**Probability Based Location Referencing Method – Statistical Evaluations and Estimated Probability Distributions**

Patric Hindenberger (patric.hindenberger@ingeo.uni-stuttgart.de), Volker Schieger (volker.schwieger@ingeo.uni-stuttgart.de)

Institute of Engineering Geodesy, University of Stuttgart

**Abstract.** Dynamic Location Referencing Methods are typically used for transferring location references from one map to another in such cases where no common databases and/or common structures are available. This paper describes the statistical evaluations and estimated probability distributions to be used in a “Probability Based Dynamic Location Referencing Method”. To obtain them, several random samples for specific criteria/attributes have been aggregated. On this basis, the histograms have been deviated and the estimated probability distributions have finally been estimated. As a result, the geometrical data follow an exponential distribution and all other criteria are binomially distributed. These probability distributions will be implemented in an algorithm for a “Probability Based Dynamic Location Referencing Method”.

**Keywords.** Dynamic Location Referencing, Digital Maps, Statistical Evaluation

1. **Location Referencing Methods**

“Location Referencing” can be understood as an umbrella term for techniques and methods to identify a “Location Reference” (Hackeloeer 2016) and is typically used to transfer a “Location Reference” from one map to another implementing a specific algorithm (“Location Referencing Method”).

There are a couple of definitions for the term “Location Reference” (Hackeloeer 2016, Lv et al. 2008, Schützle 2016, Wevers 2000) which could be summarized as follows: a “Location Reference” is the description of a geographical object, basically by its geometrical attributes (e.g. coordinates
or bearing) and could be extended by topological (e.g. node valence/degree), syntactical (e.g. name) and semantical (e.g. physical appearance) attributes.

"Location Referencing" and the algorithm-specific “Location Referencing Method” can be seen from different perspectives: Schützle (2016) and Wevers (2000) differentiate into static and dynamic LR methods. Static LR (or pre-coded) transfers the “Location Reference” between two identical maps using predefined IDs, whereas dynamic LR (or map-agnostic or on-the-fly) transfers between two different maps. Hackelooer (2016) and Zhang (2009) give an additional classification by the type of identification category: Geometrical properties like coordinates or the distances/differences between two given objects. Topological properties of a given object like the number of crossing lines for a given point (node valence, node degree). Semantical properties which describe the meaning of a given object. This could be the name of an object, the importance of a line segment, etc. Some authors define the name of an object as a syntactical or toponomical attribute (e.g. Bellahsene et al. 2011, Monge & Elkan 1996; Zaiss 2010), so the list could be extended by a fourth one. Furthermore, the methods can be distinguished by the type of algorithm or ranking which is used to identify the most likely “Location Reference”: Analytical which is typically done by a weighted power function (or a cost function) with several defined criteria (e.g. Schützle 2016) which will be maximized (or minimized) to find the most likely “Location Reference”. Probability or fuzzy-based which relies on the fact that linking data from different sources typically has some inaccuracy, inconsistency, redundancy, etc. (Gahegan & Ehlers 2000, Samal et al. 2004). Nearly all dynamic “Location Referencing Methods” use an analytical algorithm with one known exception: TPEG2-URL with “Markov Chain” introduced in Kwella et al. (2012).

2. Probability Based Dynamic Location Referencing Method

Based on the remarks in Section 1, the “Probability Based Dynamic Location Referencing Method” will be developed with the specific criteria listed below - the focus is on linear objects (e.g. streets):

- Spatial/geometrical attributes: “node distances between two nodes” and “bearing differences between line segments”;
- Topological attributes: “node valences/degrees”,
- Syntactical/toponomical attributes: “street names”,
- Semantical attributes: “Functional Road Class (FRC)”, “Form of Way (FOW)”, “speed categories” and “one-way-attribute”.

8
The criteria listed result from the outcome of former works in this context (e.g. Schützle 2016, Wartenberg 2007). In addition to the classifications in Section 1, the criteria need to be classified according to the characteristics of the probability distribution: discrete or continuous distributions as well as map supplier specific sample generation and mixed sample generation. Mixed sample generation for criteria are assumed as supplier independent.

<table>
<thead>
<tr>
<th>Geometrical</th>
<th>Topological</th>
<th>Syntactical</th>
<th>Semantical</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample supplier dependent / mixed</td>
<td>mixed</td>
<td>mixed</td>
<td>dependent</td>
</tr>
<tr>
<td>continuous / discrete distribution</td>
<td>continuous</td>
<td>discrete</td>
<td>discrete</td>
</tr>
</tbody>
</table>

Table 1. Additional classifications of LR criteria

3. Statistical Evaluations and Estimated Probability Distributions

3.1. Database and Test Areas
The following databases have been used to create sets of random samples:
- TomTom Multinet for the area of Utrecht, 2008 (TomTom B.V. 2013),
- OpenData City of Cologne, version summer 2016 (Stadt Köln),
- OpenData road network federal state NRW, summer 2016 (GovData),
- OpenData City of Rostock, version summer 2016 (Hansestadt Rostock),
- OpenStreetMap, version by date of generating (summer 2016),
- TomTom Multinet and HERE map for the federal state of Baden-Wuerttemberg provided by the map supplier.
The evaluations have been done for urban and urban hinterland scenarios.

