ARTICLE

Analysis of Positional Uncertainty of Road Networks in Volunteered Geographic Information with a Statistically Defined Buffer-zone Method

Wenbin Zhang[a], Yee Leung[b] and Jianghong Ma[a]

[a]Department of Mathematics and Information Science, Changan University, Xian, P.R. China; [b]Institute of Future Cities, Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong SAR,P.R. China

ARTICLE HISTORY
Compiled September 15, 2019

ABSTRACT
Volunteered geographic information (VGI) is crowdsourced information that can enrich and enhance research and applications based on geo-referenced data. However, the quality of VGI is of great concern, and positional accuracy is a fundamental basis for the VGI quality assurance. A buffer-zone method can be used for its assessment, but the buffer radius in this technique is subjectively specified; as result, different selections of the buffer radius lead to different positional accuracies. To solve this problem, a statistically defined buffer zone for the positional accuracy assessment in VGI is proposed in this study. To facilitate practical applications, we have also developed an iterative method to obtain a theoretically defined buffer zone. In addition to the positional accuracy assessment, we have derived a measure of positional quality, which comprises the assessment of positional accuracy and the level of confidence in such assessment determined with respect to a statistically defined buffer zone. To illustrate and substantiate the theoretical arguments, both numerical simulations and real-life experiments are performed using OpenStreetMap. The experimental results confirm the high significance of the proposed statistical approach to the buffer zone-based assessment of the positional uncertainty in VGI.

KEYWORDS
Buffer-zone method; OpenStreetMap; Positional uncertainty; Statistically defined buffer zone; Volunteered geographic information.

CONTACT Yee Leung. Email: yeeleung@cuhk.edu.hk

ABSTRACT
Volunteered geographic information (VGI) is a crowdsourced information that can enrich and enhance research and applications based on geo-referenced data. The quality of VGI, however, is of great concern and positional accuracy is the fundamental basis for VGI quality assurance. Buffer-zone method is a way to assess positional accuracy. The problem of the current buffer-zone method is that the buffer radius is subjectively specified. With different selections of buffer radius, we often end up with different conclusions on positional accuracy. To overcome such shortcomings, we propose in this paper a statistically defined buffer zone for positional accuracy assessment in VGI. To facilitate practical applications, we also construct an iterative method to obtain the theoretically defined buffer zone in practice. To go beyond positional accuracy assessment, we propose a measure of positional quality which comprises the assessment of positional accuracy and the level of confidence in such assessment with respect to the statistically defined buffer zone. To illustrate and substantiate the theoretical arguments, we perform numerical simulations and carry out real-life applications in the OpenStreetMap (OSM). The empirical results render a firm support to the significance of the proposed statistical approach to buffer-zone-based assessment of positional uncertainty in VGI.

KEYWORDS
Buffer-zone method; OpenStreetMap; Positional uncertainty; Statistically defined buffer zone; Volunteered geographic information.

1. Introduction

Volunteered geographic information (VGI) is an important crowdsourced data that can fundamentally enhance and enrich the information content in geographical research (Goodchild 2007b). On the other hand, because such information is largely provided by volunteers who are not professionals, the quality of VGI is an issue of great concern in research and applications (Van Exel *et al.* 2010). The lack of standardization or gatekeepers makes the quality of VGI a question mark (Goodchild and Li 2012). A satisfactory solution to the quality issue of VGI will make this crowdsourced geo-referenced information valuable to research and real-life decision making (Coleman 2013). Among other quality issues of VGI, positional accuracy is an essential problem that needs to be resolved. How to characterize and measure positional accuracy has thus become a fundamental problem in quality assessment of VGI (Goodchild and Li 2012).

In the literature, positional accuracy is employed as a measure of positional quality in VGI, and the buffer-zone method is a common method for assessment (Senaratne *et al.* 2017). Under the buffer-zone method, reference datasets have been used to assess positional accuracy of map-based VGIs, mainly the OSM data (Kounadi 2009, Al-Bakri and Fairbairn 2010, Ciepłuch *et al.* 2010, Haklay *et al.* 2010, Helbich *et al.* 2013, Arsanjani *et al.* 2015). The assessment results, however, directly depend on the choice of the buffer radius as there is a strong positive correlation between the size of the buffer radius and the assessment result of positional accuracy (Zhou 2017). However, the buffer-zone method is a concept that has not been subjected to rigorous theoretical and experimental investigations. Furthermore, buffer zones are subjectively specified with no statistical justifications. Without a sound theoretical and computational basis, it cannot provide a proper and reliable assessment and interpretation of positional accuracy and its impact on the products or applications of VGI (Goodchild and Li 2012). Other than the process of measurement, much more is involved in assessing positional quality of VGI, such as credibility, which makes it difficult to intrinsically assess the quality (Goodchild 2007a, Haklay and Weber 2008).

Against this background, we employ the basic measurement-error (ME) model (Le-

ung *et al.* 2004a) as the theoretical basis to formulate an objective and statistically rigorous buffer-zone method for the assessment of positional quality in VGI. This model treats the two end-point measurement-error vector of a road line segment as a random vector, and error propagation from each point of the road becomes a statistical estimation problem (Xue *et al.* 2015). Because the conventional buffer-zone method for VGI aims to capture the error band of the reference and calculate the ratio that a VGI object falls within it, it is essential to equip the measure with a rigorous statistical foundation. Using the ME model, we can estimate the confidence interval of the error band with respect to each point on the road, and the confidence intervals corresponding to all points together form the statistically defined buffer zone for the assessment of positional accuracy. Since the theoretically derived statistical buffer zone is an abstract set whose practical realization is difficult, we construct an approximation method to estimate the true envelope. The buffer zone such obtained possesses a probability (namely the confidence level) useful to the assessment of positional accuracy. The positional accuracy and the related probability together forms a natural measure of positional quality in VGI.

In brief, the aim of this research is to provide a more reasonable, objective and reliable measure for positional quality assessment in VGI with a sound statistical basis. The basic building block is the statistically defined buffer zone. In what follows, we first discuss in section 2 the problem of the conventional buffer zone concept in VGI and then construct the statistically defined buffer zone, together with a procedure for its practical implementation in positional accuracy assessment. With that, we construct an uncertainty measure of positional quality in section 3 and substantiate the conceptual arguments with OSM applications. We then conclude the paper with a summary and direction for further research in section 4.

2. Statistically defined buffer-zone method

2.1. *The basic Measurement error (ME) model*

From the modeling point of view, errors in GIS database can be reckoned mainly as system error and random error (Leung *et al.* 2004a). In the processing of geo-referenced data, random error, which is also called measurement error, is impossible to avoid entirely (Wolf and Ghilani 1997). The basic ME model aims to determine the error structures and location coordinates of objects in geo-referenced databases (Leung *et al.* 2004a).

The basic ME model can be simply expressed as:

$$\begin{cases} \boldsymbol{W} = f(\boldsymbol{V}), \\ \boldsymbol{V} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}), \end{cases} \tag{1}$$

where f is a transfer function, $\boldsymbol{W}$ the indirect measurement-value vector obtained by f, $\boldsymbol{V}$ the random measurement-value vector, $\boldsymbol{\mu}$ the true value vector, and $\boldsymbol{\varepsilon}$ the random ME vector with zero mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\varepsilon}$. The transfer function f is case specific. Through the model in (1), the uncertainty in $\boldsymbol{W}$ can be obtained from the uncertainty in $\boldsymbol{\varepsilon}$ because the distribution of $\boldsymbol{W}$ can be theoretically calculated when the distribution of $\boldsymbol{\varepsilon}$ is known (Herrador *et al.* 2005).

