ORIGINAL PAPER

Evaluation of the uncertainty of the length of energy distribution networks

Marco Scaioni

Abstract A statistical evaluation of the uncertainty of the length of a utility network is frequently omitted in current geographic information system (GIS) applications, although this information is important for management purposes. Nevertheless, the spatial database storing geographic information about the network might have been set up through a complex process (ground surveying, digitising of analogue maps, integration of different data sources), resulting in a dataset with unclassified uncertainty. This paper proposes a practical but 'rigorous' approach to evaluate the influence of the uncertainty in the determination of the length of a network. Starting from the definition of length uncertainty in terms of simple geometric elements (arc, pair of arcs, open and closed path), some rules for its evaluation in the case of complex networks have been developed. The statistical concept applied for the evaluation of uncertainty is quite standard because it is based on the covariance propagation theory. On the other hand, the influence of correlations between the lengths of adjacent elements has been considered as well. The application of this method only requires the topological structure of the network that can be built in a GIS environment. Furthermore, an alternative 'approximate' method is proposed to find an approximation of the uncertainty of the network length with a simple formula, which requires neither any particular processing step nor the network topology. Both methods have been tested and validated on simulated and real datasets.

M. Scaioni (✉)
Department of Building Environment Science and Technology,
Politecnico di Milano,
via M. d'Oggiono 18/a,
23900 Lecco, Italy
e-mail: marco.scaioni@polimi.it

Introduction

During recent years, the use of geographic information system (GIS) technology for the management of energy distribution networks has quickly increased at Italian facility management (FM) companies. Many problems are involved in the passage from analogue to digital cartographic data, and *vectorisation* and *georeferencing* of maps are the most important tasks to be solved (Burrough and McDonnell 1998). Even though GIS systems are rapidly spreading among FM operators, some critical aspects regarding the application of these instruments are worthy of analysis, especially when considering data quality. The geometric framework of spatial databases is often derived from the *vectorisation* of existing hardcopy maps, usually at the 1:1,000 or 1:2,000 scale (Russomanno 1998; De Knecht et al. 2001). The original positional error of analogue geographic information is incremented during the digitalisation process that consists of a preliminary scanning of paper sheets, followed by a feature extraction step. Different errors are introduced when a manual (Bolstad et al. 1990) or an automatic process is used (Den Hartog et al. 1996). Subsequently, some editing tasks like *line simplification* (Douglas and Peucker 1973; Vereign 2000) and *generalisation* (Blakemore 1984; João 1995) might introduce errors that must be considered. The lack of consideration for data quality along the entire vectorisation process commonly results in the unavailability of knowledge about the accuracy of geometric features in the final spatial database (Chrisman 1982).

The problem of estimating the *uncertainty of network length measurement* (UNLM) is a very important task in Geographic Information Sciences, given that this aspect is not usually implemented in commercial GIS packages. On the other hand, knowledge regarding the consistency of a network is fundamental for its management. The Italian electric power and gas distribution industry is governed by a national authority,[1] which coordinates all companies in the field. One of its main activities is the redistribution of the total income of the energy sellers, on the basis of a set of established parameters. Among these, the length of the network owned by each company plays a key role. The source of this information is given by engineering drawings and utility maps that, in many cases, have been converted into a digital format for their use in CAD or GIS environments (Keefer et al. 1991).

In the literature, several works can be found about spatial uncertainty and its propagation. For a general overview of this subject and the presentation of error models, see e.g. Caspary and Scheuring (1993), Heuvelink and Burrough (2002), Kiiveri (1997), Leung and Yan (1998), Shi (1998) and Meidow et al. (2009); for the computation of the uncertainty of line segments, see Goodchild and Hunter (1997) and Shi and Liu (2000), while Navratil and Achatschitz (2004) address that of areas; Tong et al. (2003) examine the problem of defining a positional error model for circular curve features. Some authors also proposed alternative approaches based on *fuzzy statistics* (e.g. Shi and Liu 2004). Despite the broad variety of papers, the problems of the evaluation of UNLM in real networks composed of many arcs are not yet properly investigated, at least to the author's knowledge. Some applications can be found in the field of geography, but these cover other types of linear objects like contour lines or cost lines (e.g. Galloway and Bahr 1979).

The classic statistical approach to this would consider the *covariance matrix* expressing the positional errors of the nodes of each arc of the network and their geometric layout and would compute the variances of the length of all arcs. This consists of the variance of the length of each arc, as well as all the covariances between arcs sharing a node. Here, a solution to the problem of evaluating UNLM following a formalisation based on geodetic *bearings* (Wolf and Ghilani 2006) is proposed. By combining formulas to compute the uncertainty of simple aggregations of arcs, a 'rigorous' method to evaluate that of complex networks is addressed. This method is clear and very simple from a theoretical point of view, but must be implemented in a GIS operational environment or, alternatively, in an external routine that works on GIS outputs. If the adopted software does not already implement a tool to manage this task or cannot be programmed, this approach is not useful at all.

Thus, a simplified solution, referred to as the 'approximate' method, is also presented. This yields an approximate value for UNLM on the basis of the knowledge of the number of network nodes and their multiplicity. The variance of the whole network length is the result of the sum of all variances of each single arc, added up to covariances, whose contribution may also be negative (as often happens). The first term is very trivial to compute, and the second term depends upon the geometric relationships between different arcs. Considering the global structure of FM networks in urban areas, some approximation for the covariances is introduced.

In the 'Fundamental hypotheses' section, the fundamental hypotheses that have been assumed in the development of the proposed theory are presented (definition of uncertainty and stochastic model). After the presentation of some formulas to compute the uncertainty of basic network elements ('Uncertainty of basic linear features' section), the 'rigorous' and 'approximate' methods for its calculus, in the case of large networks, are reported in the 'Uncertainty of the length of complex networks' section. The 'Systematic errors' section explores some systematic errors that could affect the estimate of UNLM and reports how these can be modelled. Some results obtained from the application of both proposed methods to simulated and real datasets are reported in the 'Results on simulated and real datasets' section, while conclusions are drawn in the 'Conclusions' section.

Fundamental hypotheses

The goal of the paper is to evaluate the UNLM on the basis of a spatial database, storing coordinates of network nodes and their topological relations (Mainguenaud 1995). The basic assumption of the positional error model of a point takes on *circular normal distribution*, which is usually adopted for cartographic data (Leung and Yan 1998). The estimated standard deviation of the total length of a network is assumed as a measure of the UNLM.

Furthermore, we assume that the positional error of all points in the spatial database is the same with the standard deviation, σ, and that no correlation between coordinates of different points exists.

Concerning the covariance σ_{xy} between x and y coordinates of the same node, in general, it could not be neglected; some considerations about the contribution of σ_{xy} to the UNLM will be given later in the 'Correlated x–y coordinates' section.

In the following, the evaluation of UNLM begins by considering the standard deviation of the length of simple linear structures, combinations of two or more linear elements and complex networks.

---

[1] AEEG—Authority per l'Energia Elettrica e il Gas (Authority for Electric Power and Gas).