3.2. Geometrical Evaluations
Following Section 2, the geometrical attributes are assumed to be supplier-independent, so a mixed random sample consisting of several supplier combinations with the same size for every subset has been aggregated. For this setting up, 100 corresponding line segments split in equal subsets have been chosen, paired and verified manually. This gives a sample size of \( n_{\text{node}} = 200 \) for node distances and \( n_{\text{bear}} = 100 \) for bearing differences. For the starting point and the endpoint just so-called “real nodes” (node valence = 1 or node valence > 2) have been selected because these nodes define typically significant points (end of a one-way-street, intersections, etc.) in a road network. Nodes with valence = 2 (bivalent) are typically intermediate points (shape points) in the road network. The node distance \( d_{\text{node}} \) has been calculated using Euclidian distance. Taking the UTM projection of
WGS84 coordinates for two given points \( P_1 \) and \( P_2 \) with their Easting \( E \) and Northing value \( N \), the distance is calculated as follows:

\[
d_{\text{node}} = \sqrt{(E_{P_2} - E_{P_1})^2 + (N_{P_2} - N_{P_1})^2}
\]  

(1)

The bearing difference \( d'_{\text{bear}} \) is given by the minimum absolute difference between the azimuths \( \omega \) of the two paired line segments:

\[
d'_{\text{bear}} = \min (360° - d_{\text{bear}}, d_{\text{bear}}) \quad \text{with} \quad d_{\text{bear}} = |\omega_{\text{line1}} - \omega_{\text{line2}}|
\]  

(2)

\[
\omega = \begin{cases} 
90° - \arctan \left( \frac{N_{P_2} - N_{P_1}}{E_{P_2} - E_{P_1}} \right) & E_{P_2} > E_{P_1} \\
270° - \arctan \left( \frac{N_{P_2} - N_{P_1}}{E_{P_2} - E_{P_1}} \right) & E_{P_2} < E_{P_1}
\end{cases}
\]  

(3)

The processing of the calculations of the node distances in m and the bearing differences in ° result in histograms (Figure 1) with relative frequencies. Visualizing the two distributions, the estimation of an exponential probability distribution is an obvious task. This is defined as (Bronstein et al. 1993):

\[
f(x) = \begin{cases} 
0, & x < 0 \\
\lambda e^{-\lambda x}, & x \geq 0
\end{cases} \quad \lambda = \frac{1}{\mu}, \quad \mu = \text{expected value}
\]  

(4)

Since the expected value \( \mu \) is unknown, this parameter has to be estimated. The average value \( \bar{X} \) will converge to the expected value \( \mu \) for infinitive sequence of random variables.

\[
\lambda = \frac{1}{\mu} = \frac{1}{E(X)} \approx \frac{1}{\bar{X}}, \quad \bar{X} = \text{mean of the random sample}
\]  

(5)

The following probability density functions for the node distances and the bearing differences (Figure 1) are estimated using (5) and calculated and visualized by (4). The validity of the exponential function has to be checked by a “goodness of fit”, here the Kolmogoroff-Smirnoff-Test (KST) has been chosen (Hartung et al. 2009). The zero-hypothesis of the KST is not rejected, meaning that the empirical data can be approximated by the estimated exponential function.

Figure 1. Histograms and exponential probability density functions for node distances and bearing differences
3.3. Topological Evaluations

To analyze the distribution of the node valence, a mixed random sample based on the named map supplier in Section 3.1 has been set up. Again, every subset has the same size. In practice, a specific node valence like val = 3 in the source system has been given and the corresponding node valence in a target system has been checked.

<table>
<thead>
<tr>
<th>Node Valence (given) $N_{V_i}$</th>
<th>Node Valence (corresponding) $N_{V_h}$</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.97 0.03 0.00 0.00 0.00</td>
<td>75 (5x15)</td>
</tr>
<tr>
<td>3</td>
<td>0.00 0.84 0.16 0.00 0.00</td>
<td>75 (5x15)</td>
</tr>
<tr>
<td>4</td>
<td>0.00 0.17 0.83 0.00 0.00</td>
<td>75 (5x15)</td>
</tr>
<tr>
<td>5</td>
<td>0.00 0.07 0.64 0.29 0.00</td>
<td>75 (5x15)</td>
</tr>
<tr>
<td>6</td>
<td>0.00 0.10 0.54 0.22 0.14</td>
<td>50 (5x10)</td>
</tr>
</tbody>
</table>

Table 2. Rel. Frequencies of node valences

Since the number of nodes with a valence = 6 is typically much lower compared to the others, a smaller but sufficient sample size has been selected. For this kind of discrete data, the corresponding value (e.g. $N_{V_h} = 4$) for a given value (e.g. $N_{V_i} = 3$) is binomially distributed with respect to all other values for this given value. This simplifies the multinomial distribution. The probabilities could be derived from the relative frequencies $f_r$ of a random sample (Hartung et al. 2009). This means in detail that for a specific combination “given criteria – corresponding criteria” the derived probability $\pi_{i,h}$ results from the relative frequency $f_r$ of this combination based on the random sample.

$$\pi_{i,h} = f_r$$  \hspace{1cm} (6)

With this derivation, the relative frequencies in Table 2 can be taken directly as probability distributions.

3.4. Other Evaluations

As pointed out in Section 2, syntactical and semantical criteria have been specified to be used. Since these criteria are discrete (Section 2), they have been analyzed analogous to the topological ones (Section 3.3) considering the classifications with respect to mixed or supplier dependent samples as well as continuous or discrete distributions.

4. Summary

In this paper, the statistical approach to find a valid base to estimate probability distributions as a base for a “Probability Based Location Referencing
Method” has been presented. For the first time, statistical distributions for different characteristics are published.

References

GovData (2016) Straßenetz Landesbetrieb Straßenbau NRW. https://www.govdata.de/daten/~/details/ac8a18de-29d2-4bd4-ba75-6a1d4b4abff. Accessed 1 April 2017
Wevers K (2000) Improved on-the-fly Location Referencing. Presented at the ITS World Congress, Turin