2.2. *The statistically defined buffer zone*

The mapped road networks are approximations of the real-world counterparts, and the same real-world road network may be represented differently in the VGI and the reference because of different conceptualizations and professionalisms of the data contributors. In general, the map road in the reference and the corresponding road in the VGI may be represented or captured differently. The two may have different topological structures. In general, a road may compose of several straight line segments with each having a start and end point. However, there may be variations. For example, the same real-world road is mapped by one road line segment in the reference data, Figure 1 (c), but five road line segments in the OSM (Figure 1 (a)). To better map the road network, one may adjust the positions of the start and end points of the road line segments in the OSM (Figure 1 (b)) or increase the number of road line segments with more vertices measured in the reference (Figure 1 (d)).

Figure 1 insert here with the caption: Different versions of the same road network mapped within OSM and reference data.

The buffer-zone method was proposed to evaluate the accuracy of a coastline (Goodchild and Hunter 1997), and it has also been applied to assess the accuracy of OSM road networks (Haklay *et al.* 2010, Goodchild and Li 2012). It is a natural idea to use the buffer-zone method to assess positional accuracy of VGI because the reference data are inaccurate, and the true value is not known. The conventional buffer-zone method treats the reference as data of higher quality. It ignores the measurement errors of the reference data and treats the measurement values of the reference data as the corresponding true values of the VGI. For illustration, Figure 2 shows the relationship between the true location, the measurement and the measurement error.

Figure 2 insert here with the caption: An illustration of the true location, the measurement and the measurement errors of the endpoints of a road line segment.

For the corresponding roads in the VGI and the reference, the buffer zone should be constructed on each road line segment of the reference in order to assess the positional accuracy of the corresponding road in the VGI. According to such conceptualization, when we employ the buffer-zone method, we need a reference which is usually a map produced, for example, by the National Mapping Agency (NMA). In general, we add a buffer of radius R on the road line segments of the reference map to evaluate the percentage of the OSM road network that falls within the buffer and treat it as its positional accuracy (Figure 3). Specifically, it is obtained as

$$positional\ accuracy = \frac{OSM\ road\ falling\ within\ the\ buffer}{total\ OSM\ road}. \tag{2}$$

However, the method has a serious problem. For example, when we assess the positional accuracy of the OSM road networks 1 and 2 in Figure 1 by the conventional buffer-zone method with the specified buffer radius R_1 and the reference road network 1, the positional accuracy of OSM road network 2 is higher than that of OSM road network 1 (as shown in Figure 3 (a) and (b), respectively). However, OSM road networks 1 and 2 will have the same positional accuracy (as shown in Figure 3 (b) and (c)) when the buffer radius is set for R_2 ($R_2 > R_1$). Apparently, different buffer

radius lead to different assessment results of the same road network, making the assessment of positional accuracy arbitrary, uncertain and unreliable. In addition, this problem also exists when the OSM and reference road networks have the same topology, e.g., the same real-world road is mapped by five road line segments in both the OSM and the reference (see Figure 1 (b) and (d)). Figure 3 (d), (e) and (f) show that when we assess positional accuracy of OSM road networks 1 and 2 by the conventional buffer-zone method and the reference road network 2, the situation is the same as that using reference road network 1. Therefore, in the worst case, one can manipulate the assessment result by subjectively specifying a desirable buffer radius in the conventional buffer zone method. In fact, the measurement errors of VGI are heterogeneous. Thus it is inappropriate to ignore the impact of the measurement errors of the reference data on the positional accuracy of VGI.

Figure 3 insert here with the caption: Concept and use of buffer zone in VGI positional accuracy assessment.

The buffer-zone method proposed by Goodchild and Hunter (1997) has not been evaluated theoretically and experimentally. To determine the size of the buffer radius and to give an appropriate theoretical basis for the rigorous analysis of positional accuracy, we need a formal basis on which the buffer can be statistically determined. With respect to error analysis in vector-based GIS, Leung *et al.* (2004a,b,c,d) have constructed a general framework for error analysis in measurement-based GIS. However, it has not been employed in VGI quality assessment, though the potential is there. There are indeed more complicated situations in VGI where the error distribution is totally random (Haklay *et al.* 2010), but the buffer is created around the reference map which is a conventional GIS product. The error analysis framework by Leung *et al.* (2004a,b,c,d) can thus serve as a theoretical basis to statistically determine the shape and size of a buffer. For easy reference, all key variables used in this paper are listed in Table 1.

Table 1 insert here with the caption: The key variables and corresponding explanation.

A road line segment is totally determined and controlled by the endpoints. Suppose the endpoints are n_0 and n_1, and the corresponding random measurement-value vectors are $\boldsymbol{V}_0$ and $\boldsymbol{V}_1$, respectively. According to the ME model in (1), they can be expressed as the sum of the true value vector $\boldsymbol{\mu}$ and the random ME vector $\boldsymbol{\varepsilon}$ with expectation $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_\varepsilon$. That is,

$$\boldsymbol{V} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon), \tag{3}$$

where $\boldsymbol{V} = (\boldsymbol{V}_0^{\mathrm{T}}, \boldsymbol{V}_1^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_0^{\mathrm{T}}, \boldsymbol{\mu}_1^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_0^{\mathrm{T}}, \boldsymbol{\varepsilon}_1^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Sigma}_\varepsilon = \begin{bmatrix} \boldsymbol{\Sigma}_{\varepsilon_{0,0}} & \boldsymbol{\Sigma}_{\varepsilon_{0,1}} \\ \boldsymbol{\Sigma}_{\varepsilon_{1,0}} & \boldsymbol{\Sigma}_{\varepsilon_{1,1}} \end{bmatrix}$. Then for any point n_λ on the road line segment, its indirect measurement-value vector $\boldsymbol{V}_\lambda$ is a linear transformation of $\boldsymbol{V}$:

$$\boldsymbol{V}_\lambda = ((1-\lambda)\mathbf{I}_2, \lambda\mathbf{I}_2)\boldsymbol{V}, \tag{4}$$

where $\lambda \in [0, 1]$ and $\mathbf{I}_2$ is a second-order identity matrix. The corresponding true value vector $\boldsymbol{\mu}_\lambda$, which is unknown, can be derived as the expectation of the indirect

measurement-value vector $\boldsymbol{V}_\lambda$:

$$\boldsymbol{\mu}_\lambda = E\boldsymbol{V}_\lambda = ((1-\lambda)\mathbf{I}_2, \lambda\mathbf{I}_2)\boldsymbol{\mu}, \tag{5}$$

and the corresponding random ME vector $\boldsymbol{\varepsilon}_\lambda$ can be derived as the difference between the indirect measurement-value vector $\boldsymbol{V}_\lambda$ and the true value vector $\boldsymbol{\mu}_\lambda$:

$$\boldsymbol{\varepsilon}_\lambda = \boldsymbol{V}_\lambda - \boldsymbol{\mu}_\lambda = ((1-\lambda)\mathbf{I}_2, \lambda\mathbf{I}_2)\boldsymbol{\varepsilon}. \tag{6}$$

