

# Calculating Route Probability from Uncertain Origins to a Destination

Carolyn von Groote-Bidlingmaier, David Jonietz  
and Sabine Timpf

**Abstract** Uncertainty in location information can affect the results of network-based route calculations to a high degree. In this study, a routing scenario is analysed where the destination is known, but the location of the point of origin can only approximately be described as “somewhere inside a polygon”. Using the concrete example of car-driving football fans arriving at a game, an approach is proposed to compute and describe the probability of them taking specific routes from their home county to the stadium. A set of candidate points of origin is created and shortest paths to the destination calculated. The observed frequency of an edge being included in a route allows inferring a routing probability for each edge. Several methods to derive a set of candidate points of origin are presented and discussed, ranging from purely geometrical to geographically weighted approaches. Our results show that the differences between the methods in determining the points of origin produce only slightly different probabilities, i.e., neither advantages nor drawbacks are to be expected from using a purely geometrical approach.

**Keywords** Routing · Probability · Network analysis

## 1 Introduction

The fact that almost all spatial information is marked by some degree of uncertainty poses particular challenges for its further use, for instance in the context of spatial analysis. The sources of uncertainty are manifold, ranging from data collection to

---

C. von Groote-Bidlingmaier (✉) · D. Jonietz · S. Timpf  
Geoinformatics Group, Department of Geography, University of Augsburg,  
Augsburg, Germany  
e-mail: carolin.vongroote-bidlingmaier@geo.uni-augsburg.de

D. Jonietz  
e-mail: david.jonietz@geo.uni-augsburg.de

S. Timpf  
e-mail: sabine.timpf@geo.uni-augsburg.de

analytical methods. One can distinguish between different types of uncertainty, including inaccuracy and error, vagueness and incompleteness (Worboys 1998). In the context of location-based services (LBS), for instance, positioning errors could result from poor GPS accuracy, conceptual vagueness from the use of imprecise spatial statements such as “near” or “far” or incompleteness in the case of missing spatial information, such as network attributes in a routing task (Basiri et al. 2012).

## 1.1 Uncertain Routing

Routing is included in the functionality of many LBS as well as a basic task for transportation planners and involves the use of algorithms to identify an optimized path from an origin to a destination through a weighted network. While in the context of LBS, the focus is typically on the individual traveller, transportation modelling generally aims at forecasting travel demand and its resulting effects such as traffic volumes at a more aggregate level (MacNally 2007). In both cases, depending on the optimization criteria, possible solutions can include the shortest, fastest, cheapest or other paths (Miller and Shaw 2001).

With regards to uncertain data, research has so far focused on edge attributes (e.g., Liao et al. 2014), destination points (e.g., Qiu et al. 2013) as well as positional uncertainty (e.g., Gonzales and Stenz 2007; Hait et al. 1999). The paper is situated in this thematic context but starts from a different flavour of uncertainty. A routing scenario is discussed which focuses on identifying the most probable route from an area of origin to a predefined destination. While the destination is known and can be represented as a fixed node on the street network, the origin of the route can only be approximately described as a distinct polygon due to either incomplete or inaccurate spatial information. In fact, a very similar challenge is posed to transportation planners during the process of route assignment, when trips are modelled for distinct traffic analysis zones (TAZ) and assigned to the road network in order to predict traffic volumes (MacNally 2007). In this case, points of origin are seldom known (e.g., home locations), but in most cases approximated as the centroid of the TAZ is used for the routing (Qian and Zhang 2012). In the past, there has also been work on the possible influences of using a population- or household-weighted centroid instead of a purely geometric one (Chang et al. 2002). In the case study presented in this paper, the aim is to determine the most probable route taken by car-driving visitors to a large-scale event, in this case a football game at a stadium. The exact origins of the visitors are unknown, although information about their home county can be retrieved from their license plates. The problem of calculating probable routes from an uncertain origin, however, is not just restricted to transportation planning, but could also be applied to positional inaccuracy in the context of LBS, i.e., using the error ellipse (circle) of uncertain GPS positioning.

In this paper we propose the use of a large set of candidate points instead of merely one representative centroid. We introduce a geographically weighted approach, which we compare with purely geometric methods in order to evaluate its

appropriateness and practicality for predicting the usage of specific routes to a stadium for a mass event. In contrast to determining a most probable starting point and calculating a single route, as in the case of TAZ centroids, our approach has the advantage to produce probabilities for each segment of all routes leading to the destination. Thus, without the need for detailed data about exact home locations, planners and mass event managers may identify potential bottlenecks and reduce congestion by applying appropriate measures and LBS may take this knowledge into account when suggesting a specific route.