Uncertainty of basic linear features

Uncertainty of a single linear feature

The basic case we consider is that of a single *linear feature*, assumed as a straight line segment defined by two points (*nodes*). Its length is expressed as l_{ij}, with associated standard deviation σ_{lij} (hereafter, indices i and j are omitted for clarity where obvious). The functional relationship between nodes and length is given by:

$$l = f(\underline{x}) = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} = \sqrt{\Delta x^2 + \Delta y^2} \quad (1)$$

where $\underline{x} = [x_i \ y_i \ x_j \ y_j]^{\mathrm{T}}$.

The stochastic model is given by the standard deviations of the coordinates of both nodes and by their covariances. At this stage, some initial assumptions have been introduced:

- Equal standard deviations for both x–y coordinates ($\sigma_{xi} = \sigma_{yi} = \sigma_i$);
- Equal standard deviations for both nodes ($\sigma_i = \sigma_j = \sigma$); and
- No correlation between coordinates of different nodes.

In general, the covariance matrix associated with the coordinate vector $\underline{x}$ has the form:

$$\mathbf{C}_l = \begin{bmatrix} \sigma^2 & \sigma_{xy} & 0 & 0 \\ \sigma_{xy} & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & \sigma_{xy} \\ 0 & 0 & \sigma_{xy} & \sigma^2 \end{bmatrix}. \quad (2)$$

The covariance σ_{xy} requires a more detailed analysis for each specific case. For methodological purposes only, the first case of $\sigma_{xy} = 0$ is discussed, and secondly, the case in which $\sigma_{xy} \neq 0$.

Uncorrelated x–y coordinates

The standard deviation (σ_l) of the length of a single linear feature can be evaluated by applying *covariance propagation* for the non-linear case (Mood et al. 1974):

$$\sigma_l = \pm\sqrt{\underline{J}_x \mathbf{C}_l \underline{J}_x^T} \quad (3)$$

where the *Jacobian* is expressed as:

$$\underline{J}_x = \frac{\partial l}{\partial \underline{x}} = \begin{bmatrix} \frac{\partial l}{\partial x_i} & \frac{\partial l}{\partial y_i} & \frac{\partial l}{\partial x_j} & \frac{\partial l}{\partial y_j} \end{bmatrix} = \begin{bmatrix} -\frac{\Delta x}{l} & -\frac{\Delta y}{l} & \frac{\Delta x}{l} & \frac{\Delta y}{l} \end{bmatrix} \quad (4)$$

and the stochastic model is given by the covariance matrix 2, which in this case reduces to the diagonal form because covariances have been considered negligible.

By applying formula 3, the variance of the length can be computed as follows:

$$\sigma_l^2 = \frac{1}{l^2}(2\Delta x^2 + 2\Delta y^2)\sigma^2 = \frac{1}{l^2}(2l^2)\sigma^2 = 2\sigma^2. \quad (5)$$

This result immediately shows the independence of σ_l from the real length l of the linear feature and from the position of the nodes.

Correlated x–y coordinates

The introduction of covariance between x and y coordinates of the nodes results in a covariance matrix $\mathbf{C}_l$ presenting a block-diagonal form 2. The application of formula 3 allows for the derivation of an expression for the variance of the length:

$$\sigma_l^2 = \underline{J}_x \mathbf{C}_l \underline{J}_x^T = \frac{2}{l^2}\left[(\Delta x^2 + \Delta y^2)\sigma^2 + 2\Delta x \Delta y \sigma_{xy}\right]$$
$$= 2\sigma^2 + 4\frac{\Delta x \Delta y}{l^2}\sigma_{xy} \quad (6)$$

By substituting correlation ρ_{xy} to covariance σ_{xy} and by introducing the bearing θ (see Fig. 1), formula 6 can be re-written as follows:

$$\sigma_l^2 = 2\left(1 + 2\frac{\Delta x}{l}\frac{\Delta y}{l}\rho_{xy}\right)\sigma^2 = 2\left(1 + \rho_{xy}\sin 2\theta\right)\sigma^2 \quad (7)$$

showing that the uncertainty of length, in the case of correlated x–y coordinates, depends on the correlation ρ_{xy} and on the orientation of the segment with respect to the adopted reference system. Furthermore, the contribution of ρ_{xy} can be positive or negative in the evaluation of length uncertainty.

Considering now that the goal of this paper is to establish a method for computing the uncertainty of length of a global network and not only of a single segment, we attempt to evaluate the total contribution of a large number of terms $\rho_{xy}\sin 2\theta$. In the hypothesis regarding a large
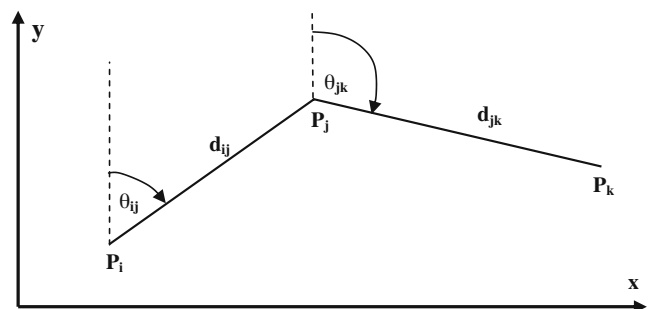

Fig. 1 The geometric model adopted for the parameterisation of straight line segments through coordinates of nodes (P_i, P_j, P_k) and bearing angles (θ_{ij}, θ_{jk})

network, diverse possible scenarios might be grouped into two main categories, as described in the following paragraphs.

Correlations ρ_{xy} are independent from the geometry of the network This case occurs when the coordinates of the network nodes have been directly measured by a geodetic technique (e.g. by global navigation satellite systems or by a theodolite). In this case, the covariance matrix of each node will be independent from its position in the network, only dependent on the geometric layout of the adopted measurement technique.

Otherwise, this is the case when the node coordinates are read from existing maps. In a large urban network, we might assume that bearings θ have a uniform probability distribution function and all correlations ρ_{xy} are similar. Then, the global contribution of the term $\rho_{xy}\sin 2\theta$ can be reasonably neglected.

However, a network presenting a particular shape is different, with one or more prevalent directions of segments. In this situation, the problem should be analysed in more detail by considering the specific geometry of the network layout.

Correlations ρ_{xy} are dependent from the geometry of the network This case might occur when nodes of the network have been introduced in the GIS environment with respect to the distances from other cartographic features (buildings, roads, pavements, etc.). For example, a pipe could be inserted on the basis of a pair of some distances of its nodes from the nearest buildings. The covariance matrix of each node would thus depend on three grounds:

- Accuracy of range measurements;
- Accuracy of reference points taken on the map; and
- Geometric scheme adopted for point positioning.

An analytical estimation of any covariance matrix that takes into account all the above-mentioned aspects is not practical because, usually, information about how each node was inserted into the GIS environment is not available. Moreover, different methods might be applied together. We attempted to draw some considerations from the analysis of the most typical configurations adopted for positioning points, with respect to other cartographic features (see examples in Fig. 2): (a) the intersection of a set of ranges from points with known coordinates, (b) the
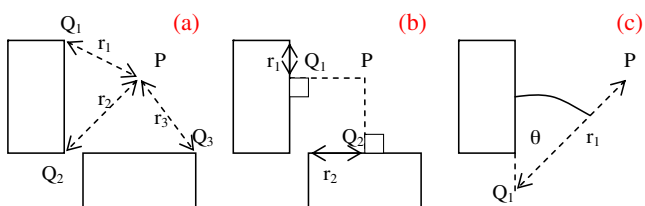

Fig. 2 Different techniques adopted for the positioning of a point P with respect to other cartographic features. Points Q_i have known positions, while distances r_j are observed on the field

use of orthogonal ranges with respect to a pair of known points on well-defined straight lines and (c) the *polar method*. In all these cases, the resulting covariance matrices of positioned point are strictly dependent on the position of reference points, and usually, values of correlations are quite small. Considering that, in a large network, several geometric configurations for reference points are possible, the global contribution of correlations can still be neglected.