Because the expectation and covariance of $\boldsymbol{\varepsilon}$ are known, we can derive the expectation $E\boldsymbol{\varepsilon}_\lambda$ and covariance $\boldsymbol{\Sigma}_{\varepsilon_\lambda}$ of $\boldsymbol{\varepsilon}_\lambda$ as:

$$E\boldsymbol{\varepsilon}_\lambda = ((1-\lambda)\mathbf{I}_2, \lambda\mathbf{I}_2)E\boldsymbol{\varepsilon} = \mathbf{0}, \tag{7}$$

$$\boldsymbol{\Sigma}_{\varepsilon_\lambda} = ((1-\lambda)\mathbf{I}_2, \lambda\mathbf{I}_2)\boldsymbol{\Sigma}_\varepsilon((1-\lambda)\mathbf{I}_2, \lambda\mathbf{I}_2)^{\mathrm{T}}, \tag{8}$$

Thus, we have

$$\boldsymbol{V}_\lambda = \boldsymbol{\mu}_\lambda + \boldsymbol{\varepsilon}_\lambda,\ \boldsymbol{\varepsilon}_\lambda \sim (\mathbf{0}, \boldsymbol{\Sigma}_{\varepsilon_\lambda}), \tag{9}$$

where $\boldsymbol{V}_\lambda$ represents the statistically derived measurement-value vector at point n_λ, with $\boldsymbol{\mu}_\lambda$ being the corresponding true-value vector, and $\boldsymbol{\varepsilon}_\lambda$ being the ME vector having expectation $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_{\varepsilon_\lambda}$. The conventional buffer-zone method uses a subjective buffer radius to assess positional accuracy because it assumes that the random ME vectors of each point are the same while in general it is not the case. Such assumption and the subjectivity in determining the buffer radius thus lead to the loss of information intrinsically carried by the random ME vectors of the endpoints, causing a problematic, weak and sometimes contradictory assessment of positional accuracy. Indeed, there is a probability distribution characterizing each measurement-value vector. If we could involve such information, we could improve the assessment of positional accuracy. In brief, the buffer radius should be determined by the structure of the measurement error of the endpoints.

When we want to measure a specific quantity, we usually make repeated measurements to estimate the quantity by the mean. Here the repeated measurements can be considered as the samples of the random measurement-value vector, the mean can be considered as the estimation of the true value vector and the sample covariance can be considered as the covariance estimation of ME. These repeated measurements can help us understand the corresponding random measurement-value distribution. Therefore, the problem of deciding on the accuracy of a random measurement-value vector is converted into the problem of statistical inference, i.e. the interval estimation. Suppose the measurement error of the reference road endpoints n_0 and n_1 are normally distributed, e.g., under the central limit theorem (Heuvelink 1998). For the point n_λ, we have

$$\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{-1/2}\boldsymbol{\varepsilon}_\lambda \sim N_2(\mathbf{0}, \mathbf{I}_2), \tag{10}$$

where $\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{-1/2}$ is the inverse of $\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{1/2}$ satisfying $\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{1/2}(\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{1/2})^{\mathrm{T}} = \boldsymbol{\Sigma}_{\varepsilon_\lambda}$. We can see that the two components of $\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{-1/2}\boldsymbol{\varepsilon}_\lambda$ are independent standard normal random variables.

Thus, we have

$$(\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{-1/2}\boldsymbol{\varepsilon}_\lambda)^{\mathrm{T}}(\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{-1/2}\boldsymbol{\varepsilon}_\lambda) = \boldsymbol{\varepsilon}_\lambda^{\mathrm{T}}\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{-1}\boldsymbol{\varepsilon}_\lambda \sim \chi_2^2, \tag{11}$$

where χ_2^2 is a Chi-square statistic with 2 degrees of freedom. As a result, the confidence interval of the random measurement-value vector $\boldsymbol{V}_\lambda$ with confidence level $1-\alpha$ can be constructed as

$$U_\lambda^\alpha = \{\boldsymbol{V}_\lambda : \boldsymbol{V}_\lambda = (v_{\lambda x}, v_{\lambda y})^{\mathrm{T}}, (\boldsymbol{V}_\lambda - \boldsymbol{\mu}_\lambda)^{\mathrm{T}}\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{-1}(\boldsymbol{V}_\lambda - \boldsymbol{\mu}_\lambda) \le \chi_{2,\alpha}^2\}, \tag{12}$$

since $P(\boldsymbol{V}_\lambda \in U_\lambda^\alpha) = P[(\boldsymbol{V}_\lambda - \boldsymbol{\mu}_\lambda)^{\mathrm{T}}\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{-1}(\boldsymbol{V}_\lambda - \boldsymbol{\mu}_\lambda) \le \chi_{2,\alpha}^2] = P[\boldsymbol{\varepsilon}_\lambda^{\mathrm{T}}\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{-1}\boldsymbol{\varepsilon}_\lambda \le \chi_{2,\alpha}^2] = 1-\alpha$, where $\chi_{2,\alpha}^2$ is the upper α-quantile of the Chi-square distribution χ_2^2 with 2 degrees of freedom. The confidence interval U_λ^α (i.e. the ellipse in Figure 4) means the region within which the measurement-value vector will fall with probability $1-\alpha$. Now, the positional accuracy of VGI should be assessed by examining whether the measurement-value vector of each road point will fall within the relative ellipse and the statistically defined buffer zone is the set of all these ellipses

$$\begin{aligned} U^\alpha = \{&\boldsymbol{V}_\lambda : \textit{there is a } \lambda, 0 \le \lambda \le 1, \textit{such that} \\ &(\boldsymbol{V}_\lambda - \boldsymbol{\mu}_\lambda)^{\mathrm{T}}\boldsymbol{\Sigma}_{\varepsilon_\lambda}^{-1}(\boldsymbol{V}_\lambda - \boldsymbol{\mu}_\lambda) \le \chi_{2,\alpha}^2\}, \end{aligned} \tag{13}$$

where $\boldsymbol{V}_\lambda = (1-\lambda)\boldsymbol{V}_0 + \lambda\boldsymbol{V}_1$. In the statistically defined buffer zone, the reference data generating a region within each point can be treated as the corresponding measurement-value vector with probability $1-\alpha$. Essentially the union of all error ellipses, U^α, could comprise all possible lines that one would get when repeatedly measuring the start and end points with random measurement error and connecting with straight lines all the point pairs within the $1-\alpha$ confidence interval. Then, the positional accuracy of VGI assessed by the statistically defined buffer-zone method aims to characterize whether the measurement-value vectors of VGI fall within the proper region derived from the reference data, i.e. it treats the proportion of the VGI road length, falling within the statistically defined buffer zone, in the total VGI road length as the relative positional accuracy.

Figure 4 insert here with the caption: Error ellipses of two endpoints and an intermediate point.