## ***1.2 Approaches to Uncertainty in Spatial Data***

Today, the term uncertainty is increasingly used instead of error to describe the difference between spatial data and their corresponding real world entities, thereby acknowledging the fact that representations are always merely approximations of the truth (Zhang and Goodchild 2002). Except for geographical abstraction, uncertainty may also result from processes of approximation, measurement or generalization (Goodchild 2009). At the same time, however, users of GIS are often unaware of these issues, a fact which may at worst lead to decisions being made on the basis of questionable spatial information. As a result, extensive research has focused on measuring, modelling, visualizing and analysing the propagation of uncertainty in spatial data.

In the specific context of modelling uncertainty, a range of theoretical frameworks have been applied, among others probability theory and statistics, fuzzy set theory, rough sets, as well as possibility theory. While fuzzy and rough sets are useful tools to extend set membership and set boundaries beyond crisp values, and possibility theory is concerned with measuring the feasibility of some event, probability theory describes methods to calculate its probability (Wang et al. 2005). Traditionally, values ranging from 0 (null event) to 1 (entire sample space  $\Omega$ , i.e., every possible outcome) are used to denote the probability  $p(E)$  of a particular event  $E$  (Jaynes 2003). Among several others, there is a most common view of  $p(E)$ , which is prevalent in statistics and relates it to the relative frequency  $f(E)$  of the occurrence of  $E$  in a number of trials of an experiment, such that  $p(E)$  equals  $f(E)$  (Hajek 2012).

This paper is structured as follows: we first present the case study that inspired us to start this investigation. In Sect. 3, the method for calculating route frequency is described in detail showing the two steps of first determining potential points of origin and second calculating the probability from the frequency. Sect. 4 compares and discusses the results from the different methods and Sect. 5 presents conclusions and future work.

## 2 Case Study: Predicting Usage of Access Routes to a Stadium

The specific problem presented in this study is set in the context of traffic management for large-scale events. In our specific study area, authorities have to cope with the dense flow of visitors moving to and from the specific site, in our case a soccer stadium. Figure 1 shows the street network of the counties (Landkreise) Augsburg, Aichach–Friedberg, and Dachau as well as the location of the SGL arena in the south of Augsburg.

Despite increased provision and marketing of public transport, many visitors still arrive in their own car and within a relatively short time interval. Thus, volumes of motorized traffic are multiple times higher than usual. In order to avoid negative effects such as traffic jams and prolonged waiting times at the main access points or unannounced shortages of parking space, it would be useful for planners to predict the visitors' access route to some degree, for instance in order to use traffic guidance systems to increase the efficiency of traffic flows in the vicinity of the stadium. For this, however, information about the visitors' access route is critical. Personal interviews are not practical, mainly due to the sheer mass of visitors. However, limited information about their home county can be retrieved from their license plates. In this case there are no origin-destination pairs, but the location of the origin



**Fig. 1** Overview over the study area in Bavaria. The *red circle* indicates the SGL arena and the *green area* is the extent of the county Aichach–Friedberg, Bavaria

can be approximated using the administrative border of the respective county. Thus, predicting access routes poses a practical problem to the analyst. With regards to the mere extent and location of the polygon relative to the destination point, it can be argued that the use of a centroid as representative point of origin seems inappropriate in this case. In fact, our aim is to analyse the influence of different spatial distributions of the origins within the polygon of origin on the most probable route. In the following calculation we work with the assumption that the origin  $v'$  is supposed to be somewhere in Aichach–Friedberg and the destination  $v_d$  is the SGL arena in the south of Augsburg (red circle in Fig. 1).

