General Entropy Model of Line Segments Uncertainty in GIS

SHI Yufeng  SHI Wenzhong

Abstract  Spatial data uncertainty can directly affect the quality of digital products and GIS-based decision making. On the basis of the characteristics of randomicity of positional data and fuzziness of attribute data, taking entropy as a measure, the stochastic entropy model of positional data uncertainty and fuzzy entropy model of attribute data uncertainty are proposed. As both randomicity and fuzziness usually simultaneously exist in linear segments, their omnibus effects are also investigated and quantified. A novel uncertainty measure, general entropy, is presented. The general entropy can be used as a uniform measure to quantify the total uncertainty caused by stochastic uncertainty and fuzzy uncertainty in GIS.

Keywords  uncertainty; line segment; general entropy; stochastic entropy; fuzzy entropy; uniform measure

CLC number  P208

Introduction

With the development of GIS, people are more and more interested in spatial data uncertainty. The purpose of spatial data uncertainty research is to discuss the occurrence, propagation and control of uncertainty. Spatial data uncertainty generally refers to error, imprecision, fuzzy and vagueness[1]. Generally, it can be classified as positional uncertainty, attribute uncertainty, temporal uncertainty, logistic inconsistency and incompleteness of data, and the uncertainty of spatial data characteristic is one of the hot issues in GIS research fields[1-3]. At present, research on the positional uncertainty is mainly focused on the model and propagation of line segment and area unit. Based on statistical theory, some researchers have studied the uncertainty models of point, line segment and polygon[2-6]. On the basis of “ε -band” model proposed by Chrisman, Shi developed it and presented a universal model to describe positional uncertainty of GIS data, and he also presented a general statistical description of these uncertainties[2]; based on randomized graph algebra, Shi also studied a probabilistic paradigm for handling uncertain objects[7]. Dai defined the visual indexes of point ellipse, line error band and polygon error donut to assess effect scope of their positional uncertainty according to the probability that the feature points dropping into their error ellipses based on the error ellipse expressing the positional uncertainty in surveying and mapping[8]. Cheung and Shi developed a model and continuous index, the probability value, to indicate the extent of the uncertainty point located inside the uncertain polygon[6]. Based on Shannon information theory, Fan proposed a new uncertainty band of error entropy (H-2 band), H-2 band can be used as an objective index of uncertainty[9]. Li proposed the information entropy models of spatial data positional uncertainty and discussed the error entropy models of point, line segment and polygon respectively[10]. From the viewpoint of pure mathematics, the uncertainty model
based on information entropy is indeed of statistical uncertainty. Because error entropy is educed on the basis of probability density, and error entropy model is a kind of stochastic entropy uncertainty model. But there are many fuzzy geographical entities in GIS, and due to these entities have intrinsic ambiguity, they can not be expressed by statistical theory. On the basis of field’s theory and model, Zhang combined positional uncertainty and attribute uncertainty to describe and analyze the spatial data uncertainty\[11,12\]. Owing to the procedure of spatial data storing and processing resembles information transmission very much, so based on the characteristics of information entropy and fuzzy entropy, this paper proposes stochastic entropy model of positional uncertainty of spatial data and fuzzy entropy model of attribute uncertainty of spatial data respectively, and takes randomicity and fuzziness into consideration comprehensively, proposes general entropy model of spatial data uncertainty. As for some ambiguous geographical phenomena, due to the randomicity and fuzziness exist in continuous form, so general entropy can better embody their uncertainty.

1 Stochastic entropy and fuzzy entropy

1.1 Definition and properties of stochastic entropy

Entropy is a kind of measurement to a system’s disorder, instability, imbalance, uncertainty, etc. Historically, entropy has three origins: thermodynamics, statistical mechanics, and information theory. Information entropy is an important concept in information source\[13\].

Definition of information entropy (discrete sample space): let $A$ be a probabilistic experiment with sample space $X$ and probability distribution $P$, where $p_i$ is the probability of outcomes $x_i \in X$ and they satisfy $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. Then Shannon information entropy is given by:

$$H_r(X) = E[-\log p_i] = -\sum_{i=1}^{n} p_i \log p_i$$

(1)

Information entropy $H_r(X)$ has the following main properties.

Symmetry: $H_r(X)$ is symmetric. $H_r(p_1, p_2, \cdots, p_i) = H_r(p_2, p_1, \cdots, p_i) = H_r(p_i, p_2, \cdots, p_1)$. That is, the ordering of the probabilities $p_1, p_2, \cdots, p_i$ does not influence the value of $H_r(X)$.