2.2.1. The simplified buffer-zone envelope

In practice, it is usually difficult to use U^α directly because it is a family of plane ellipses. For the practical application of our statistically defined buffer-zone method, the envelope of the ellipse family must be captured to represent the buffer-zone. Mathematically, an envelope means a curve, surface, or higher-dimensional object tangent to a given family of lines or curves or surfaces or higher-dimensional objects, respectively. Thus, the envelope here means a curve tangent to the ellipse family. Obviously, the curve that connects the farthest points from the ellipse to the reference road is the envelope we want (see Figure 5). For each error ellipse, e.g. $U_\lambda^\alpha, \lambda \in [0,1]$, its farthest points to the reference road are the points with the same slope of the reference road,

which is obtained by solving the simultaneous equations:

$$\begin{cases} (\boldsymbol{V}_\lambda - \boldsymbol{\mu}_\lambda)^{\mathrm{T}} \boldsymbol{\Sigma}_{\varepsilon_\lambda}^{-1} (\boldsymbol{V}_\lambda - \boldsymbol{\mu}_\lambda) = \chi^2_{2,\alpha}, \\ \dfrac{\mathrm{d}v_{\lambda y}}{\mathrm{d}v_{\lambda x}} = (v_{1y} - v_{0y})/(v_{1x} - v_{0x}), \end{cases} \tag{14}$$

where $(v_{0x}, v_{0y})^{\mathrm{T}} = \boldsymbol{V}_0$, $(v_{1x}, v_{1y})^{\mathrm{T}} = \boldsymbol{V}_1$ and $(v_{\lambda x}, v_{\lambda y})^{\mathrm{T}} = \boldsymbol{V}_\lambda$. In fact, consider the standard parametric representation of an ellipse $(x, y) = (\mathrm{acos}t, \mathrm{bsin}t), 0 \leq t \leq 2\pi$, and a general line function $\mathrm{A}x + \mathrm{B}y + \mathrm{C} = 0$, the distance from any point of the ellipse to the line is

$$d = \frac{|\mathrm{Aacos}t + \mathrm{Bbsin}t + \mathrm{C}|}{\sqrt{\mathrm{A}^2 + \mathrm{B}^2}}, 0 \leq t \leq 2\pi, \tag{15}$$

and equation (15) reaches the extremum when $\mathrm{Aasin}t = \mathrm{Bbcos}t$. Then, the slope of the ellipse

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}t} \Big/ \frac{\mathrm{d}x}{\mathrm{d}t} = -\frac{\mathrm{bcos}t}{\mathrm{asin}t} \tag{16}$$

shows that the slope of the extremum point is $-\mathrm{A}/\mathrm{B}$, which is exactly the slope of the line function. With enough points obtained from equation (14), the curve connecting these farthest points from the ellipses to the reference road line segment can accurately capture the envelope. However, in practice we cannot afford to calculate for infinite number of points. Thus, the envelope needs to be fitted by a finite number of selected points and represented by a general expression, such as the Cubic Spline Interpolation or Fourier Series. For practical applications, a line in the GIS is produced by the straight line connecting the input points. So, an envelope in GIS applications can be fitted by serval line segments. What remains to be determined is the specific number of points to be generated to produce an acceptable envelope.

Figure 5 insert here with the caption: Approximate envelope in buffer zone. The envelope in bold is the statistically defined buffer-zone; $\mathrm{c}_1, \mathrm{c}_2, \mathrm{c}_3$ and c_4 are the farthest points of two end-points. L is the distance between the midpoint of an approximation and the midpoint of the relative true envelope.

Undoubtedly, the farthest points, $\mathrm{c}_1, \mathrm{c}_2$ and $\mathrm{c}_3, \mathrm{c}_4$, from the ellipses to the reference road line segment obtained by equation (14) with respect to the two end-points are necessary for the close approximation of the envelope because the parts of the error ellipses from c_1 to c_2 and from c_3 to c_4 belong to the true envelope (see Figure 5). As the part of the envelope from c_1 to c_3 is symmetric to the part from c_2 to c_4, we only use the part from c_2 to c_4 for our discussion. Consider in the most straightforward way, we directly connect c_2 with c_4 as an approximation, l_{c_2,c_4}, to the bottom part of the envelop in Figure 6 (a), i.e. it may be used as the envelope by the conventional buffer-zone method (It should be noted that the conventional buffer-zone method does not ever use error ellipse to define the radius). The approximation is obviously not an acceptable approximate envelope in this case because there is a visible gap between the approximation (in blue) and the true envelope, especially at the midpoint (see Figure 6 (d)).

Figure 6 insert here with the caption: Illustration of buffer-zone envelope approximation. (a) l_{c_2,c_4} is the approximate envelope connecting two points c_2 and c_4. (b) l_{c_2,c_5}, l_{c_5,c_4} is the improved approximate envelope connecting three points c_2, c_5 and c_4. (c) l_{c_2,c_6}, l_{c_6,c_5}, l_{c_5,c_7}, l_{c_7,c_4} is the further improved approximate envelope connecting five points c_2, c_6, c_5, c_7 and c_4. (d) The comparison of the approximation precision of the approximate envelope in (a), (b) and (c).

Based on the approximate envelope l_{c_2,c_4} in Figure 6 (a) that connects c_2 with c_4 directly, we construct the acceptance level (AL) index to measure the acceptance level of the approximate envelope as follows:

$$AL = 1 - \frac{MPE}{\mathrm{L}}, \tag{17}$$

where L is as indicated in Figure 5 which is the distance between the midpoint of l_{c_2,c_4} and the corresponding point of the true envelope, and MPE is the mean position error between the midpoint of the current two adjacent points, which are used for approximation, and the corresponding point on the true envelope. In fact, the AL index indicates how much error of the approximate envelope l_{c_2,c_4} (in Figure 6 (a)) has been eliminated by the current approximate envelope. With a smaller MPE, the current approximate envelope will have a larger AL which indicates that it has further eliminated the error than the previous approximate envelope, leading to a better approximate envelope. For the initial approximation, l_{c_2,c_4}, its AL obviously is 0 because its MPE is exactly L. It means that the farthest points, c_2 and c_4, used for approximation are not enough to produce an acceptable approximate envelope. Thus, we further select the point on the true envelope which is used to calculate MPE to produce a better approximate envelope as shown in Figure 6 (b). Now, its $AL = 1 - \frac{L_1+L_2}{2\mathrm{L}}$ (see Figure 6 (d)), where MPE is $\frac{L_1+L_2}{2}$. Starting with the initial points c_2 and c_4, we can calculate the acceptance level of the current approximate envelope. To obtain a closer approximation to the true envelope, we can perform the above procedure iteratively until an acceptable approximate envelope is obtained. That is, to better the first-step approximation (in Figure 6 (a)), we select the point on the true envelope which will be used to calculate MPE in next step (as shown in Figure 6 (b)). The above procedure will be iteratively applied as shown in Figure 6 (c). Apparently, the AL index will get to be sufficiently large when more points are involved throughout the iterative process. Specifically, we specify a level of acceptance AL to which the iterative procedure is required to achieve for the final approximate envelope.