### 3 Calculating Most Probable Routes

Based on the intention to identify the most probable route from an uncertain point of origin to a predefined destination, our proposed approach is as follows: We assume a street network represented in the form of a graph  $G = (V, E)$ . There is an unknown point of origin  $o \in V$ , the position of which is approximated by a set of points within a polygonal boundary  $P$  such that  $o \in P$ . Furthermore, there exists a known destination vertex  $v_d$ , which is part of the set of points on the graph but outside the set of points representing the polygon:  $v_d \in V$  and  $v_d \notin P$ . Since  $o$  is unknown, we introduce a subset  $V' \in V$  and  $V' \in P$  of vertices as candidate origins on the street network or projected onto the network and within the polygon, and compute a set of shortest paths  $SP = \{v'_0 e_1 v_1 e_2 \dots e_k v_d \text{ where } e_i = v_{i-1}^{(i)} v_i, \forall 1 \leq i \leq k\}$  from each  $v' \in V'$  to  $d$ . Of course, the restriction to the network distance as single impedance represents a simplification, especially since a wide range of additional path optimization criteria such as time or easiness are used by humans during route planning. In addition, assuming that humans have perfect knowledge about the road network is certainly not fully realistic (Prato 2009). Considering the issue of model simplicity, however, as well as the fact that distance is in fact widely used as impedance in transportation models and thinking of the decision maker as a LBS, these limitations seem acceptable. On this basis, we calculate normalized values ranging from 0 to 1 for each  $e \in E$  and  $e \in P$  to denote the relative frequency  $f(e)$  of  $e \in SP$ . Following the most prevalent interpretation of probability, we argue that the higher the value of  $f(e)$ , the higher the probability  $p(e)$  of  $e$  to be visited when a traveller starts a trip to  $v_d$  from an unknown  $o$ . The values  $f(e_{1 \dots n})$  are used in the second step as impedance for a route calculation from  $s$  to the furthest  $v'$ , in order to determine the most probable route for all member vertices of  $V'$ .

#### 3.1 Methods of Generating Points of Origin $V'$

The results of the computation described above can be expected to depend to a high degree on certain characteristics of the vertex subset  $V'$ , such as the number of

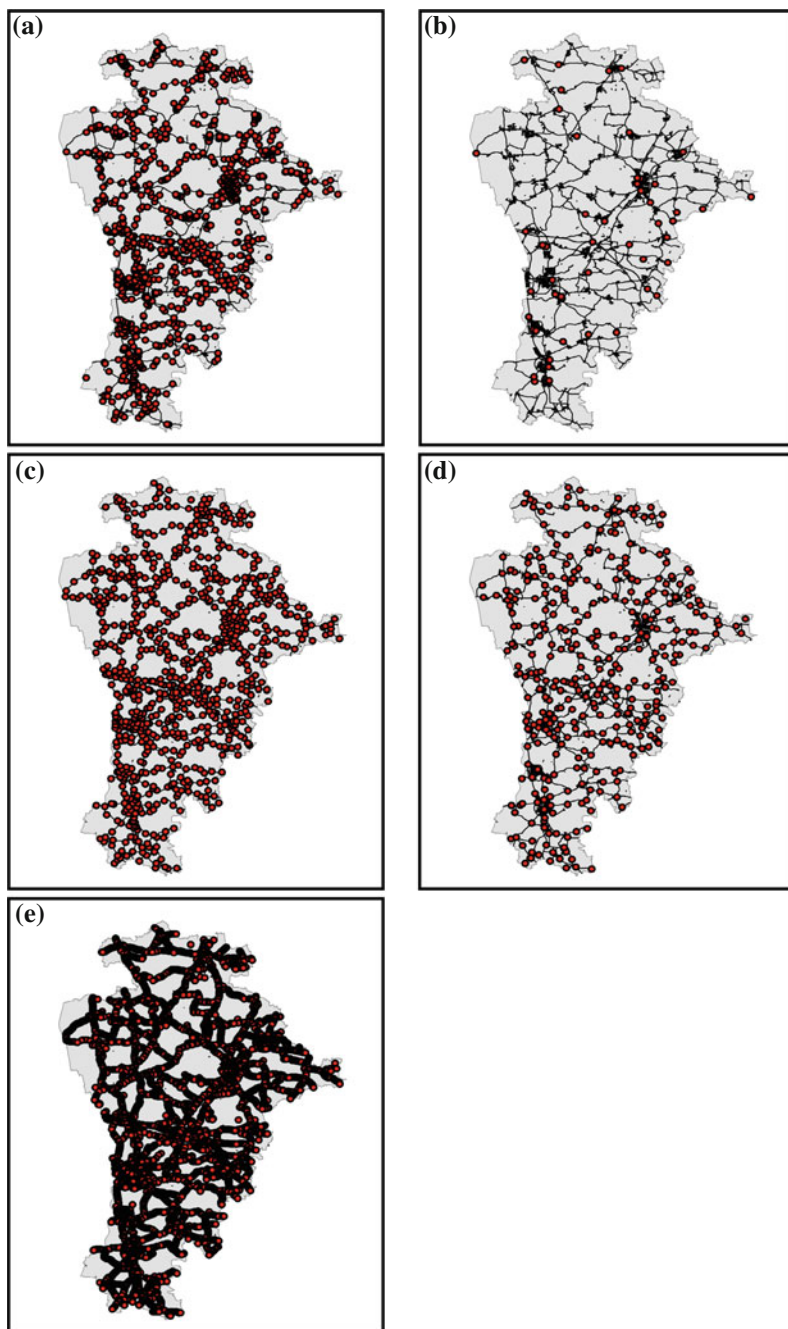
vertices included, their mutual distance and their distribution on the road network. Moreover, in reality, the chance of being the point of origin for a trip is not equal among all  $v'$ . Rather, vertices located in areas of high population or road density can be assumed to be of a higher relative importance for the calculation than others, a fact which can be expressed in the form of individual weight coefficients assigned to specific vertices. In this study, a range of possible approaches is analysed, including purely geometrical and geographically weighted methods. For the purpose of this paper we chose 50 and 1,000 randomly distributed points on the network, regularly dispersed points at a distance of 500 and 1,000 m in a weighted and unweighted version (see Sect. 3.1.2), and points derived from street network density (see Sect. 3.1.3).