Uncertainty of the length of two adjacent arcs

The following step considers two adjacent straight line segments of a network sharing a common intermediate point (see Fig. 1). Nodes are called indices i, j and k, respectively. We consider the same assumptions established for the case of a single segment presented in the 'Uncertainty of a single linear feature' subsection, resulting in a covariance matrix 2 for each point in diagonal form.

The mean value for the total length $l_{ijk} = l_{ij} + l_{jk}$ of both arcs can be computed by the sum of two terms of type 1. Consequently, the variance of the length d is given by:

$$\sigma^2_{l_{ijk}} = \sigma^2_{l_{ij}} + \sigma^2_{l_{jk}} + 2\sigma_{l_{ij}l_{jk}}. \tag{8}$$

The variances of each component can be expressed as a function of node variances due to Eq. 5, obtaining the following:

$$\sigma^2_{l_{ij}} = \sigma^2_{l_{jk}} = 2\sigma^2, \tag{9}$$

while the covariance can be computed as follows:

$$\sigma_{l_{ij}l_{jk}} = \left(\frac{\partial l_{ij}}{\partial x_j}\right)\left(\frac{\partial l_{jk}}{\partial x_j}\right)\sigma^2 + \left(\frac{\partial l_{ij}}{\partial y_j}\right)\left(\frac{\partial l_{jk}}{\partial y_j}\right)\sigma^2$$
$$= -\left(\Delta x_{ij}\Delta x_{jk} + \Delta y_{ij}\Delta y_{jk}\right)\frac{\sigma^2}{l_{ij}l_{jk}}. \tag{10}$$

In a similar way to that adopted in Eq. 7, by introducing the bearings θ_{ij} and θ_{jk} and by applying a simple trigonometric relation, the covariance between lengths of components can be re-written as:

$$\sigma_{l_{ij}l_{jk}} = -\left(\sin\theta_{ij}\sin\theta_{jk} + \cos\theta_{ij}\cos\theta_{jk}\right)\sigma^2$$
$$= -\cos\left(\theta_{ij} - \theta_{jk}\right)\sigma^2. \tag{11}$$

The expression of the variance of the total length of the pair of consecutive segments then becomes:

$$\sigma^2_{l_{ijk}} = 4\sigma^2 - 2\cos\left(\theta_{ij} - \theta_{jk}\right)\sigma^2 \tag{12}$$

where the contribution of the covariance term might play a positive or negative role, according to the difference between bearings and then depending on the relative rotation between two adjacent arcs. In Fig. 3, all configurations corresponding

Fig. 3 Contribution of the correlation term in the uncertainty of the length of two consecutive segments (Eq. 11). The first segment is P_iP_j, and the second $P_jP_k^n$ ($n=1, 2, 3, 4$)

$$\rho = -\cos(\theta_{ij} - \theta_{jk})$$
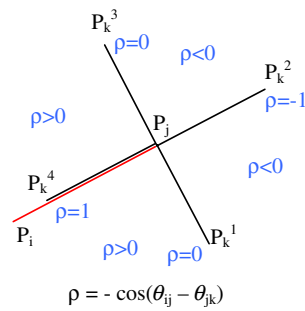
to a change in the sign of covariance term 11 are reported. The most favourable case is that with two aligned segments because a possible error in one of two lengths is compensated for by the other. The sign of the covariance term remains negative, as long as the second segment has the same direction as the first; the sign changes in the case of inversion (moreover, in real FM networks, this case seldom happens). When the pair of network arcs is orthogonal, the covariance is naught and the variance of the length can be computed by merely summing up the variances of each segment.

Uncertainty of the length of open and closed polylines

A *polyline* is a set of consecutive and adjacent straight line segments. When the initial node is coincident with the last one, the polyline is said to be *closed*, but is otherwise termed *open*. The evaluation of the uncertainty of the length of a polyline can be considered as a further development of the case concerning a single linear segment.

First, we define the variance of the length for an *open polyline* made up of n arcs and $n+1$ nodes. All the assumptions previously established still hold. By generalising formula 10, the variance of the length is the sum of the variances of each arc and correlations between pairs of adjacent arcs:

$$\sigma_l^2 = 2n\sigma^2 - 2\sigma^2 \sum_{i=1}^{n-1} \cos(\theta_{i,i+1} - \theta_{i+1,i+2}). \tag{13}$$

It is very interesting to note that the correlation term is nil in the case of a polyline with arcs that are always mutually orthogonal, a situation very common in real FM networks in urban areas.

In the case of a *closed polyline*, the last node is identical to the first, and correlations between this pair of arcs must be taken into account. Then, the variance can be found through the following expression:

$$\sigma_l^2 = 2n\sigma^2 - 2\sigma^2 \sum_{i=1}^{n} \cos(\theta_{i,i+1} - \theta_{i+1,i+2}). \tag{14}$$

With the analysis of polylines, all the variances of length for any basic arc configuration have been presented. Now,

the paper will continue to discuss the use of these fundamental relations in the evaluation of UNLM in real networks.

Uncertainty of the length of complex networks

Modern energy distribution systems are based on complex networks, which can be modelled by a *graph* structure, made up of a set of nodes and arcs with their topological relations. The graph may be either *connected* or *unconnected*. Nodes can be classified as *external* and *internal*, the latter prevalently grouping nodes linking two or three arcs; nodes with a higher number of connections are rare.

The process of evaluating the uncertainty in GIS can be carried out by a 'rigorous' approach ("'Rigorous' method" subsection) by using the theory proposed in the first part of the paper. The procedure will be presented in the next subsection, and it is quite applicable to operational GIS environments. Given a table of node coordinates, the topological structure of the network and some assumptions about the stochastic model, the global UNLM, can be easily evaluated.

Some simplifications that can be introduced in the computation of covariances for configurations of real networks allow for a rough estimation of the UNLM by an 'approximate' method ("'Approximate' method" subsection). This can be useful when a first estimate of the UNLM is needed or when functions for its computation by the 'rigorous' method cannot be implemented.

'Rigorous' method

Let us consider a network made up of n arcs and m nodes; at this turn, each node may link k arcs, usually up to three. According to the results of the first part of the paper, the UNLM is given by the sum of two components:

- Sum of the variance of every arc length (S^2); and
- Sum of the covariances between arcs connected to the same node (COV).