Example 1. Assumed that there are two reference observations, $\boldsymbol{V}_0$ and $\boldsymbol{V}_1$, of the road endpoints, n_0 and n_1, where $\boldsymbol{V} = (\boldsymbol{V}_0^{\mathrm{T}}, \boldsymbol{V}_1^{\mathrm{T}})^{\mathrm{T}} = (0, 0, 6, 4)^{\mathrm{T}}$. The related ME is $\boldsymbol{\varepsilon} \sim N_4(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$ which is a four-dimensional normal random variable with expectation $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_\varepsilon$. Let us consider three different correlation situations of ME between two endpoints to illustrate how the statistically defined buffer zone is obtained under each situation with 0.99 acceptance level:

Correlation situations:

a. Positive: $\boldsymbol{\Sigma}_{\varepsilon_{0,0}} = \boldsymbol{\Sigma}_{\varepsilon_{1,1}} = \sigma^2\mathbf{I}^2,\ \sigma = 0.4,\ \boldsymbol{\Sigma}_{\varepsilon_{0,1}} = \boldsymbol{\Sigma}_{\varepsilon_{1,0}} = \begin{bmatrix} 0.16 & 0.018 \\ 0.018 & 0.16 \end{bmatrix}$;
b. Negative: $\boldsymbol{\Sigma}_{\varepsilon_{0,0}} = \boldsymbol{\Sigma}_{\varepsilon_{1,1}} = -\boldsymbol{\Sigma}_{\varepsilon_{0,1}} = -\boldsymbol{\Sigma}_{\varepsilon_{1,0}} = \sigma^2\mathbf{I}^2,\ \sigma = 0.1$;
c. Independent: $\boldsymbol{\Sigma}_{\varepsilon_{0,0}} = \boldsymbol{\Sigma}_{\varepsilon_{1,1}} = \sigma^2\mathbf{I}^2,\ \sigma = 0.1,\ \boldsymbol{\Sigma}_{\varepsilon_{0,1}} = \boldsymbol{\Sigma}_{\varepsilon_{1,0}} = \mathbf{0}$.

The results are depicted in Figure 7. It can be observed that the simplified approximate envelope has a good performance in approximating the true envelope. In situation a, the approximate envelope connecting five equally spaced points, c_1, c_3, c_5, c_7 and c_9, can approximate the true envelope with an acceptance level higher than 0.99. In Figure 7 (a), the first-step approximation (in blue) has $AL = 0$, thus we further connect the initial points, c_1 and c_9, to the midpoint, c_5, of the true envelope to form the second-step approximation (in red). As the second-step approximation has $AL = 0.954$ which is smaller than the pre-specified acceptance level 0.99, we again better the second-step approximation with the third-step approximation (in green) which has $AL = 0.9969$. As the acceptance level of the third-step approximation is higher than the pre-specified acceptance level, we finally use the third-step approximation to approximate the true envelope. Situation (c) is like the situation a (as shown in Figure 7 (c)). While in situation (b), the second-step approximation (in red) connecting, c_1 and c_3, to the midpoint of the true envelope, c_2, already has $AL = 1$, which means that we can produce an acceptable approximate envelope in two-step iteration (as shown in Figure 7 (b)).

Figure 7 insert here with the caption: Illustration of buffer zone approximation under three correlation situations. $(1 - \alpha = 0.95; AL = 0.99)$.

With the above practical realization of the statistical buffer zone stipulated in equation (13), the ratio that an OSM road network falls within the buffer zone of the reference road network is equivalent to the ratio that the OSM falls within the approximate envelope, which means

$$positional\ accuracy = \frac{OSM\ road\ falls\ within\ the\ approximate\ envelope}{total\ OSM\ road}. \quad (18)$$

The buffer zone obtained by the conventional method thus becomes a special case of our statistically defined buffer zone when we have equal buffer width for each point of the reference road. Therefore, the AL-index-based statistical buffer zone renders a more reasonable, objective and reliable assessment of positional accuracy with a strong statistical basis.

3. Positional uncertainty of OSM road networks

3.1. *Positional uncertainty*

Error and accuracy are used to assess the quality of digital positional data in geographic information science. The idea is borrowed to assess positional quality in VGI. In conventional GIS, positional measurements are made by the professionals, which means measurements can be modified by the measurement process. However, the positional measurements in VGI come from various sources, making their modifications by the measurement process difficult if not impossible (Flanagin and Metzger 2008). It is believed that positional uncertainty is more proper to describe the quality of VGI (Goodchild 2007a). From the statistical perspective, uncertainty can be characterized by the level of confidence, a probability argument, we have on an estimation. With positional accuracy defined in (18), we can build on it to construct a measure to characterize positional uncertainty with respect to the buffer zone. It is on this basis we propose the following characterization of positional uncertainty.

As discussed in section 2, the statistically defined buffer-zone method constructs a region within which all the points could be the corresponding measurement values with probability $1-\alpha$. Assessment results of positional accuracy obtained are related to the pre-specified confidence level, i.e. the region has different forms with different confidence levels. We thus need to take the confidence level into the characterization of uncertainty in quality assessment by using positional accuracy and the associated level of confidence to characterize positional uncertainty as follows:

$$positional\ uncertainty = (positional\ accuracy\ ;\ confidence\ level), \tag{19}$$

where positional accuracy is obtained from equation (18) and confidence level is specified in equation (13). Therefore, positional uncertainty in VGI can describe the precision of the results of positional accuracy assessment based on the statistically defined buffer zone. In summary, positional accuracy and positional uncertainty are two measures of positional quality of VGI. Positional accuracy can be assessed more accurately with a statistical buffer zone and positional uncertainty is the improvement of positional accuracy as it further incorporates the uncertainty embedded in the statistical buffer zone.

3.2. *Positional uncertainty assessment of the intersection in OSM road networks*

In section 2, we have analyzed positional uncertainty of the line segment of a road. It should be noted that a road network is defined (approximately) by the collection of line segments joining relevant points (Kiiveri 1997). To analyze positional uncertainty of a road network, we need an additional measure: positional uncertainty of an intersection. Intersections can be divided into two categories in terms of the way, direct and indirect, they are generated.

The direct intersection is generated by the corresponding measurements, it is the start or end points of four road line segments. Despite almost all the intersections in OSM are direct intersections, their positional uncertainty should be assessed according to the matching intersections in the reference map. With sufficient measurements, a reasonable region of the random measurement-value vector of a direct intersection in the reference map can be constructed on the basis of the statistically defined buffer-zone method, and the positional uncertainty of the corresponding intersection in OSM can then be assessed by equation (19).

However, the indirect intersection is just the intersection of two road line segments, its location is not determined by the direct measurements but the four start and end points of the two road line segments. Within such situation, the location of the indirect intersection is a nonlinear transformation of the related start and end points. Due to the nonlinearity of the transfer function f, we can only give an approximation to the ME of the indirect intersection. Leung *et al.* (2004a,b,c,d) used the first-order Taylor series expansion to replace the nonlinear transfer function. Xue *et al.* (2015) extended the work by using higher-order Taylor series expansion method. However, the statistically defined buffer zone provides an easy and direct way to capture positional uncertainty of the intersection.

Since the statistically defined buffer zone effectively captures positional uncertainty of a road line segment, positional uncertainty of the intersection of line segments can be captured by overlaying the associated statistically defined buffer zones of the relevant line segments. The approximation method aims to find the covariance matrix

of the corresponding random measurement error of the intersection, which we use to characterize positional uncertainty of the intersection. By connecting the selected points in turn, positional uncertainty of the intersection is represented by a closed area (Figure 8 (a)), and the buffer zone of its positional uncertainty can be obtained by overlaying the relevant road line segments.

Figure 8 insert here with the caption: Determination of positional uncertainty of the intersection of two road segments with simulated data. (a)The simplified statistically defined buffer zone for the intersection road line segments. (b) The simulated endpoints and related intersection points. (c) The simulated points falling within the simplified statistically defined buffer zones. (d) The positional uncertainty of the intersection point captured by the simplified statistically defined buffer zone. $(1-\alpha = 0.95; AL = 0.99)$.