Non-negativity: $H_r(X)$ is non-negative. $H_r(p_1, p_2, \cdots, p_i) = -\sum_{i=1}^{n} p_i \log p_i \geq 0$.

Additivity: $H_r(X)$ is additive. If $X$ and $Y$ are two sample spaces, where outcomes in $X$ are independent of those in $Y$, then the information entropy of joint events $(a_i, b_j)$ is given by:

$$H_r(XY) = -\sum_{i=1}^{r} \sum_{j=1}^{s} p(ab) \log p(ab) = -\sum_{i=1}^{r} p_i \log p_i - \sum_{j=1}^{s} p_j \log p_j = H_r(X) + H_r(Y)$$

(2)

where $p_i$ and $p_j$ are the probability of space $X$ and $Y$ respectively, $0 \leq p_i \leq 1$, $\sum_{i=1}^{r} p_i = 1$ and $0 \leq p_j \leq 1$, $\sum_{j=1}^{s} p_j = 1$.

Maximality: $H_r(X)$ is maximum if all probabilities are equal. That is, $H_r(p_1, p_2, \cdots, p_i) \leq \log r$. This corresponds with the situation where maximum uncertainty exists. $H_r(X)$ is minimum if one outcome has a probability equal to 1.

Definition of information entropy (continuous sample space): for the continuous stochastic variable $x$ with probability density function $p(x)$, the amount of information is equal to

$$H_r(x) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx$$

(3)

Clearly, the definition of the continuous information entropy is based on analogy with the discrete one.

From the definition and characteristics of information entropy, we can see that information entropy is a function of probability density function and can be used as the measure of stochastic uncertainty. In order to differentiate and compare with fuzzy entropy in the following sections, information entropy is called stochastic entropy in this paper.

1.2 Definition and properties of fuzzy entropy

Fuzzy entropy is a proper measure of fuzzy subset
in fuzzy set theory\cite{14}.

Definition of fuzzy entropy: let $U$ be a non-empty finite set, fuzzy entropy, $H_f$ be a mapping that maps power set $2^U$ into non-negative real number space. That is, $H_f: 2^U \rightarrow [0, \infty)$, and for any two fuzzy subsets $A$ and $B$, they satisfy the following properties\cite{15}.

Sharpness: $\forall x \in X$, $H_f(A) = 0 \Leftrightarrow \mu_A(x) = 0$ or 1.

Maximality: $\forall x \in X$, $H_f(A)$ is maximum $\Leftrightarrow \mu_A(x) = 0.5$.

Resolution: if $B$ is a sharpened venison of $A$, then $H_f(B) \ll H_f(A)$; that is, if $\mu_B(x_i) < 0.5$ , then $\mu_A(x_i) < \mu_A(x_i)$; and if $\mu_A(x_i) > 0.5$, then $\mu_A(x_i) \geq \mu_A(x_i)$.

Symmetry: if $A'$ is the complement of fuzzy subset $A$, then $H_f(A) = H_f(A')$.

Valuation: $H_f(A \cup B) + H_f(A \cap B) = H_f(A) + H_f(B)$.

Many scholars have been studying the modeling of fuzzy entropy and proposed many fuzzy entropy models\cite{15-19}. With no consideration of probability density function, Deluca and Termini proposed the following fuzzy entropy model\cite{15}:

$$H_f(A) = -k \sum_{i=1}^{n} \mu_i(x_i) \log \mu_i(x_i) + (1 - \mu_i(x_i)) \log (1 - \mu_i(x_i)) \tag{4}$$

where $k$ is a positive constant and $\mu_i(x_i)$ represents the grade of membership of $x_i$ in fuzzy set $A$. It can be validated that Eq.(4) is satisfied the former four properties of fuzzy entropy.

Resembling Eqs.(3) and (4), we can define fuzzy entropy model of continuous variable as following:

$$H_f(A) = - \int_a^b \mu(x) \log \mu(x) + (1 - \mu(x)) \log (1 - \mu(x)) \, dx \tag{5}$$

where $\mu(x)$ is the membership function of fuzzy continuous variable, and $a$, $b$ are the lower limit and upper limit of fuzzy continuous variable respectively.

### 2 General entropy model

There may be stochastic uncertainty and fuzziness uncertainty simultaneously in a system. When the two kinds of uncertainty exist at the same time, Deluca and Termini proposed a measure to evaluate the total uncertainty of this system, and they called it total entropy.