Independent from the actual method of creating  $V'$ , the aim was to minimize the number of points in order to reduce the computational effort required for the analysis. Since the only impedance value used for the routing process will be network distance, it can be assumed that shortest paths will be similar for all  $v'$  which are located on the same edge  $e_i$ . Accordingly, if for a particular  $e_i$  the number of newly created vertices exceeds 1, i.e.,  $N(v') > 1$ , they will be combined to one representative vertex and their absolute number stored as an attribute. In case of weight coefficients being assigned to  $v'$ , the mean weight value will be calculated and saved as well, to make sure that the following calculations are not affected by any loss of information.

### 3.1.1 Randomly Distributed Points of Origin

As stated previously, among other potentially influential factors, we expect the results of the route calculations to depend on the number and distribution of  $v'$  as well. As a first approach, therefore, 50 and 1,000 random vertices were created on the network (see Fig. 2a, b), and used as input origins for a shortest path calculation to the destination  $v_d$ . To acknowledge the potential effect of the number of  $v'$ , several trial runs were conducted.

According to the law of large numbers, a general principle which describes how, in a large number of trials, the observed frequency of an event will tend towards its theoretical probability, it can be expected that above a certain number of  $v'$ , there will be no significant change in route frequency (Hazewinkel 2001). In this routing scenario, no additional weighting function for generating the vertices was applied. There is, however, an implicit influence of street network density, since on denser segments of the graph the chance of a point of origin being created is higher. The influence of population density will be included in the second approach, described in the following section.



**Fig. 2** Points of origin derived through different methods **a–e**, **a** randomly distributed points of origin-1,000 points, **b** randomly distributed points of origin-50 points, **c** regularly dispersed points of origin with 500 m, **d** regularly dispersed points of origin with 1,000 m, **e** points of origin derived from street network density



### 3.1.2 Regularly Dispersed Points of Origin

In contrast to a random distribution, for this approach an approximately even dispersion of points of origin was chosen. Vertices  $v'$  were created with a mutual distances of 500 and 1,000 meters on the graph edges  $e \in E$  and  $e \in P$  (see Fig. 2c, d). In order to avoid higher point densities in the vicinity of edge junctions, additional processing steps were necessary to arrive at a point dispersion, which approximates an even distribution on the network. In a first step, shortest paths were calculated from each  $v' \in V'$ . In a second step, weights were assigned to each point of origin based on the normalized population numbers of their respective municipal area, since we assume the chance of starting a trip as well as the number of potential visitors to be higher in more densely populated areas. The process of calculating route frequency, a step which is described in Sect. 3.2, incorporates the resulting weight coefficients.

### 3.1.3 Points of Origin Derived from Street Network Density

A further approach involves deriving the points of origin based on the density of the street network. This approach has the advantage that no other input data are necessary while, at least to some degree, an approximation to population density distribution can be achieved. First, the polylines from the street network are used to calculate line density values and then normalized to a range from 0 to 1 on a cell-by-cell basis. The following formula is used for the normalization:

$$D_{norm} = \frac{D_i - D_{min}}{D_{max} - D_{min}}.$$

Subsequently, all cells with  $D_{norm} > 0$  are converted to points and projected onto the closest edge  $e \in P$ . Thus, in this specific approach, urban agglomerations are not only favored by higher weight factors  $D_{norm}$ , but also by the fact that within these areas, a comparatively higher number of points are created (see Fig. 2e).

## 3.2 Calculating Route Frequency

Based on the created set of points of origins  $V'$ , shortest paths are calculated for each origin-destination pair  $v'-v_d$ . Each segment on the path is assigned an individual weight coefficient, which corresponds to 1 for unweighted origins and to the weight of the origin in the weighted calculations. Route frequency  $f(e_{1...n})$  is measured by calculating the number of overlapping routes per network segment, counting each route either once (following unweighted approaches) or according to its individual weight coefficient.