Example 2. Assumed that there are two intersecting road line segments L_a and L_b, with endpoints measurement-value vectors $\boldsymbol{V}^a = (0, 0, 6, 4)^{\mathrm{T}}, \boldsymbol{V}^b = (2, -2, 3, 7)^{\mathrm{T}}$, respectively, and having the same ME vector $\boldsymbol{\varepsilon} \sim N_4(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$, where $\boldsymbol{\Sigma}_{\varepsilon_{0,0}} = \boldsymbol{\Sigma}_{\varepsilon_{1,1}} = \sigma^2\mathbf{I}_2, \sigma = 0.1, \boldsymbol{\Sigma}_{\varepsilon_{0,1}} = \boldsymbol{\Sigma}_{\varepsilon_{1,0}} = \mathbf{0}$.

The indirect intersection measurement-value vector $\boldsymbol{V}_i$ is a nonlinear transformation of four endpoints $\boldsymbol{V}^a$ and $\boldsymbol{V}^b$ (Leung *et al.* 2004a). It is generally difficult to directly use the basic ME model (Leung *et al.* 2004c). However, the simplified statistically defined buffer zone provides a way to capture the positional uncertainty about the intersection. By 1000 numerical simulations, positional uncertainty of the intersection is shown in Figure 8 (b). The simplified statistically defined buffer zone can capture all the positional uncertainty of the intersecting road line segments (see Figure 8 (c)). And for the intersection, its positional uncertainty can be captured by the overlaid region of the simplified statistically defined buffer zones on the road line segments (see Figure 8 (d)). It should be noted that there are some points falling outside the overlaid region because the statistically defined buffer-zone is derived with 0.95 confidence level.

3.3. *OSM Experiment*

For a better understanding of the concept and computation of our proposed measure of positional uncertainty, we use it to examine the quality of the road data in the OSM of the Kowloon peninsula of Hong Kong. The experimental data are extracted from china-latest.osm.pbf, which is a file downloaded from Geofabrik containing all OSM data up to 2018-03-26T20:43:02Z. The reference is the iC1000 Digital Land Boundary Map prepared by the Land Information Centre, Survey and Mapping Office at the Lands Department (the official land information office) of the Hong Kong Special Administrative Region.

3.3.1. Positional uncertainty assessment of road line segment

To illustrate in details the practical implementation of the statistically defined buffer zone, we used Nathan Road, a major road, as a test case first. We collected all the Kowloon OSM nodes with a tag "name = Nathan Road" and put them in ArcGIS to generate the road for comparison. We, then, projected the data onto the Hong Kong 1980 Grid coordinate system as depicted in Figure 9 (b). For illustration, Figure 9 (a) and (c) show a good road section of the OSM data (when the OSM data fall within the reference boundary) and a bad road section of the OSM data (when part of the

OSM data intersect or fall outside the reference boundary), respectively. It shows that the positional error in the OSM road network is distributed unevenly. Thus, we cannot correct the data through a simple displacement or rotation. Furthermore, the conventional buffer-zone method cannot characterize the error distribution through a pre-defined buffer radius whereas the statistically defined buffer zone can because the buffer zones are constructed through the reference itself while the conventional buffer-zone method considers the OSM data to subjectively set the buffer radius. Because the road map in the reference data use boundaries to represent road but the OSM road data are captured in two lines (e.g., the thin red lines in Figure 9 (a) and (c)) running in opposite directions along a road, we need to use the corresponding center lines to represent the road in both the reference and the OSM and use the positional uncertainty of the OSM centerline to represent the positional uncertainty of the OSM data.

Figure 9 insert here with the caption: Road representation in the reference map and OSM. (a)Blown-up view of the upper box demarcated in (b), which is the good section of the OSM data. (b)Original data under Hong Kong 1980 Grid coordinate system. (c)Blown-up view of the lower box demarcated in (b), which is the bad section of the OSM data that intersect the reference boundary. (d), (e) and (f)are the center lines of the road in (a), (b) and (c), respectively. The red lines are the OSM data and the green lines are the reference data.

The Figures 9 (d), (e) and (f) show the relative center lines with respect to Figures 9 (a), (b) and (c). We merge the two thin red OSM road lines and the boundaries of the reference road into one center line through ArcGIS, respectively. Clearly, the center line of the good section of the OSM has a better positional quality than that in the bad section because the OSM center line in Figure 9 (d) completely overlaps with the reference center line while there is an apparent offset of the two center lines in Figure 9 (f). For the good section, no matter how small the buffer radius is specified by the conventional method, positional quality is extremely high because the OSM center line completely overlaps with the reference line. For the same reason, similar result is obtained by our proposed method. For the bad section of the road, however, different buffer radius subjectively specified by the conventional method lead to different positional accuracies as shown in Figure 10 (a). The OSM center line (in red) falls within the 4-m buffer (in green) but falls outside the 2-m buffer (in blue). Therefore, using two subjectively specified radius of the buffer zone will give two different conclusions about the positional uncertainty. It is thus confusing whether we should use the 2-m buffer zone or the 4-m buffer zone to assess the positional accuracy of the road in such case. Apparently, such specification of buffer radius by the conventional method is unacceptable. On the other hand, the statistically defined buffer-zone (in yellow) as shown in Figure 10 (b) is constructed from the data themselves. It aims to capture the true position of the road and the buffer width is independent of the OSM data. It should be noted that the covariance used in this section is estimated by the simulated measurements made with reference to the measurement on each node.

Figure 10 insert here with the caption: Illustration of the paradox arises from the subjectively specified buffer zone. (a) The conventionally specified buffer zone with a 2-m buffer (in blue) and 4-m buffer (in green). (b) The statistically defined buffer zone (in yellow) constructed through simulated measurements.

Figure 11 insert here with the caption: Contrasting the assessments of positional uncertainty via the conventional and the statistical buffer zones. (a) The OSM road center falls within the conventional 4-m buffer zone (in green) but the right side of the OSM road intersects the reference boundary, the statistically defined buffer zone is completely covered by the conventional 2-m buffer zone (in blue). (b) The OSM road center falls within the statistically defined buffer zone (in yellow) but outside the conventional 2-m buffer zone (in blue). The solid red line is the OSM road center and the dotted red lines are the left and right sides of the OSM road, respectively.

