Paper submitted to PTRC conference, Brunel University, 1-5 September 1997

# PEDESTRIAN ROUTE CHOICE: AN EMPIRICAL STUDY

Neville Q Verlander
Benjamin G Heydecker
Centre for Transport Studies, University College London

## 1. INTRODUCTION

There has been relatively little work done on route choice for *pedestrians*. The present paper addresses this issue by using a sample survey of daily walks in a UK urban area. The walks undertaken are reconstructed using a geographical information system and compared with the shortest available route. It was found that about 75 per cent of walkers in the sample chose the shortest available route. Two strategies were used to synthesise sets from which pedestrians could have chosen their routes. These choice sets can then be used in discrete choice modelling to study route choice and to determine which factors are important to pedestrians in this. At the time of writing, it is proposed to proceed with this modelling.

The structure of the paper is as follows. Section 2 describes the various sources of data used in this work, section 3 discusses the choice set generation strategies that were developed, section 4 briefly compares the walks with the corresponding shortest routes, while section 5 presents the conclusions that were drawn from this.

## 2. SOURCES OF DATA

This section describes the three kinds of data that were used in this work and their integration: these are the pedestrian survey data, the Ordnance Survey data in Landline digital form, and paper maps. The survey data informs about where people walked and the routes that were taken, the Ordnance Survey Landline data enables the generation of a set of possible routes from which pedestrians chose and the paper maps serve as visual aids. These are discussed in greater detail in subsections 2.1, 2.2 and 2.3 respectively.

### 2.1 Pedestrian Survey Data

The survey data used were derived from diary records of all activities on foot undertaken during a 24 hour period by members of about 400 households containing 737 individuals and were collected during 1992 in Northampton. The primary purpose of the data collection was the study of pedestrian activity and its relationship with accident risk, which is fully described in Ward, Cave, Morrison, Allsop, Evans, Kuiper and Willumsen (1994). These data are useful for our purposes as they inform about the walks, and in particular the routes taken, by each individual during a specified day. This is done by detailing the roads either traversed along or crossed for each walk by using a unique five-letter code for each road name. In the case where a person walked on a path, track, park or any other open area where individuals may walk other than a road, then the term "footpath" was used with the corresponding special road name code "00000". The term "line segment" is used here to denote part of a road or footpath.

There are about 2200 walks in the dataset, of which about one third have a total distance walked of less than 50 metres and over two thirds less than 350 metres. A detailed breakdown of all the walks by sex, age, types of road crossed, etc is provided in Ward et al (1994).

Many walks (about 250) were non-contiguous as they appeared in the database. Correction of this entailed making additional entries so as to reflect all the roads and footpaths traversed by the pedestrian in the course of their journey. The original questionnaires were used to decide which additional road name codes were required.

Because the purpose for collection of the data was pedestrian activity, not all of the 737 individuals were always walking; for example, an individual might have been washing the car in the street. Any such non-walking activities are not of interest to us for this work and so were excluded from the analysis. Other walks eliminated were those that took place outside Northampton and those that were of excessive duration, as judged in relation to the distance walked.

The road names were (mainly) entered in the order in which they were either traversed or crossed for each trip. It was thus decided to take as the starting point of the part of each journey to be studied in detail the intersection of the first pair of differing road names mentioned for that walk. Likewise, the finishing point was taken as the intersection of the last pair of differing road names mentioned for that walk. This approach has the effect of giving only the inner part of the walk that was undertaken, as the individual is unlikely to have started and ended at an intersection. However, this recovers that part of the journey with a high degree of confidence. The advantage of the method was the ease of locating the origin and destination of each partial walk.

Several difficulties arose with this strategy. Of the approximately 2200 walks, about 1200 used only one or two roads or footpaths, 250 had at least one non-intersection of two roads, 80 had more than one intersection for either the starting point, ending point or both, 40 had the same start and end point, and 60 had at least one of the intersection roads as a driveway, where a driveway is a road without a name that goes into premises such as a petrol station or supermarket and which carries a reasonable amount of traffic. In addition, 330 had an intersection of a footpath with a road. This last category typically resulted in more than one intersection point with a road or no intersection at all. In these cases, the Ordnance Survey paper maps mentioned in subsection 2.3 were used to determine the location of the intersection point. There were roughly 200 walks where this strategy yielded the origin and destination uniquely.