As has been described previously, in the case that several  $v'$  were located on one edge  $e$ , in order to minimize computational effort, the vertices are combined into one representative point. In this case, the accumulated mean weight value will determine how one route is counted when calculating the frequency.

According to the frequency interpretation of probability, one can infer the probability  $p(e_{1...n})$  from the frequency  $f(e_{1...n})$ . Finally, both a prediction of vehicle distribution in the network and the most probable route can be deduced from the resulting network using probabilities as weights.

## 4 Results and Discussion

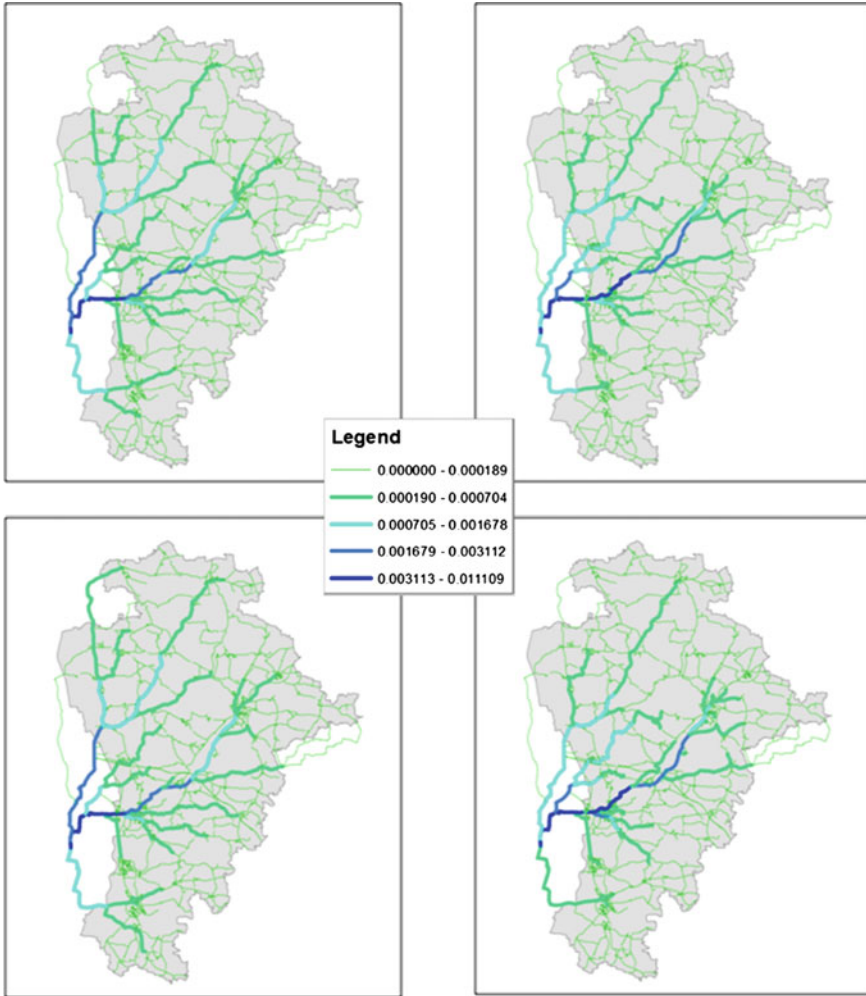
As expected, the results vary depending on the approach used. The main differences occur between the weighted and unweighted methods, whereas the contrast between the various weighted methods is comparatively small (see Fig. 3). When using the unweighted regularly-dispersed-method the calculated routes are more dispersed on the network. In comparison, the calculated routes for the weighted regularly-dispersed-method lead through the agglomeration areas (in this case the two bigger cities Aichach and Friedberg). The routes from unweighted methods also pass more often through the northern network, whereas the routes resulting from the weighted methods make more use of the southern network. However, the southern network is less frequented with the weighted 1,000 m method than with the weighted 500 m method. This distinction might make a difference in planning and managing traffic for a huge event at the stadium.

If the number of random points is too small (i.e., 50), all routes that are calculated are represented as higher frequency (compared to no frequency), whereas in the case of 1,000 random points enough routes are calculated across a segment that a meaningful frequency pattern can emerge and outliers are detected as such.

The method using 1,000 random points and the method using line density produce almost identical frequencies (see Fig. 4). This is remarkable since the line density uses about 7,000 points in contrast to the 1,000 points of the random method. An explanation might be that the randomly distributed points on the network implicitly factor in the density of the network, because they are projected onto the network after creation.