The assessment results of positional accuracy with respect to the three types of buffer zones in Figure 10 are shown in Table 2. It shows that the statistically defined buffer-zone method gives the positional accuracy like the conventional 2-m buffer-zone, but the conventional 4-m buffer-zone produces a much higher positional accuracy assessment. Table 2 cannot display the superiority of statistically defined buffer-zone method but reveals a problem in using conventional buffer-zone method, i.e. which buffer zone, 2-m or 4-m, should we use? They are both subjectively defined but deliver completely different results. The conventional method will probably use the 4-m buffer zone because it will pre-define the positional accuracy as about 0.95. To further substantiate the issue, let's look at Figure 11. Here, Figure 11 (a) shows a problem with the 4-m buffer zone. This part of the OSM road has a poor positional quality because the right side of the OSM road (dotted line) intersects the boundary of the reference road. However, the related center road (red line) falls within the 4-m buffer, and in this way, the 4-m buffer zone gives the OSM road a high positional accuracy which is in contradiction with the ground truth. In contrast, the 2-m buffer zone gives this OSM road section a poor positional accuracy which is consistent with the actual situation because the related center line falls outside the buffer zone. In this sense, the 2-m buffer zone more faithfully assesses the positional quality of this part of the OSM road than the 4-m buffer zone. On the other hand, when we look at Figure 11 (b), there is also a problem using the 2-m buffer zone. This part of the OSM road does not have a bad positional quality as they are in a reasonable position in a crossroad, but the related road center falls outside the 2-m buffer zone. It means that the 2-m buffer zone gives this part of the OSM road a poor positional quality which is in contradiction with the ground truth. In contrast, the 4-m buffer zone gives this part of the OSM road a good positional accuracy which is consistent with the actual situation. So, we cannot give a convincing answer to the above paradox if the radius of the buffer zone is subjectively defined.

Table 2 insert here with the caption: Assessment results of positional accuracy with three kinds of buffer zone.

Such problem, however, does not exist in the use of the statistically defined buffer zone (as shown in Figure 11). The statistically defined buffer-zone method assigns probability to the reference road, giving the positional accuracy assessment a probabilistic and objective characterization that builds on the data themselves. According to equation (19), the positional uncertainty of Nathan Road in the OSM is:

$$positional\ uncertainty = (0.6709\ ;\ 0.95).$$

This gives an informative assessment of positional uncertainty in VGI. That is, we are 95% confident that the positional accuracy is 0.6709. Because the buffer zone is

constructed through equation (13) where $\alpha = 0.05$, it means that the probability of the OSM road measurement-value vectors falling within the buffer zone is 0.95. This is a very important information about the assessment result because it describes the precision of the result with a specific buffer radius (see Figure 8 (d)). According to equation (14), the buffer radius is negatively related to the precision, meaning that we must make a trade-off between the two. In short, positional accuracy only provides the assessment of accuracy, 0.6709, without the level of precision. Positional uncertainty, on the other hand, provides both accuracy and precision so that we know the level of confidence in the assessment, i.e. the buffer radius.

3.3.2. Positional uncertainty assessment of road networks

The Positional uncertainty of road networks can be characterized by the road line segments and intersection point. In section 3.1, we have investigated the positional uncertainty of the road line segment, and the positional uncertainty of the intersection point is obtained by overlaying the statistically defined buffer zones of the intersecting roads. In this section, we illustrate the positional uncertainty of the road network in the Kowloon OSM data obtained by the conventional buffer-zone method and the statistically defined buffer-zone method. Figure 12 is the positional accuracy assessment results based on different subjectively specified buffer widths. It shows that 6 m is the buffer width whose positional accuracy is closest to 95%. Since the quality distribution in VGI is heterogeneous (Goodchild and Li 2012), the test area is clipped into a 40×40 fishnet to represent the quality difference.

Figure 12 insert here with the caption: The positional accuracy assessment results with different pre-specified buffer widths.

Figure 13 shows the grid-cell-based distribution map of positional accuracy assessment of the OSM Kowloon road network. The positional accuracy assessment results are evenly divided into five levels, the darker the color the higher the positional accuracy. The blank grid cell means that there is no standard reference road in it. The conventional buffer-zone method used in Figure 13 (a) has a pre-specified buffer width of 6 m, because we can make the test data reach 95% positional accuracy. However, the related positional accuracy distribution map cannot provide the information on the heterogeneous quality among the grids because the positional accuracy in most of the cells are beyond 0.8. Since the conventional buffer-zone method aims to make the test data achieve 95% positional accuracy, the buffer width must be made large enough before the grids with poor quality could get a proper positional accuracy assessment. This automatically takes away the quality difference among the cells. On the contrary, the positional accuracy distribution map based on the statistically defined buffer zone can clearly differentiate the quality difference among the cells. Since it is constructed from the data themselves, it can faithfully and precisely represent the quality of the data in each cell.

Figure 13 insert here with the caption: Distribution of the positional accuracy assessment results based on (a)The conventional buffer-zone method and (b)The statistically defined buffer-zone method.

It should be noted that the statistically defined buffer zone used in the experiment is derived with $\alpha = 0.05$, which means the positional accuracy assessment has a 95%

credibility. Therefore, for each grid in Figure 13 (b), the positional uncertainty gives the positional accuracy assessment with 0.95 confidence probability. Not only can we objectively assess the accuracy of the OSM road network but can also assess it with a level of confidence.

4. Discussion and conclusion

In this paper, we have discussed how to assess positional quality of road networks in VGI. We argue that the conventional buffer-zone method does not have a sound theoretical foundation and statistical basis for the construction of the buffer zone and the sound interpretation of the assessment results. To overcome this problem, we have built on the basic measurement-error model of Leung *et al.* (2004a) to give a statistical and experimental analysis of buffer-zone construction and application. However, the buffer zone such derived involves a family of ellipses which is difficult to obtain by direct calculation. For practical implementation, we have simplified the procedure to obtain an approximate envelope for the ellipse family through an iterative process based on the AL index. The approximate envelope is obtained by iteratively connecting adjacent points by a straight line until enough number of points is involved, i.e. a specified AL value is reached, in determining the envelope that can be used to replace the theoretically defined buffer zone. This greatly facilitates the use of the buffer-zone method in real-life positional quality assessment. Furthermore, the simplified statistically defined buffer zone helps us to capture the positional uncertainty of the intersection points directly. The statistically defined buffer-zone method therefore provides a sound statistical buffer-zone construction and applications. To better assess positional quality, we have also proposed a measure for positional uncertainty which is a pairwise measure that gives the assessment of positional accuracy and the associated level of confidence with respect to the buffer zone.

However, it should be noted that positional accuracy assessment and positional uncertainty is made with respect to a reference. While VGI is a crowdsourced data with real-time or near-real-time updates, the reference, on the other hand, may not be as frequently updated. Further research should investigate how to synchronize the two to make the assessment of positional quality even more reliable.

Acknowledgement

This project was supported by the the earmarked grant CUHK 14406514 of the Hong Kong Research Grants Council and the Vice-Chancellor's One-off Discretionary Fund of The Chinese University of Hong Kong.

References

Al-Bakri, M. and Fairbairn, D., 2010. Assessing the accuracy of crowdsourced data and its integration with official spatial datasets. *In*: *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Leicester, UK*. vol. 2023, 317320.

Arsanjani, J.J., *et al.*, 2015. Quality assessment of the contributed land use information from openstreetmap versus authoritative datasets. *In*: *Openstreetmap in giscience*. Springer, Cham, 37–58.

Ciepłuch, B., *et al.*, 2010. Comparison of the accuracy of openstreetmap for ireland with google maps and bing maps. *In*: *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*. University of Leicester, 337.

Coleman, D.J., 2013. Potential contributions and challenges of vgi for conventional topographic base-mapping programs. *In*: *Crowdsourcing geographic knowledge*. Springer, 245–263.

Flanagin, A.J. and Metzger, M.J., 2008. The credibility of volunteered geographic information. *GeoJournal*, 72 (3-4), 137–148.

Goodchild, M.F., 2007a. Beyond metadata: Towards user-centric description of data quality. *In*: *Proceedings, Spatial Data Quality 2007 International Symposium on Spatial Data Quality, June*. Citeseer, 13–15.