Definition of total entropy: let $A$ be not only a probabilistic but also a fuzzy experiment within sample space $X$, then the total entropy model is given by:

$$H_{total} = H_r(A) + H_f(A) \tag{6}$$

where $H_r(A)$ is the stochastic entropy of sample space and can be calculated by Eq.(1); $H_f(A)$ is the fuzzy entropy of sample space and can be evaluated by Eq.(4).

From Eq.(6), it can be seen that the stochastic entropy and fuzzy entropy of a system can be evaluated in probabilistic space and fuzzy space respectively, and there are no systemic relationship between fuzzy entropy and stochastic entropy. But randomness and fuzziness often mix with together, for example, the boundary of soil in GIS has not only randomness but also fuzziness\cite{12,11,12,20}. Therefore, in order to wholly consider the total uncertainty caused by randomness and fuzziness, the product space, $R \times F$, defined by statistical space $R$ and fuzzy space $F$, should be considered.

Based on Shannon information entropy model and Deluca-Termini total uncertainty model, a novel entropy model, general entropy model, is proposed which can quantify the co-uncertainty caused by stochastic uncertainty and fuzzy uncertainty.

Definition of general entropy: let $A$ be not only a probabilistic but also a fuzzy experiment with sample space $X$, then its general entropy model in product space, $R \times F$, is given by Eq.(7), and the general entropy model should satisfies the following property.

When there is no fuzziness in the sample space, $H_r(R,F)$ degenerates to stochastic entropy model; if there is no randomness in the sample space, $H_f(R,F)$ degenerates to fuzzy entropy model.

$$H_g(R,F) = - \sum_{i=1}^{n} \left[ p_i \mu_i \log p_i \mu_i + p_i (1 - \mu_i) \log (p_i (1 - \mu_i)) \right] \tag{7}$$

where the meanings of all the items in Eq.(7) are equal to Eqs.(1) and (4).

Eq.(7) can be decomposed as following:

$$H_g(R,F) = - \sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} \left[ \mu_i \log (1 - \mu_i) + (1 - \mu_i) \log (1 - \mu_i) \right] + \sum_{i=1}^{n} (1 - p_i) \left[ \mu_i \log (1 - \mu_i) + (1 - \mu_i) \log (1 - \mu_i) \right] = H_r + H_f - H_f \tag{8}$$
From Eqs. (1), (14) and (8), it is easy to see that \( H_r \) and \( H_f \) are stochastic entropy and fuzzy entropy respectively, \( H_{rf} \) can be regarded as the mixing entropy which is the mixing-effect of fuzzy uncertainty and stochastic uncertainty, it is given by:

\[
H_{rf} = - \sum_{i=1}^{n} (1 - p_i) \left\{ \mu_i \log \mu_i + (1 - \mu_i) \log (1 - \mu_i) \right\}
\]  (9)

In this way, general entropy is equal to the summation of stochastic entropy, fuzzy entropy and minus mixing entropy. It can be shown by Fig. 1.

\[
H_s(R,F) = - \int_{-\infty}^{\infty} \{ p(x) \log p(x) + p(x) \log \mu(x) \} dx
\]
\[
H_f = - \int_{-\infty}^{\infty} \{ \mu(x) \log \mu(x) + (1 - \mu(x)) \log (1 - \mu(x)) \} dx
\]
\[
H_{rf} = - \int_{-\infty}^{\infty} (1 - p(x)) \{ \mu(x) \log \mu(x) + (1 - \mu(x)) \log (1 - \mu(x)) \} dx
\]  (14)

It is easy to see that Eq. (12) is the stochastic entropy model of continuous stochastic variable and Eq. (13) is the fuzzy entropy model of fuzzy continuous variable. Eq. (14) can be regarded as the cross-effect of fuzzy entropy and stochastic entropy, and analogy to Eq. (9) it can be called mixing entropy. From Eqs. (12)-(14), we can obtain the same conclusion resembling Eq. (8). Eqs. (12)-(14) is the general entropy model of continuous variable.

### 3 General entropy of line segment uncertainty

Spatial data is usually managed by object-oriented model in GIS, and it is a kind of common and standard spatial data model. The distribution of spatial data can be expressed by a set of discrete points, line segments and polygons in object-oriented model. The object-oriented model can be used to express clear defined spatial entities. Based on mathematical statistical theory and surveying adjustment method, many researchers have studied the object-oriented model of spatial data, and they focused on the positional uncertainty and proposed the error ellipse model of point uncertainty, “\( \epsilon \)-band” and “\( \epsilon \sigma \)-band” models of line segments, “\( g \)-ring” and “\( g\sigma \)-ring” models of polygons, etc. Whereas, there are many problems that can not be effectively expressed by statistical object model in GIS, such as the uncertainty analysis of landcover and soil classification produced by remote image. Due to the boundary of different vegetations or soils are seldom clear, the statistical object model can not be used to measure their uncertainty anymore, the statistical object model and fuzzy model should be taken into account. Based on the definition and characteristics of entropy in Section 1 and Section 2, this paper takes account of the co-effects of positional uncertainty and attribution uncertainty, and utilizes general entropy model to measure the total uncertainty caused by randomicity and fuzziness.

