization of qualitative locations in geographic information systems," *Cartography and Geographic Information Science,* vol. 32, 2005: 219-229.

[5] J. Xu, "Formalizing natural-language spatial relations between linear objects with topological and metric properties," *International Journal of Geographical Information Science,* vol. 21, 2007: 377-395.

[6] S. Wenzhong, Principal of Modelling Uncertaintiesin Spatial Data and Analysis vol. 1. Beijing: Science Press, 2005.

[7] S. Du, Q. Wang, and Z. Li, "Definitions of natural-language spatial relations in GIS," *Geomatics and Information Science of Wuhan University,* vol. 30, 2005: 533-538.

[8] S. Du, Q. Qin, D. Chen, and L. Wang, "Spatial data query based on natural language spatial relations," Seoul, South Korea, 2005: 1210-1213.

[9] S. Wang, W. Shi, H. Yuan, and G. Chen, "Attribute Uncertainty in GIS Data," in F*uzzy Systems and Knowledge Discovery*, 2005: 614-623.

[10] O. Bonin, "Attribute uncertainty propagation in vector geographic information systems: Sensitivity analysis," Capri, Italy, 1998: 254-259.

[11] T. F. Kalt, "Control Models of Natural Language Parsing," in *Graduate School of the University of Massachusetts Amherst*. vol. Doctor of philosophy Amherst: University of Massachusetts Amherst, 2005: 149.

[12] P. Gorniak and D. Roy, "A visually grounded natural language interface for reference to spatial scenes," Vancouver, BC, Canada, 2003: 219-226.

[13] D. Shihong, W. Qiao, and Q. Qiming, F*uzzy Description and Composite Reasoning of Spatial Relations*. Beijing: Science Press, 2007.

[14] F. Zhiwei, "The current situation and problems in machine translation," in *Some important issuse of Chinese information processing* Beijing: Science press, 2003: 448.

[15] L. Xiaoqiu, Y. Chongjun, and Y. Wenyang, "Spatial Concept extraction Based on Spatial Semantic Role in Natural Language," *Geomatics and Information Science of Wuhan University,* vol. 30, p. 4, 2005-11-03 2005.

[16] W. Shi, C.-K. Cheung, and X. Tong, "Modelling error propagation in vector-based overlay analysis," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 59, 2004: 47-59.

[17] B G Zhang, Z. Ling, and G Zhu, "The uncertainty propagation model of vector data on "buffer" operator in GIS," *Acta G eodaetica et Cartographica Sinica,* vol. 27, 1998:259-266, 2007-12-18.