Goodchild, M.F., 2007b. Citizens as voluntary sensors: spatial data infrastructure in the world of web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2, 24–32.

Goodchild, M.F. and Hunter, G.J., 1997. A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, 11 (3), 299–306.

Goodchild, M.F. and Li, L., 2012. Assuring the quality of volunteered geographic information. *Spatial statistics*, 1, 110–120.

Haklay, M., *et al.*, 2010. How many volunteers does it take to map an area well? the validity of linus' law to volunteered geographic information. *The Cartographic Journal*, 47 (4), 315–322.

Haklay, M. and Weber, P., 2008. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7 (4), 12–18.

Helbich, M., *et al.*, 2013. Does altitude moderate the impact of lithium on suicide? a spatial analysis of austria. *Geospatial health*, 7 (2), 209–218.

Herrador, M.Á., Asuero, A.G., and González, A.G., 2005. Estimation of the uncertainty of indirect measurements from the propagation of distributions by using the monte-carlo method: An overview. *Chemometrics and intelligent laboratory systems*, 79 (1-2), 115–122.

Heuvelink, G.B., 1998. *Error propagation in environmental modelling with gis*. CRC Press.

Kiiveri, H.T., 1997. Assessing, representing and transmitting positional uncertainty in maps. *International Journal of Geographical Information Science*, 11 (1), 33–52.

Kounadi, O., 2009. Assessing the quality of openstreetmap data. *Msc geographical information science, University College of London Department of Civil, Environmental And Geomatic Engineering*.

Leung, Y., Ma, J.H., and Goodchild, M.F., 2004a. A general framework for error analysis in measurement-based gis part 1: The basic measurement-error model and related concepts. *Journal of Geographical Systems*, 6 (4), 325–354.

Leung, Y., Ma, J.H., and Goodchild, M.F., 2004b. A general framework for error analysis in measurement-based gis part 2: The algebra-based probability model for point-in-polygon analysis. *Journal of geographical systems*, 6 (4), 355–379.

Leung, Y., Ma, J.H., and Goodchild, M.F., 2004c. A general framework for error analysis in measurement-based gis part 3: Error analysis in intersections and overlays. *Journal of geographical systems*, 6 (4), 381–402.

Leung, Y., Ma, J.H., and Goodchild, M.F., 2004d. A general framework for error analysis in measurement-based gis part 4: Error analysis in length and area measurements. *Journal of Geographical Systems*, 6 (4), 403–428.

Senaratne, H., *et al.*, 2017. A review of volunteered geographic information quality assessment

methods. *International Journal of Geographical Information Science*, 31 (1), 139–167.

Van Exel, M., Dias, E., and Fruijtier, S., 2010. The impact of crowdsourcing on spatial data quality indicators. *In*: *Proceedings of the GIScience 2010 Doctoral Colloquium, Zurich, Switzerland*. 14–17.

Wolf, P.R. and Ghilani, C.D., 1997. *Adjustment computations: statistics and least squares in surveying and gis*. Wiley-Interscience.

Xue, J., Leung, Y., and Ma, J.H., 2015. High-order taylor series expansion methods for error propagation in geographic information systems. *Journal of Geographical Systems*, 17 (2), 187–206.

Zhou, Q., 2017. Rethinking the buffering approach for assessing openstreetmap positional accuracy. *In*: *International Cartographic Conference*. Springer, Cham, 435–448.

Table 1. The key variables and corresponding explanation.

Variables	Explanation
V	The random measurement-value vector of a location.
V^a	The random measurement-value vector of the location of road line segment a.
V_i	The random measurement-value vector of the location of the intersection.
μ	The true value vector of a location.
ε	The random measurement error of a location.
U_λ^α	The statistical buffer-zone on any point λ with the level of significance α.

Table 2. Assessment results of positional accuracy with three kinds of buffer zone.

Buffer-zone type	Length within buffer	Total length	Positional accuracy
Conventionally defined 2-m buffer zone	2311.37 m	3573.75 m	64.68%
Conventionally defined 4-m buffer zone	3377.06 m	3573.75 m	94.50%
Statistically defined buffer zone	2397.63 m	3573.75 m	67.09%

Figure 1. Different versions of the same road network mapped within OSM and reference data.

Figure 2. An illustration of the true location, the measurement and the measurement errors of the endpoints of a road line segment.

Figure 3. Concept and use of buffer zone in VGI positional accuracy assessment.

Figure 4. Error ellipses of two endpoints and an intermediate point.

Figure 5. Approximate envelope in buffer zone. The envelope in bold is the statistically defined buffer-zone; c_1, c_2, c_3 and c_4 are the farthest points of two end-points. L is the distance between the midpoint of an approximation and the midpoint of the relative true envelope.

Figure 6. Illustration of buffer-zone envelope approximation. (a) l_{c_2,c_4} is the approximate envelope connecting two points c_2 and c_4. (b) l_{c_2,c_5}, l_{c_5,c_4} is the improved approximate envelope connecting three points c_2, c_5 and c_4. (c) l_{c_2,c_6}, l_{c_6,c_5}, l_{c_5,c_7}, l_{c_7,c_4} is the further improved approximate envelope connecting five points c_2, c_6, c_5, c_7 and c_4. (d) The comparison of the approximation precision of the approximate envelope in (a), (b) and (c).

Figure 7. Illustration of buffer zone approximation under three correlation situations. ($1 - \alpha = 0.95; AL = 0.99$).

Figure 8. Determination of positional uncertainty of the intersection of two road segments with simulated data. (a)The simplified statistically defined buffer zone for the intersection road line segments. (b) The simulated endpoints and related intersection points. (c) The simulated points falling within the simplified statistically defined buffer zones. (d) The positional uncertainty of the intersection point captured by the simplified statistically defined buffer zone. ($1 - \alpha = 0.95; AL = 0.99$).

Figure 9. Road representation in the reference map and OSM. (a)Blown-up view of the upper box demarcated in (b), which is the good section of the OSM data. (b)Original data under Hong Kong 1980 Grid coordinate system. (c)Blown-up view of the lower box demarcated in (b), which is the bad section of the OSM data that intersect the reference boundary. (d), (e) and (f)are the center lines of the road in (a), (b) and (c), respectively. The red lines are the OSM data and the green lines are the reference data.

Figure 10. Illustration of the paradox arises from the subjectively specified buffer zone. (a) The conventionally specified buffer zone with a 2-m buffer (in blue) and 4-m buffer (in green). (b) The statistically defined buffer zone (in yellow) constructed through simulated measurements.

Figure 11. Contrasting the assessments of positional uncertainty via the conventional and the statistical buffer zones. (a) The OSM road center falls within the conventional 4-m buffer zone (in green) but the right side of the OSM road intersects the reference boundary, the statistically defined buffer zone is completely covered by the conventional 2-m buffer zone (in blue). (b) The OSM road center falls within the statistically defined buffer zone (in yellow) but outside the conventional 2-m buffer zone (in blue). The solid red line is the OSM road center and the dotted red lines are the left and right sides of the OSM road, respectively.

Figure 12. The positional accuracy assessment results with different pre-specified buffer widths.

Figure 13. Distribution of the positional accuracy assessment results based on (a)The conventional buffer-zone method and (b)The statistically defined buffer-zone method.