The southern network is most frequented in the 50 random, 1,000 random and line density approaches (the last southern segment is dark blue in contrast to light blue or even green in the point dispersed approaches).

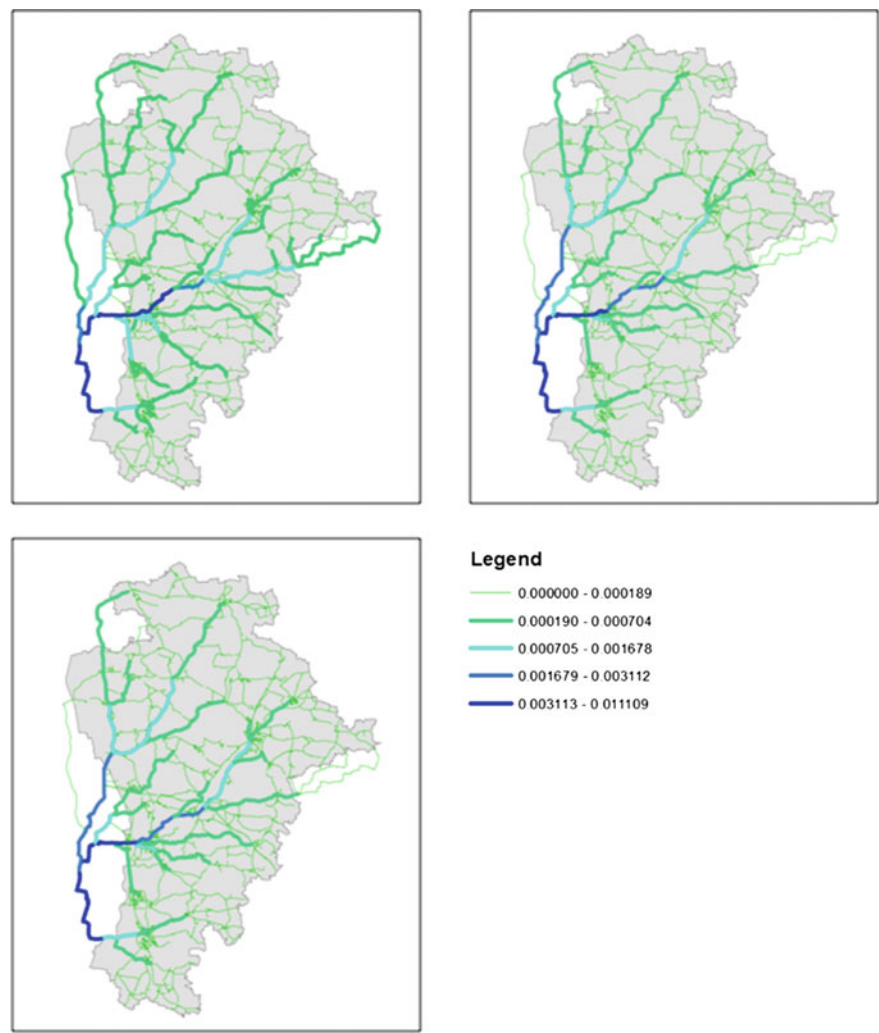
Factoring in the aim to use as few additional sources as possible and to reduce computing time without losing information an approach based on deriving the potential points of origin from the street network seems most reliable. With respect to the original question of using geometric versus geographically weighted information, we have to conclude that (at least in this case study) the geometric information is sufficient to produce a satisfying route probability.



**Fig. 3** Frequency of routes in the network—calculations for regularly dispersed origins. *Top row* 500 m distance, *bottom row* 1,000 m distance, *Left column* unweighted, *right column* weighted

#### 4.1 Comparison of Frequencies

The frequencies are calculated for each segment, i.e., edge, within the network. For a comparison of frequencies per segment, a normalization procedure needs to be carried out. We have used two methods for normalization. First, we normalize across the distribution of frequency values of all segments within the network, i.e., computing a relative frequency: within calculation method 1 (e.g., 1,000 randomly distributed points) look for the maximum value  $frequ_{max,method1}$  and divide the