Suppose we have a polygon data classification model and it is illustrated by Fig. 2, \( A \) and \( B \) are denoted two different soils, and their boundary is line \( l \).
It is obvious that the position of boundary \( l \) has stochastic uncertainty due to the faulty of surveying method and technology in the process of fixing on boundary, whereas the soil properties in the respective sides of boundary have fuzziness due to the changing of soil property is continuous. So, there are two different kinds of uncertainties, stochastic uncertainty and fuzzy uncertainty, in the boundary. Suppose the error distribution of point in the boundary is equal to normal distribution, that is to say, positional uncertainty takes on randomicity, and its probability density function is equal to
\[
2 \sigma \exp \left( -\frac{x^2}{2\sigma^2} \right).
\]
Then the stochastic entropy is given by:
\[
H(x) = -\int \! p(x) \log p(x) \, dx = \ln \sigma \sqrt{2\pi} e \sigma (15)
\]
Form Eq.(15), it is can be seen that stochastic entropy of normal distribution has no relation with expectation of probability density function and only has relation with variance of probability density function. Then the radius of entropy uncertainty interval can be given by:
\[
R_{\text{entropy}} = \frac{1}{2} e^{H(x)} = \sqrt{\frac{\ln \sigma \sqrt{2\pi} e \sigma}{2}} = 2.066 \sigma (16)
\]
It is evidence that interval is \((-2.066 \sigma, 2.066 \sigma)\) the point error distribution interval based on stochastic entropy, and this interval concentrates the main uncertainty of stochastic variable, so it is the objective measure of stochastic uncertainty.

![Fig.2 Polygon data classing model](image)

Suppose that the changes of soil attributes at the respective side of boundary \( l \) are fuzzy distributions, and the membership function of their attribute is an invert triangular membership function, and it is shown in Fig.3. From Fig.3, it is easy to see that the value of membership function is equal to 0.5 at the point \( m \), that is to say, the fuzziness is maximum at point \( m \); the longer the distance from point \( m \), the less the fuzziness until reaching the maximum membership grade, 1. The fuzzy distributions of attribute \( A \) and \( B \) may be different in the respective sides of the boundary; their united fuzzy distribution membership function is \( \mu_{\text{A,B}}(x) \) which is given by Eq.(17), where \((c,d)\) are the interval terminals of soil \( A \) and \( B \)'s union distribution.

\[
\mu_{\text{A,B}}(x) = \begin{cases} 
\frac{1}{2(c-m)}(x-c-2m), & c \leq x \leq m \\
\frac{1}{2(d-m)}(x+d-2m), & m < x \leq d \\
1, & x < c \text{ or } x > d
\end{cases} (17)
\]

From Eq.(13), we can evaluate the fuzzy entropy which is given by Eq.(18); and from Eq.(18), it is easy to see that fuzzy entropy of triangular membership function has only relation with the fuzzy distribution interval and its value is equal to the half of interval.

\[
H_f(A \cup B) = -\int \! \left[ \mu(x) \log f(x) + (1-\mu(x)) \log (1-\mu(x)) \right] \, dx = -\int \! \left[ \mu(x) \log f(x) + (1-\mu(x)) \log (1-\mu(x)) \right] \, dx - \int \! [\mu(x) \log f(x) + (1-\mu(x)) \log (1-\mu(x))] \, dx = \frac{1}{2} (d-c) (18)
\]

From Eq.(14), the mixing entropy can evaluated as following:

\[
H_f = -\int \! \left[ (1-p(x)) \left[ \mu(x) \log f(x) + (1-\mu(x)) \log (1-\mu(x)) \right] \right] \, dx = -\int \! \left[ (1-p(x)) \left[ \mu(x) \log f(x) + (1-\mu(x)) \log (1-\mu(x)) \right] \right] \, dx - \int \! \left[ (1-p(x)) \left[ \mu(x) \log f(x) + (1-\mu(x)) \log (1-\mu(x)) \right] \right] \, dx = \frac{1}{2} (d-c) = 2 (19)
\]

Eq.(19) can not be integrated by symbolic integration, so the approximate value of \( H_f \) can only be calculated by numeric integral in the case of knowing distribution function, membership function and integrating interval.