The first set of 1200 walks was discarded because of the impracticality of determining the route taken. Because these walks are generally short, there is no sensible choice between routes, so this does not represent any great loss. The second set of 250 was dealt with by examining the questionnaires and the database. In this case, either typing errors, non-contiguity or incorrect order of road names were found to be the cause of the apparent non-intersection. The third set of 80 was investigated making use of the questionnaires to decide the most likely intersection. Some of the fourth set of 40 were round trips and so were subdivided into an "outgoing" and an "incoming" partial journey. Finally, the fifth set of 60 required additional programming to ignore the driveways. During the course of this work, further walks were excluded from analysis due either to lack of clear specification of the route taken or because the respondent was just wandering, in other words, no route choice was involved; about 30-40 dog walks were deemed to be in this latter category. These corrections and deletions resulted in 820 walks that were considered to be suitable for further analysis.

### 2.2 Ordnance Survey Landline Data

Digital maps in the form of Ordnance Survey Landline data in NTF format (OS) were used for those areas of Northampton where the sample of pedestrians had walked. These maps comprised 206 tiles, most of which are at the 1:1250 urban scale and the remaining few are at the 1:2500 rural scale. The dates of these tiles vary from October 1992 to June 1996 with most in the years 1993 to 1995. These data then postdate the walk data in all cases.

The relevant map features for pedestrians were road centre-lines and footpaths which were extracted from the OS data. Road edges, buildings and non-footpath data were also extracted to facilitate the identification of footpaths, and to yield a contiguous set of road edges; this is discussed further below. These data were imported into the ArcInfo (ESRI, 1992a) geographical information systems (GIS) package using an NTF to ARC converter.

As mentioned in subsection 2.1, the walk diaries inform about which roads were walked along. Each road name needed to be associated with all of the road sections of that particular road in order that the GIS could reproduce the route taken by each pedestrian. All told, there are 545 different roads mentioned in the walk diaries, yielding 4099 named road sections.

A problem was encountered with the OS data in extracting the footpaths. Each type of OS data has what is called a *feature code* associated with it. Unfortunately, footpaths appear under several different feature codes and, for each code, is only one of several different data sub-types; for example, power lines might have the same code as tracks under the feature code "general line detail". Furthermore, there is a hierarchy which determines what kind of data are shown; for instance, if a building overhangs a portion of road, the building will be recorded but the road will not.

In an attempt to build as complete a footpath network as possible, the following strategy was pursued. First, for each eligible feature code, those features labelled "path", "track", "park" and so on were extracted. However, this procedure omitted numerous footpaths because they are not labelled as such. Thus, in an effort to get as many of these latter features as possible, a visual examination of a digital plot of the buildings, roads and footpath and non-footpath data for Northampton was undertaken and the appropriate features were then captured interactively. Line segments that were within 5 metres of each other were then joined together, using the ArcInfo "arcsnap" operator. Finally, working in conjunction with the paper maps to be described in subsection 2.3 and the walk diaries, further line segments were added manually where this was necessary to achieve adequate coverage.

In order to associate a name with a road section, each name has to be linked to the corresponding road section number. This was difficult to achieve with road edges due to the way they are digitised. Thus it was decided to use road centre lines rather than road edges. One problem with this was that in the OS map data, some road centre lines had feature codes corresponding to road edges and vice versa. A further problem was ensuring that each road name is linked to *all* the corresponding road sections comprising that road. The snapping mentioned above resulted in the characteristics of some of the line segments making up a road being replaced by the characteristics of a footpath. This was clearly undesirable and so required redefinition of the characteristics for the affected line segments. All anomalies of these kinds had to be identified and corrected so as to have a contiguous walk network of road centre lines, paths, tracks, fields and other open areas. This needed to be done prior to reconstruction of walks undertaken and further analysis.

Once the list of walks with their starting and ending positions had been constructed as detailed in subsection 2.2, ArcInfo was then used to reconstruct the actual routes taken on the walk network. This was done by calculating the shortest path between the origin and destination points using only those roads and footpaths mentioned by the respondent for that journey. Discussion of the shortest path algorithm used by ArcInfo is provided by ESRI (1992b). Furthermore, the set of shortest distance routes for all walks using the whole of the walk network was also generated, and a comparison made with the routes that were taken.

About 30-35 walks could not be reconstructed by this shortest distance strategy, so these had to be reconstructed manually using both the database and the questionnaires. Figure 2.1 shows an example of this where the walk actually undertaken is in bold. This walk started at the right-hand end, which is near home, and ended at the left-hand point, near a child's school. Clearly this walk is neither the shortest route available between the given origin and destination, nor the shortest route using only the roads mentioned by the respondent. No reason was apparent for the circuitous nature of this walk. A map of Northampton is shown in Figure 2.2; all line segments used in the 820 walks are shown in bold.