The first component can easily be evaluated by a list of all arcs with related nodes, as an extension of formula 7:

$$S^2 = 2 \sum_{i=1}^{n} \sigma_i^2 \tag{15}$$

where σ_i^2 is the variance of the generic coordinate of a single arc. Nevertheless, when all nodes have the same covariance matrix $\mathbf{C}_i = \mathrm{diag}(\sigma^2, \sigma^2)$ and considering the hypotheses presented at the beginning of the 'Fundamental

hypotheses' section, the term S^2 can be simplified as follows:

$$S^2 = 2n\sigma^2. \tag{16}$$

The second term of the UNLM accounts for all correlations between arcs related to the same node. The covariance introduced by a link between two arcs in node j can be derived from expression 11:

$$(\text{COV}_{\text{double}})_j = -2\cos(\theta_{ij} - \theta_{jk})\sigma^2. \tag{17}$$

When a node is composed of three related arcs, the total covariance term will be the sum of each combination of terms (Eq. 17):

$$(\text{COV}_{\text{three}})_j = \sum_{p=1}^{3} (\text{COV}_{\text{double}})_p. \tag{18}$$

This formula can also be adapted to encompass nodes having more than three linking arcs; in this case, it is enough to consider all covariances arising from any possible pair-wise combinations of arcs.

The global covariance term COV will be the sum of the components of all nodes.

Finally, the total variance of the network length can be evaluated, resulting in the following:

$$\sigma_{\text{NL}}^2 = S^2 + \text{COV} = S^2 + \sum_{p=1}^{m} (\text{COV})_p. \tag{19}$$

Formula 19 shows that, in UNLM evaluation, the contribution of the term COV is not a priori negligible; moreover, covariances in most cases reduce σ_{NL}, so their omission could result in overestimating the UNLM. Examples reported in the 'Results on simulated and real datasets' section will demonstrate the real contribution of the COV term.

The proposed method is completely suitable for implementation in a GIS environment. While the structure of the network can be modelled by a graph, the stochastic problem can be reduced to one of the following cases.

All coordinates of nodes with the same planimetric standard deviation

The term S^2, representing the sum of the variance of each arc length, can be computed by formula 16. In all terms contributing to the global covariance COV, the specific variance of each node coordinates can be replaced by the general value σ^2. This term can be introduced by the user on the basis of considerations about the construction process of the map (see De Knecht et al. 2001).

Coordinates of nodes with different standard deviations

Even though only the case with $\sigma_{xi} = \sigma_{yi} = \sigma_i$ is dealt with here, formulas 17 and 18 can still be applied, considering the real standard deviation of each node. These values can be stored as attributes of nodes in the geodatabase from which data for network modelling are retrieved.

'Approximate' method

To define a method to evaluate an approximate value for UNLM, some preliminary considerations should be mentioned.

First of all, in the "'Rigorous' method" subsection, the global UNLM was given as a function of the variance of each arc considered as a single element and of covariances between different arcs converging to the same node. The term S^2 can be easily computed through Eq. 16 in the case all nodes would present the same diagonal covariance matrix. The term COV must be simplified because its computation requires knowledge of the bearing of each arc.

The second consideration arises from observing that in an urban utility network, the mutual directions of arcs are prevalently parallel or orthogonal. Nevertheless, also when these conditions are not exactly verified, directions are very close to them. As the covariances are due to nodes, three main cases must be analysed; the first two will concern double nodes, and the third, the T-junctions (triple nodes).

Double nodes can be divided between those encompassing two aligned arcs (1) and two orthogonal arcs (2). In the case 1, the covariance introduced by the node is $-2\sigma^2$. This results in a nil contribution of the node to the computation of UNLM because variances of both arcs' length and covariance are exactly opposite. Practically, a double node of type 1 could be neglected. Unfortunately, very often, adjacent arcs do not exactly lie on the same straight line, even though differences between their bearings are negligible. As a result of Eq. 17, the COV term is based on the cosine function, which assumes a value very close to 1, even in the case of arguments significantly far from 0° (e.g. a difference of $\alpha = \pm 20°$ results in a value of $\cos\alpha = 0.94$). Then, it is possible to erase from the graph all nodes presenting a difference between the bearings of arcs converging to it, which are under a fixed threshold (e.g. 180±20°). The underestimation of the contribution of the cancelled covariances could be compensated for by introducing a corrective term:

$$C = n_e \times \rho_{st}\sigma^2 \tag{20}$$

where n_e is the number of eliminated nodes of type 1 and ρ_{st} is an average value for correlation between arcs converging to each node that was eliminated. On the other hand, in the 'approximate' method, the evaluation of bearings is omitted to simplify its application. Consequently, all double nodes

are excluded from the computation of both S^2 and COV terms.

However, double nodes could be categorised as a second type (2), i.e. those featuring a difference of bearings close to 90° or 270° (mutually orthogonal arcs). These give a covariance term that is approximately nil. We observed that, in a utility network, the number of double nodes (2) is small, with respect to the total number of double nodes. Neglecting them would result in a negligible underestimation of the term S^2, which is compatible with the level of approximation of this method.

In the case of nodes made up of three arcs (3), such as T-junctions, the 'rigorous' evaluation of the covariance should be based on three terms: the covariance between two aligned arcs, i.e. $-2\sigma^2$, and two zero covariances between two pairs of orthogonal arcs. The covariance introduced by each T-junction is then $-2\sigma^2$, resulting in a global contribution depending on their number n_T; this can be derived by the graph information from the GIS system. Thus:

$$\mathrm{COV_T} = -2n_T\sigma^2. \qquad (21)$$

By considering the hypothesis of a large network, the remaining configurations of nodes are negligible.

The approximate evaluation of global length uncertainty can be given as follows:

$$\sigma'^2_{NL} = 2(n - n_T)\sigma^2. \qquad (22)$$

In the following, all steps required for the computation of σ'^2_{NL} are reported:

1. *Input data*: A network stored in a geodatabase is needed. The stochastic problem is reduced to the simplified conditions reported in the 'Fundamental hypotheses' section, with the assumed standard deviation of the generic x–y coordinates and $\sigma_{xy}=0$;
2. *Graph analysis*: The total number of nodes (n) and that of T-junctions (n_T) must be evaluated (see the 'Results on simulated and real datasets' section); and
3. *UNLM evaluation*: Once all basic elements have been acquired, the UNLM can be evaluated by applying the simplified formula 22.

Some examples reported in the 'Results on simulated and real datasets' section will show the amount of error introduced in the evaluation of UNLM when using the 'approximate' method with respect to the 'rigorous' method.

Systematic errors

The proposed method for the UNLM evaluation is based on the retrieval of information from a geodatabase. Here,

the problem has been considered only for position, neglecting the influence of elevation. This introduces two sources of possible systematic errors due to (1) the reduction to the ellipsoid of any distance represented on the map and (2) the slope gradient of each segment. Furthermore, the length of an arc on the map is scaled to a factor depending on the adopted mapping projection (3).

In the evaluation of the total length of a network, the distances between connected points are computed through Eq. 1. To correctly consider the aforementioned effects, all distances evaluated this way must be analytically compensated for by using information from the geodatabase, i.e. elevation of points for errors 1 and 2, grid coordinates for error 3. On the other hand, in the following, an estimate of errors introduced when omitting these corrections is made. This is relevant when applying the 'approximate' method presented in the ''Approximate' method' subsection, as this tries to give a simple evaluation of the UNML.

In the following paragraphs, error 3 is discussed in the 'Error due to linear deformation of the adopted map projection' subsection because its compensation is a preliminary task that must be completed before correcting errors 1 and 2. The discussion of these is presented in the 'Errors due to elevation' subsection, while in the 'Considerations' subsection, some considerations will be explored.