Suppose that \( p(x) \sim N(0,1) \), that is to say the variance of stochastic distribution of point error is equal to 1, that is \( \sigma = \pm 1 \); fuzzy membership function is in symmetry and let \( m = 0, -c = d = \pm 2 \); then stochastic entropy is equal to \( H_s = \ln \sqrt{2\pi e} \sigma = 1.4189 \), fuzzy entropy is equal to \( H_f = \frac{1}{2} (d-c) = 2 \), and
mixing entropy is equal to $H_p = 0.720 \ 5$, which is evaluated by numeric integral. The general entropy is given by

$$H_g(R, F) = H_r + H_f - H_p = 1.418\ 9 + 2 - 0.720\ 5 = 2.698\ 4$$

From Eq.(16), the radiuses of stochastic entropy band, fuzzy entropy and general entropy of line segment are as follows respectively: 2.006 3, 3.694 5 and 5.616 7. The uncertainty distributions of line segment with stochastic entropy band radius, fuzzy entropy band radius and general entropy band radius are respectively illustrated in Fig.3.

![Fig.3 Stochastic entropy band, fuzzy entropy band and general entropy band of line segment](image)

References

[1] Goodchild M F(1995) Attribute accuracy[M]/Guptill S C, Morrison J L. Elements of Spatial Data Quality. Oxford: Elsevier Scientific:59-79

[2] Shi Wenzhong, Liu Wenbao(1998) The stochastic process model for handling positional uncertainty of line segments in GIS[J]. *Acta Geodaetica et Cartographica Sinica*, 27(1):37-44 (in Chinese)

[3] Shi Wenzhong(1998) Theory and methods for handling errors in spatial data[M]. Beijing: Science Press (in Chinese)

[4] Blakemore M(1984) Generalization and error in spatial databases[J]. *Cartographica*, 21:131-139

[5] Chrisman N R(1982) A theory of cartographic error and its measurement in digital databases[C]. Auto-Carto, Utrecht

[6] Cheung C K, Shi Wenzhong, Zhou Xian(2004) A probability-based uncertainty model for point-in-polygon analysis in GIS[J]. *Geoinformatica*, 8(1):71-98

[7] Shi Wenzhong, Liu Wenbao(1998) The stochastic process model for handling positional uncertainty of line segments in GIS[J]. *Acta Geodaetica et Cartographica Sinica*, 27(1):37-44 (in Chinese)

[8] Dai Honglei, Xiaozangguo, Huang Xingyuan(2002) Analyzing on visual indexes of positional uncertainty in GIS[J]. *Journal of Image and Graphics*, 4(3):239-249 (in Chinese)

[9] Fan Aimin, Guo Dazhi(2001) The uncertainty band model of error entropy[J]. *Acta Geodaetica et Cartographica Sinica*, 30(1): 48-53 (in Chinese)

[10] Li Dajun, Gong Jianya, Xie Gangsheng, et al.(2002) Error entropy band for linear segments in GIS[J]. *Geomatics and Information Science of Wuhan University*, 27(5):462-465 (in Chinese)

[11] Zhang Jingxiong, Du Daosheng(1999) Field–based models for positional and attribute uncertainty[J]. *Acta Geodaetica et Cartographica Sinica*, 28(3):244-249 (in Chinese)

[12] Zhang Jingxiong, Goodchild M F(2002) Uncertainty in geographical information[M]. London:Taylor & Francis

[13] Shannon C E(1948) A mathematical theory of communication[J]. *Bell System Technical Journal*, 21: 379-423, 623-656

[14] Zadeh L A(1965) Fuzzy sets[J]. *Information and Control*, 8:338-353

[15] Deluca A, Termini S(1972) A definition of nonprobabilistic entropy in setting of fuzzy sets theory[J]. *Information control*, 20:301-312

[16] Kosko B(1986) Fuzzy entropy and conditioning[J]. *Information Science*, 40:165-174

[17] Pal N R, Bezdek J C(1994) Measuring fuzzy uncertainty[J]. *IEEE Transactions on Fuzzy System*, 2(2): 107-118

[18] Yager R R(1995) Measure of entropy and fuzziness related to aggregation[J]. *Information Science*, 82: 147-166

[19] Zadeh L A(1968) Probability measures of fuzzy events[J]. *Journal of Mathematics Analysis and Application*, 23: 421-427

[20] Foody G M(1996) Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data[J]. *Int J Remote Sensing*, 17: 1317-1 340, 1996