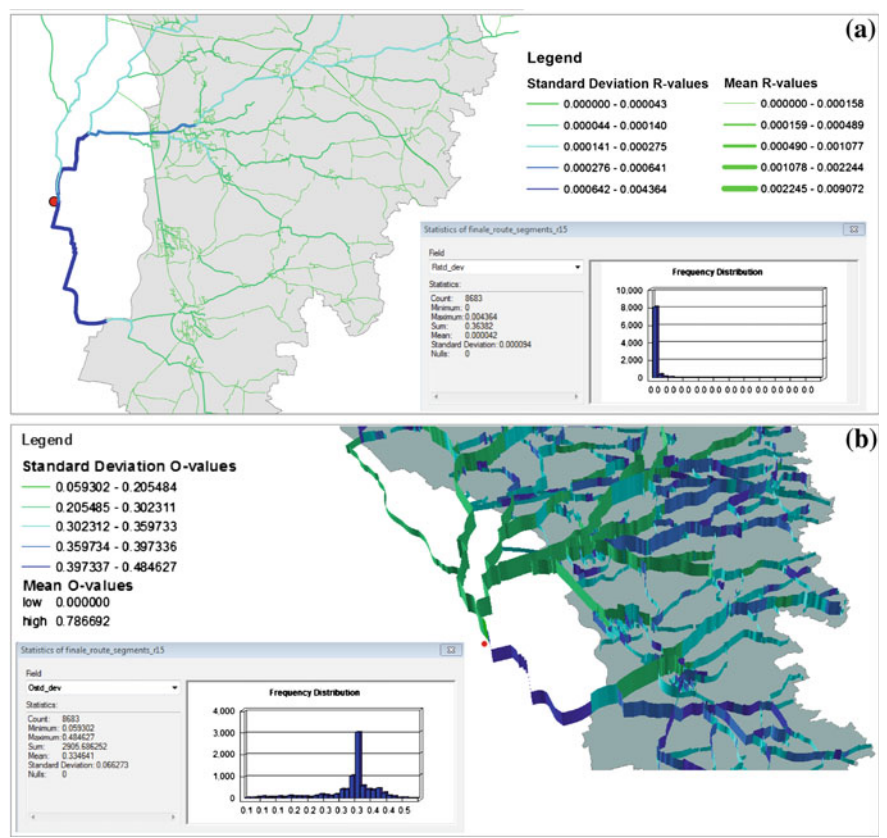
**Fig. 4** Frequency of routes in the network—calculations for randomly dispersed origins (*top row left 50, right 1000 points*) and line density

calculated frequency  $frequ_i$  of each segment  $i$  by the maximum value, thus yielding a relative frequency for each segment  $relfrequ_i$  within the range 0–1. We call this relative frequency the R-value of a specific segment within a specific method.

Second, we normalize across all existing relative frequency values on one single segment for each of the seven methods used to determine the distribution of the points of origin. Thus, we divide the  $relfrequ_i$  of a segment  $i$  by the maximum relative frequency across all seven methods for a specific segment  $i$ , resulting in a comparison of the relative frequency across all methods. This is only possible

because the relative frequencies are normalized and thus within the range of 0–1. We call this the O-value.

In addition to the relative frequency per method (R-value) and the relative frequency per segment (O-value), a mean value of all relative frequencies as well as a standard deviation of the relative frequency can be calculated for each segment, again within the network (R-values) and across all methods (O-values). In Fig. 5 it is immediately visible that the mean relative frequency is particularly high in the last segment from the south although this segment only caters to a small portion of the total area of potential origins. In the 3D figure the last segment from the south shows a high standard deviation, which means that it is the least stable segment within the network. Since this segment carries a relatively high load, any measures of event managers or traffic planners should consider this segment very carefully.



**Fig. 5** Mean and standard deviation of route frequency: R-values and O-values. **a** Width of line equals mean of R-value per segment and **b** height of line equals mean of O-value per segment. Colour represents standard deviation

## 5 Conclusions and Future Work

The aim of this research was to derive the variability in the most probable routes from a specified area (represented as polygon) to a specific destination without knowing the exact points of origin. This kind of question occurs for example during mass events, when many people from different geographic areas arrive for the event. Alternatively, there could be an inability to derive the current location, e.g., due to continued missing GPS signals, while still needing to provide a route to a specified destination. Within the context of event management, event planners want to be able to provide participants with up-to-date information along the routes they most probably will take. Traffic managers would use the information on the routes most probably taken to manage traffic in order to avoid congestion or at least manage the flow of vehicles.

The approach taken here differs from traditional route assignment since it derives a set of potential points of origin through several methods and calculates a set of shortest routes starting from these points of origin. Each of the segments within the road network can then be attributed with the number of routes passing through the segment. This number is then used to calculate a probability of the segment being part of a route to the destination starting from within the specified area of origin.

Our results show that the differences between the methods in determining the points of origin produce only slightly different probabilities. Considering that the calculation of a potentially high number of shortest paths takes up resources, we chose as the best the method with the least amount of necessary calculations. For this case study at least it turns out, that the geometric information, i.e., the road network itself, is sufficient to generate the probability values for route segments.