Error due to linear deformation of the adopted map projection

A first systematic error that could be introduced in the evaluation of network length depends on the linear deformation (m) of maps. Disregarding the specific map projection adopted, linear deformation usually depends on two main factors: a global isotropic scale (m_{isot}), which is used to reduce the uncertainty of points under the *draughting error*, and an anisotropic deformation, which depends on the projection (Pearson 1990). In a *Gauss projection*, the relation between a differential length on the ellipsoid (dL_e) and the corresponding one on the map (dL_{map}) is given by:

$$dL_{map} = m \times dL_e = m_{isot}\left(1 + \frac{(E - E_0)^2}{2\rho N}\right) \times dL_e \qquad (23)$$

where E is the eastern coordinate of the infinitesimal line dL_{map}, E_0 is the *East false origin* adopted in the mapping grid (e.g. E_0=500,000 m for UTM projection) and ρ and N are the two *principal curvatures* of the reference ellipsoid in the region where the network is located (Wolf and Ghilani 2006). Because a complex utility network can be considered as the sum of many small line segments located in a region featuring roughly the same value for E, ρ and N, Eq. 23 can

be extended from differential to finite quantities. Then, the total length of the network on the ellipsoid surface would be:

$$L_e = \frac{1}{m} L_{map} = \frac{2\rho N}{m_{isot}\left[(E - E_0)^2 + 2\rho N\right]} L_{map}. \quad (24)$$

Thus, the error ε_m due to the neglect of this scaling effect can be expressed as:

$$\varepsilon_m = 1 - \frac{L_{map}}{L_e} = 1 - m. \quad (25)$$

In Fig. 4, the amount of error ε_m is shown for different distances from the central meridian of the projection fuse. Here, the influence of the latitude is not taken into account because this does not significantly influence the result of Eq. 24. As one can see, the presence of an isotropic scale $m_{isot}=0.9996$ results in a map contraction close to E_0 that might reach 0.4‰ and a dilatation towards the borders of each fuse up to 1.1‰. Considering as an example a network with $L_{map}=100,000$ m, errors due to the neglect of this scaling effect would result in an underestimation of the length of 40 m or in an overestimation of 110 m, according to the distance from the central meridian.

In case any another projection was adopted instead of the Gauss projection, Eqs. 23 and 24 must be replaced by the corresponding formulas for the implemented projection (Pearson 1990).

Errors due to elevation

The map plane is considered to be a 2D transformation of the reference ellipsoid, so if the real length of the network (and its uncertainty) has to be evaluated, some corrections must be considered. First of all, the relation between the length (L_Q) of a finite line located at an average elevation

Q_{av} (with respect to the reference ellipsoid) and the length (L_e) of the corresponding segment on the ellipsoid can be computed through the formula (Bezoari et al. 1989):

$$L_Q = \left(\frac{Q_{av}}{R} + 1\right) L_e \quad (26)$$

where $R = \sqrt{\rho N}$ is the average radius of the local sphere. As demonstrated by Eq. 27, the error ε_Q committed when the real length of a line L_Q is replaced by the length L_e, read on the map without correction, results in underestimating L for positive elevations:

$$\varepsilon_Q = 1 - \frac{L_e}{L_Q} = \frac{Q_{av}}{Q_{av} + R}. \quad (27)$$

Equations 26 and 27 still hold when considering the total length of a network, instead of a single line segment. Then, to give a rough evaluation of the error ε_Q, we have considered a set of locations at different elevations ranging from 0 to 500 m over the reference ellipsoid. The influence of the radius variation according to latitude is completely negligible for the problem under investigation. As shown in Fig. 5, the underestimation of L_Q can reach up to 0.08‰ of the total length of the network located at an elevation of 500 m on the ellipsoid. Practically, for a network with a length $L_e=100,000$ m, this means a maximum error of 8 m.

A second systematic error related to the elevation is due to the inclination of line segments on the real world, an effect that cannot be compensated for by Eq. 26 only. Taking the drawing in Fig. 6 as an example, the real 3D length (L_{3D}) of the straight line segment AB is a function of the length L_Q of its projection to a surface parallel to the ellipsoid (locally approximated by a tangent plane) and passing through the medium point of AB. The length L_Q is that derived from the geodatabase, but corrected by
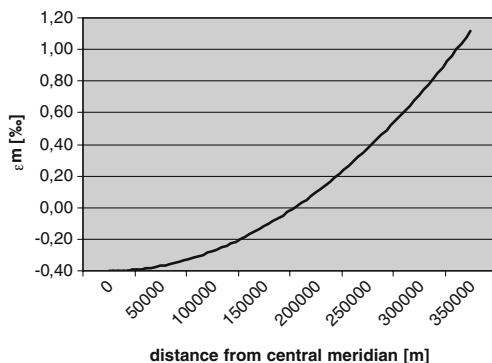

Fig. 4 Error (ε_m) in the evaluation of the network length, when the correction for scaling due to the linear deformation introduced by the map projection is omitted
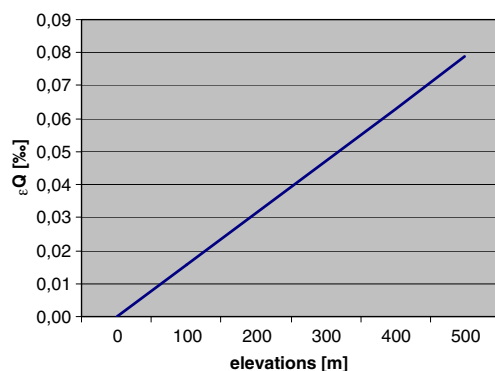

Fig. 5 Underestimation error (ε_Q) of the length of a network, when the correction for the reduction to the ellipsoidal surface is omitted
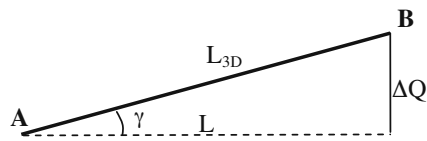
Fig. 6 Real length of a line (L_{3D}) and the one readable from maps (L)

means of Eq. 26. The relation between L_Q and L_{3D} is as follows:

$$L_{3D} = \frac{L_Q}{\cos \gamma} = \frac{L_Q}{\cos a \tan s} \qquad (28)$$

where s is the *mean slope gradient* $\Delta Q / L_Q$. When considering the total length of a network in urban areas with smooth slopes, Eq. 28 still holds. In a similar manner as in Eq. 27, a new error ε_{3D} can be defined as follows:

$$\varepsilon_{3D} = 1 - \frac{L_Q}{L_{3D}} = 1 - \cos a \tan s. \qquad (29)$$

By considering a slope average gradient s ranging from 0% to 5%, the maximum increment of the 3D length with respect to the planimetric length will reach 1.2‰ (see Fig. 7). For a network featuring L_Q=100,000 m, this means an underestimation of its length up to 120 m.

Considerations

In this section, three different kinds of systematic errors imputable to the uncorrected or uncompleted transformation from the map plane to the real world have been analysed. Considering some common conditions of the local topography, typical of most urban areas, the maximum amount of these errors is 1.1‰ for the scaling due to the adopted mapping projection, 0.08‰ for neglecting the compensation for the mean elevation and 1.2‰ for the mean ground inclination. In a network featuring a length of 100,000 m, the pessimistic concurrence of all these errors will lead to a maximum error not larger than 150–200 m, i.e. 1.5–2‰ the network length.

Results on simulated and real datasets

In order to assess the effectiveness of both presented methods to evaluate the UNLM, three different tests were conducted.

The first case study concerned a simulated small network (test 'SIM'), while the second and third tests were based on data derived from two existing gas networks in urban areas (tests 'REAL1' and 'REAL2'), which are located in a small town (~15,000 inhabitants) and in a mid-sized town (~100,000 people), respectively. For the sake of reservation, the locations of both test sites are anonymous. However, they

are both in flat areas at mean geographic coordinates φ=45°, λ=10°, h=100 m. In both tests, either the 'rigorous' or 'approximate' method was applied to compare the results. In test 'REAL2', the computation of UNLM was carried out on the overall network, as well as on a portion of it. With this approach, the influence of the network size on the accuracy of results could also be evaluated.

The objectives of these experiments are twofold: firstly, a comparison between both proposed methods ('rigorous' and 'approximate') and, secondly, to give an evaluation of the UNLM in the case of operational distribution gas networks and to compare it with the systematic errors described in the 'Systematic errors' section (see the 'Effects of systematic errors' subsection).

The computation of the UNLM was performed by using a code developed and running it in MATLAB®. This basically requires, as input, three tables describing the structure of the network graph: the *node table* (ID_NODE_i, E_i, N_i), which gives the position of each node i; the *arc table*, which lists all the elementary straight line segments j of the network with their nodes (ID_ARC_j, ID_NODE1_j, ID_NODE2_j); and the *link table*, which stores the nodes connected to each node i and the number of connections (ID_NODE_i, ID_NODEJ_i, ID_NODEK_i, ID_NODEL_i, $NLINK$). A data model where each node can have no more than three connected nodes was implemented here. All tables needed for the evaluation of UNLM can be easily derived from data exportable from all commercial GIS packages where the topology of the considered network has been previously constructed.

Test 'SIM'

A graph corresponding to a small size gas distribution network was generated, featuring a global length of 2,922.1 m (Fig. 8). The simulation was performed by directly building up the three tables described in the
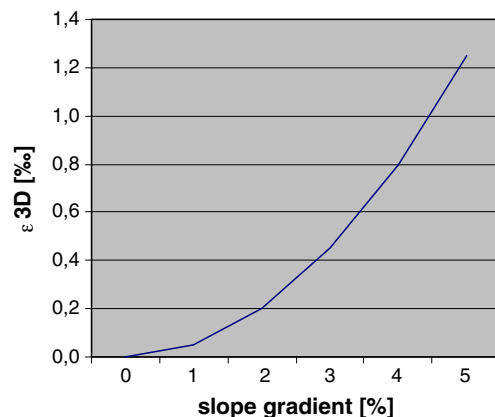

Fig. 7 Underestimation error (ε_{3D}) of the length of a network when neglecting the average incline of the network lines
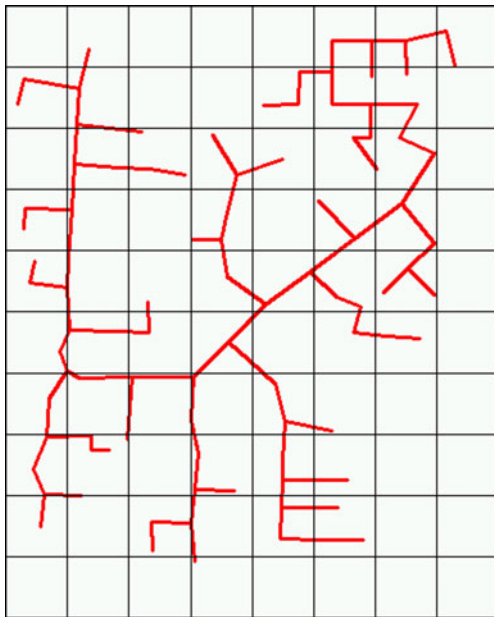
Fig. 8 Geometric layout of the simulated network of the test 'SIM'; the size of each grid cell is 100×100 m

introduction of this section. The geodatabase of the network was thought to be derived from digitising the analogue maps, resulting in a hypothesis for the standard deviations of point coordinates of 0.40 m. Both coordinates have been assumed as not being correlated. This value is typical of maps at the 1:2,000 scale that are frequently adopted as reference for utility network representation. All characteristics of the simulated graph are reported in Table 1.

The application of the 'rigorous' method to evaluate the UNLM resulted in 4.3 m, corresponding to 1.5‰ of the network length (see Table 2). The contribution of all the covariances (term COV) is approximately 44% of the sum of the arc variances (term S^2), showing that neglecting the COV term would result in an exaggerated estimation of the UNLM.

By using the 'approximate' method, a value of the UNLM=4.9 m was obtained, which overestimates the uncertainty by 14% (see Table 3).

Tests 'REAL1' and 'REAL2'

Both tests were carried out on gas distribution networks that are fully operational in two Italian towns. Layouts of both can be seen in Figs. 9 and 10, respectively. To apply the algorithm for the computation of UNLM, preliminary data pre-processing was needed to create the tables of nodes, arcs and links. The network digital maps were exported in CAD format (DWG) from the GIS systems, where they were stored and used, and then loaded into Autodesk Autocad Map 3D®. The maps featured a scale of 1:2,000, then a standard deviation of point coordinates $\sigma=\pm0.40$ m was defined, as common for Italian standards. On the other hand, in all the elementary formulas adopted in the 'rigorous' (Eqs. 16 and 17) and 'approximate' methods (Eq. 22), results are linearly proportional to σ^2. Consequently, a change in σ^2 will affect only the absolute value of UNLM, but not the ratio between results obtained with different approaches. No grounds were found to consider correlations between both coordinates of each node.

A preliminary check of the data geometry was performed in order to identify possible duplicate features, false crossings, imperfect connections to nodes, etc. Furthermore, each polyline was decomposed into its elementary straight line segments because intermediate nodes might also contribute to the UNLM. Secondly, a network topology was defined and applied to both datasets. This step allowed for the extraction of all data needed for computing the UNLM. Indeed, by a topological query, it was possible to generate in Autocad Map 3D® an *output table* where each arc j of the network is associated with its nodes and with their coordinates (ID_NODE_j, ID_NODE1_j, ID_NODE2_j, $X1_j$, $Y1_j$, $Z1_j$, $X2_j$, $Y2_j$, $Y2_j$). From this table, it is then possible to automatically derive the *node, arc* and *link tables*, a task that is performed by a MATLAB® routine. This procedure generates two main datasets ('REAL1' and 'REAL2') whose characteristics are depicted in Table 1. From the case study 'REAL2', four subsets have been derived ('REAL2a', 'REAL2b', 'REAL2c' and 'REAL2d') by considering only a portion of it. Approximately each of

Table 1 Characteristics of the networks adopted for tests described in the 'Results on simulated and real datasets' section