Taking the results a step further, we could determine the most probable route through the whole road network, starting from the destination  $v_d$  and using the calculated route segment probabilities. In this case a breadth first search needs to be conducted which allocates probability values to each edge. The result shows the most probable route from an area to the destination.

What should be done in the near future is the comparison of the probabilities of the route segments with centrality network measures as well as results of TAZ centroid-based routing processes at a number of structurally different networks. This would correlate the geometric properties of the road network structure with the usage properties as well as test the comparability of methods. The results could be validated for example by deriving the observed frequency of use of a specific road segment from floating car data or traffic counts obtained from road-side traffic counters.

Additional variations in the calculations performed in this study may be enlightening: a distinction between road types in the shortest path (i.e., hierarchical shortest path) or using different path optimizations (fastest, most beautiful, least complex...) could produce a different result in the final probabilities. Another distinct approach could be the use of a probability surface instead of a distinct polygon in the delineation of the origin area. Future work could also encompass the derivation of the most central route through the network, resulting in a centrality

measure for routes within the network. Another variation of the computation could incorporate temporal measures instead of distance measures for the calculation of the relative frequencies.

**Acknowledgements** We gratefully acknowledge support of Carolin von Groote-Bidlingmaier through the program “Chancengleichheit von Frauen in Forschung und Lehre” of the University of Augsburg.

## References

- Basiri A, Winstanley A, Sester M, Amirian P, Kuntzsch C (2012) Uncertainty handling in navigation services using rough and fuzzy set theory. In: Kroeger P, Renz M (eds) QUESST '12 Proceedings of the Third ACM SIGSPATIAL international workshop on querying and mining uncertain spatio-temporal data. Redondo Beach, CA, USA, 07 Nov 2012
- Chang KT, Khatib Z, Ou y (2002) Effects of zoning structure and network detail on traffic demand modelling. *Environ Plan* 29:37–52
- Gonzales JP, Stentz A (2007) Planning with uncertainty in position using high-resolution maps. In: Proceedings IEEE international conference on robotics and automation, Rome, Italy, 2007
- Goodchild MF (2009) Methods: uncertainty. In: Kitchin R, Thrift M (eds) *International encyclopedia of human geography*. Springer, New York
- Hait A, Simeon T, Taix M (1999) Robust motion planning for rough terrain navigation. In: Proceedings. IEEE/RSJ International. Conference. Robotics and Systems, Kyongu, Korea
- Hajek A (2012) Interpretations of Probability. In: Zalta EN (ed) *The stanford encyclopedia of philosophy*. <http://plato.stanford.edu/entries/probability-interpret/#FreInt>. Accessed 30 May 2014
- Hazewinkel M (2001) Law of large numbers. In: Hazewinkel M (ed) *Encyclopaedia of mathematics*. Springer, Berlin
- Jaynes ET (2003) *Probability theory—the logic of science*. Cambridge University Press, Cambridge
- Liao F, Rasouli S, Timmermans H (2014) Incorporating activity-travel time uncertainty and stochastic space-time prisms in multistate supernetworks for activity-travel scheduling. *Int J Geogr Inf Sci* 28(5):928–945
- MacNally MG (2007) The four step model. In: Hensher DA, Button KJ (eds) *Handbook of transport modeling*. Elsevier, Oxford
- Miller HJ, Shaw S-L (2001) *Geographic information systems for transportation: principles and applications*. Oxford University Press, Oxford
- Prato CG (2009) Route choice modelling: past, present and future research directions. *J Choice Modeling* 2(1):65–100
- Qian ZS, Zhang HM (2012) On centroid connectors in static traffic assignment: their effects on flow patterns and how to optimize their selections. *Transp Res Part B* 46:1489–1503
- Qiu D, Papotti P, Blanco L (2013) Future locations prediction with uncertain data. In: Blockeel H, Kersting K, Nijssen S, Zelezny F (eds) *Machine learning and knowledge discovery in databases*, LNCS 8188. Springer, Berlin, pp 417–432
- Wang S, Wenzhong S, Yuan H, Chen G (2005) Attribute uncertainty in GIS data. *Fuzzy Syst Knowl Discov*, LNCS 3614:614–623
- Worboys M (1998) Imprecision in finite resolution spatial data. *Geoinformatica* 2(3):257–279
- Zhang J, Goodchild MF (2002) *Uncertainty in geographical information*. Taylor and Francis, New York

Progress in Location-Based Services 2014

Gartner, G.; Huang, H. (Eds.)

2015, XII, 282 p. 111 illus., 92 illus. in color., Hardcover

ISBN: 978-3-319-11878-9