Case study	No. of arcs	No. of nodes	No. of single nodes	No. of double nodes	No. of T-junctions
SIM	101	102	30	46	26
REAL1	6,707	6,318	2,127	3,143	1,048
REAL2	8,448	7,823	2,078	4,165	1,580
REAL2a	4,252	4,262	1,017	2,417	828
REAL2b	1,715	1,605	514	765	326
REAL2c	788	753	238	363	152
REAL2d	427	413	149	179	85

Table 2 Results of the application to the case studies of the 'rigorous' method to evaluate the UNLM

Case study	NL [m]	UNLM [m]	UNLM/NL [‰]	S^2 [m²]	COV [m²]	COV/S^2 [%]
SIM	2,922.1	4.3	1.5	33.28	−14.8	45
REAL1	40,572.7	34.3	0.8	2,146.2	−969.7	49
REAL2	219,850.8	37.1	0.2	2,703.4	−1324.6	52
REAL2a	121,431	25.5	0.2	1,360.6	−708.2	45
REAL2b	50,393.2	17.4	0.3	548.8	−246.7	49
REAL2c	24,348	11.3	0.5	252.2	−124.0	47
REAL2d	12,663.2	8.5	0.7	136.6	−63.6	44

these accounts for a number of elements (arcs and nodes), which is half of the dataset from which it originates. The new datasets have been generated to test both proposed methods for UNLM computation with respect to the size of the network.

The application of the 'rigorous' method to both original datasets resulted in a very similar value of UNLM (34.3 m for 'REAL1' and 37.1 m for 'REAL2'), despite the different lengths (40 km for 'REAL1' and 220 km for 'REAL2'), leading to a very different ratio between the UNLM and the network length (0.8‰ versus 0.2‰, respectively). On the other hand, an important difference exists between the two network structures due to the number of single nodes that are bigger, in the case of dataset 'REAL1' (see Table 1). Indeed, this network is smaller than that of 'REAL2', but it comprehends all the final pipes connecting the gas network to buildings; network 'REAL2' does not include all final pipes because they are stored in a separated layer, which has not been considered here. In formula 19, single nodes do not contribute to increase the negative term COV, while they contribute positively to S^2. The contribution of all the covariances (COV) with respect to the arc variances (S^2) is slightly smaller for case study 'REAL1' (49%) than for 'REAL2' (52%). The comparison of the results obtained for different datasets of case study 'REAL2' provides evidence of a relative increment in the ratio between UNLM and the network length, which ranges from 0.2‰ for the full

dataset up to 0.7‰ for the smallest dataset, 'REAL2d'. This difference is motivated by the larger fraction of single nodes in the total number of nodes featured in the smaller datasets (see Table 1).

The use of the 'approximate' method gave an underestimation of the UNLM that was quite similar in both cases (see Table 3). However, the size of the error (about 7–8%) committed when using this method, instead of the 'rigorous' one, is acceptable for most applications. It should be observed that, the smaller the size of the network, the lower the underestimation error. For the smallest dataset ('REAL2d'), this became a slight overestimation error, as occurred with the simulated dataset 'SIM'.

Effects of systematic errors

Different systematic errors that can affect the evaluation of the network length can be properly corrected if the information needed to model them is available from the geodatabase. Scaling errors due to the adopted mapping projection can be compensated for by the knowledge of grid coordinates, and

Table 3 Results of the application to the case studies of the 'approximate' method to evaluate the UNLM, and a comparison with those achieved by using the 'rigorous' method

Case study	UNLM′ [m]	(UNLM−UNLM′)/UNLM [%]
SIM	4.9	−13.9
REAL1	31.9	7.0
REAL2	34.2	7.8
REAL2a	24.3	4.7
REAL2b	16.4	5.7
REAL2c	11.2	0.9
REAL2d	8.7	−2.3
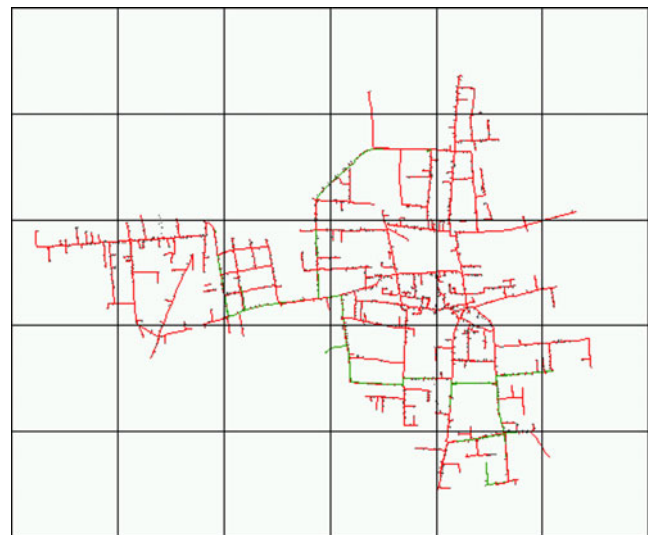
Fig. 9 Geometric layout of the real network of test 'REAL1'; the size of each grid cell is 500×500 m, while different colours (*red* and *green*) represent two coexisting sub-networks, both considered in the evaluation of UNLM
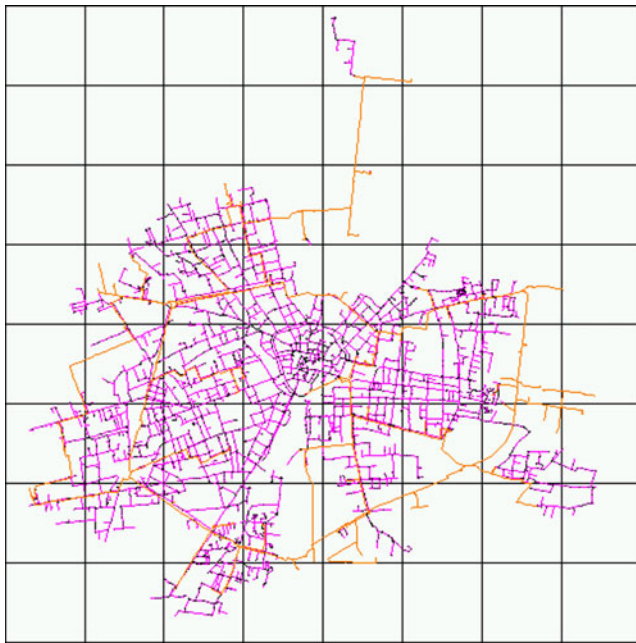
Fig. 10 Geometric layout of the real network of test 'REAL2'; the size of each grid cell is 1,000×1,000 m, while different colours represent the main (medium pressure, in *orange*) and the secondary (low pressure, in *purple*) distribution networks, respectively

elevation errors can be corrected if nodes have 3D coordinates. Furthermore, the error due to the reduction to the ellipsoidal surface (see the 'Errors due to elevation' subsection) can also be compensated for by using an average elevation on the area of the network.

Here, an evaluation of the amount of these errors when they are not compensated for in all case studies is presented. As mentioned in the introduction of this section, all case studies are located in flat areas at approximate geographic coordinates $\varphi=45°$, $\lambda=10°$, $h=100$ m. The map projection adopted is UTM (Fuse 31), based on the WGS84 ellipsoid, with a global isotropic scale $m_{isot}=0.9996$. Application of formula 25 gives an error $\varepsilon_m=-3‰$; note that, due to the distance from the central meridian ($\Delta E=78{,}815$ m), in this case, ε_m results in an overestimation of network length. The elevation error computed through formula 26 gives $\varepsilon_Q=0.016‰$. If we consider a flat topography, an average slope of 1% can be used for the

accounting of ramps and bridges. By introducing this value in Eq. 28, an error $\varepsilon_{3D}=0.05‰$ is obtained.

The results of the systematic error evaluation are reported in Table 4 where a comparison to the estimated UNLM is done. As evident, omitting the correction for the map projection scale always results in a significant error; moreover, in the case of large networks (like case study 'REAL2'), its effect is more than one time the UNLM. Other errors are less relevant in the cases of short and medium-length networks, but they cannot be neglected for large networks. On the other hand, all case studies are in favourable locations; when a network is positioned in a hilly area and where map deformation m are larger, the correction of systematic errors becomes extremely vital in order to avoid a biased estimation of the UNLM.

Conclusions

In the paper, two different approaches for computing the UNLM have been proposed, discussed and validated on simulated and real datasets. The first method ('rigorous') is based on propagating the variance of the length of each arc of the network, which is computed on the basis of the standard deviations of both nodes. Furthermore, the contributions of the covariances due to adjacent arcs are considered here, which usually reach approximately 50% of the sum of the variances. Due to the geometric layouts of the utility networks in urban areas, covariances reduce the UNLM, so they cannot be omitted. Even though, in the literature, several more involved theoretical models to define the positional accuracy of points and arcs have been presented, the simple model assumed here makes the proposed approach very operational and completely suitable for implementation in a GIS environment. Indeed, the structure of the network can be modelled by a graph and a table of point coordinates, and the topological relationships between different arcs can be exported with ease. This information is enough for the computation of UNLM that can be performed in the GIS where data are stored by programming a suitable routine or in an external mathematical environment or stand-alone software.

Table 4 Total amount of systematic errors described in the 'Systematic errors' section on the case studies

Case study	NL [m]	UNLM [m]	Total effects of different systematic errors on NL evaluation (% of the NL)		
			Map projection scale [m]	Elevation [m]	Slope [m]
			$\Delta\varepsilon_m$	$\Delta\varepsilon_Q$	$\Delta\varepsilon_{3D}$
SIM	2,922.1	4.3	−0.8 (−20%)	0.0 (1%)	0.1 (3%)
REAL1	40,572.7	34.3	−12.2 (−35%)	0.6 (2%)	2.0 (6%)
REAL2	219,850.8	37.1	−66.0 (−178%)	3.5 (9%)	3.5 (30%)

The second method ('approximate') allows for the evaluation of the UNLM on the basis of the total number of nodes and of T-junctions, in addition to the standard deviation of point coordinates. Despite the simplicity of this approach, it has yielded errors that are in the order of 7–8% (underestimation) for the largest networks, while they have decreased to become overestimation errors for the smallest networks. This method can then be properly used to give a preliminary and rapid evaluation of the UNLM.

Future work will study in more detail the error distribution of points and the effect of line simplification (Douglas and Peucker 1973) during the digitisation process, here omitted.

References

Bezoari G, Monti C, Selvini A (1989) Topografia Cartografia. Hoepli, Milan

Blakemore M (1984) Generalization and error in spatial databases. Cartographica 21:131–139

Bolstad PV, Gessler P, Lillestand TM (1990) Positional uncertainty in manually digitized map data. Int J Geogr Inf Sci 4:131–139

Burrough PA, McDonnell RA (1998) Principles of geographic information systems. Oxford University, New York

Caspary W, Scheuring R (1993) Positional accuracy in spatial databases. Comput Environ Urban Syst 17:103–110

Chrisman NR (1982) A theory of cartographic error and its measurement in digital data bases. Proceedings of Auto-Carto 5, American Congress on Surveying and Mapping, Falls Church, pp 159–168

De Knecht JP, Schavemaker JGM, Reinders MJT, Vossepoel AM (2001) Utility map reconstruction. Int J Geogr Inf Sci 15:7–26

Den Hartog J, ten Kate T, Gerbrands J (1996) Knowledge-based interpretation of public utility maps. Comput Vis Image Underst 63:105–117

Douglas DH, Peucker TK (1973) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Can Cartogr 10:112–122

Galloway RW, Bahr ME (1979) What is the length of Australian coast? Aust Geogr 14:244–247

Giussani A, Scaioni M (2004) Evaluation of uncertainty of network length on the basis of GIS data. Proceedings of the International Symposium on Spatial Data Quality, Bruck an der Leitha, Austria, 15–17 April, pp 279–280

Goodchild MF, Hunter GJ (1997) A simple positional accuracy measure for linear features. Int J Geogr Inf Sci 11:299–306

Heuvelink GBM, Burrough PA (2002) Developments in statistical approaches to spatial uncertainty and its propagation. Int J Geogr Inf Sci 16:111–113

João EM (1995) The importance of quantifying the effects of generalization. In: Müller J-C, Weibel R, Legrange J-P (eds) GIS and generalization. Taylor & Francis, London, pp 183–193

Keefer BJ, Smith JL, Gregoire TG (1991) Modeling and evaluating the effects of stream mode digitizing errors on map variables. Photogramm Eng Remote Sensing 57:957–963

Kiiveri HT (1997) Assessing, representing and transmitting positional uncertainty in maps. Int J Geogr Inf Sci 11:33–52

Leung Y, Yan J (1998) A locational error model for spatial features. Int J Geogr Inf Sci 12:607–620

Mainguenaud M (1995) Modelling the network component of geographical information systems. Int J Geogr Inf Sci 9:575–593

Meidow J, Beder C, Förstner W (2009) Reasoning with uncertain points, straight lines, and straight line segments in 2D. ISPRS J Photogramm Remote Sens 64:125–139

Mood A, Graybill F, Boes D (1974) Introduction to the theory of statistics. McGraw Hill, New York

Navratil G, Achatschitz C (2004) Influence of correlation on the quality of area computation. Proceedings of the International Symposium on Spatial Data Quality, Bruck an der Leitha, Austria, 15–17 April, pp 291–304

Pearson F (1990) Map projections: theory and applications. CRC, Boca Raton

Russomanno DJ (1998) Utility network derivation from legacy source data for feature-based AM/FM systems. Int J Geogr Inf Sci 12:445–463

Shi WZ (1998) A generic statistical approach for modelling errors of geometric features in GIS. Int J Geogr Inf Sci 12:131–143

Shi W, Liu W (2000) A stochastic process-based model for the positional error of line segments in GIS. Int J Geogr Inf Sci 14:51–66

Shi W, Liu K (2004) Modeling fuzzy topological relations between uncertain objects in a GIS. Photogramm Eng Remote Sensing 70:921–929

Tong X, Shi W, Liu D (2003) An error model of circular curve features in GIS. Proceedings of 11th ACM International Symposium on Advances in GIS, New Orleans, Louisiana, pp 141–146

Vereign H (2000) Quantifying positional error induced by line simplification. Int J Geogr Inf Sci 14:113–130

Wolf PR, Ghilani CD (2006) Elementary surveying: an introduction to geomatics. Pearson Prentice Hall, Upper Saddle River