The strategies considered to synthesise a set of possible routes from which walkers choose is discussed in section 3.

### 2.3. Paper Maps

Two paper map sources were used to aid the tasks detailed above. These were the Northampton Streetfinder Atlas (1993) and the Street AZ Plan Northampton (1995). Their main use was to identify the area in Northampton where each road mentioned in the walk diaries is located and its extent, and to identify some further footpaths.

## 3. CHOICE SET GENERATION STRATEGIES

We considered two distinct strategies to generate a reasonable set of routes that the pedestrians could have used for each of their walks, only some of which were actually chosen. These strategies are discussed in subsections 3.1 and 3.2 below.

### 3.1. *k*-shortest routes

One method of generating choice sets is to identify the shortest, second shortest, third shortest routes and so on, until a certain number (often denoted by *k*) of routes has been found. This is the *k-shortest route* strategy, and generates a choice set in which each pair of paths differs by at least one line segment.

By identifying all routes in order of increasing length, this method will eventually find the one that was used. The *k*-shortest algorithm which is based on the double-sweep method and the FORTRAN code on which our implementation of the algorithm is based is described by Phillips and Garcia-Diaz (1981). The algorithm identifies all paths up to the *k*-shortest including those with loops, in other words routes in which at least one node is reached more than once. We do not desire to have a route choice set containing such routes. They can easily be excluded in general but this results in fewer than k routes without loops being generated. The FORTRAN code uses integers for the lengths and so there is the possibility of more than one route with the same integer length. This motivates the use of *k* as the number of *distinct* route lengths and “route” as the number of routes. We chose a value of 50 as the maximum number of routes to generate.

The *k*-shortest strategy was applied for a *k* value of 20 to each of the 820 origin-destination pairs comprising the walk data. Table 3.1 shows the number of walks whose length is no greater than that of the route with the number indicated. It is clear from this table that 616 out of 820 walks were on the shortest route available. We can also see that for 24 of the walks, 50 routes were generated but even the longest of these was shorter than the route actually taken.

Table 3.1: table of number of walks for each *k*-shortest route number.

| route number | number of walks with length recovered | number of walks with length not recovered |
|---|---|---|
| 1 | 616 | 204 |
| 2 | 667 | 139 |
| 3 | 684 | 104 |
| 4 | 686 | 86 |
| 5 | 686 | 74 |
| 6 | 693 | 63 |
| 7 | 694 | 58 |
| 8 | 694 | 57 |
| 9 | 694 | 55 |
| 10 | 694 | 55 |
| 11 | 694 | 53 |
| 12 | 694 | 51 |
| 13-14 | 694 | 48 |
| 15 | 694 | 47 |
| 16-18 | 694 | 45 |
| 19-30 | 694 | 42 |
| 31 | 694 | 34 |
| 32 | 694 | 31 |
| 33-36 | 694 | 29 |
| 37-40 | 696 | 27 |
| 41-50 | 696 | 24 |

### 3.2. κ-dissimilar routes

Another method of generating choice sets is to identify in turn the paths that differ from the shortest in that they do not contain a certain number of line segments of it and are the shortest such paths. This is known as the *κ-dissimilar path* strategy, where κ denotes the number of line segments of the shortest path that were excluded. This method can be implemented by penalising the use of each line segment in the shortest path by the addition of the same artificial cost, denoted by λ, to each such link: the value of this cost is related to (and increases with) κ and can be determined experimentally (Scott, Pabon-Jimenez and Bernstein, 1997).

Two points about this strategy are worth noting. First, a κ-dissimilar and a κ+1-dissimilar route, for instance, can be identical. This will occur whenever the shortest route between the origin and destination without using κ links of the shortest route does not use κ+1 links of the shortest route. The second point is that a 0-similar route has no links in common with the shortest route and is the longest route that can be generated by this method; the value of λ required to achieve this depends on the origin-destination pair and the network.

This method will generate certain routes in increasing order of length. The maximum number of routes that it can generate is equal to the number of line segments in the shortest one. The length of the 0-similar route might or might not exceed that of the one that was taken, and even if it does, the route taken might not be identified by this method. The principal advantage of this method over the *k*-shortest one is that when the same number of routes are generated, those generated by the κ-dissimilar method will generally have a greater spread of cost.

Table 3.2 below shows the number of walks whose lengths are at most equal to, are less than at 0-similarity, and are less than at some non-zero similarity, the κ-dissimilar route lengths in columns 2, 3 and 4, respectively, for each λ value as given in column 1. It can be seen that 26 walks reported have a similarity with the shortest path of zero and exceed the length of paths generated for λ values of 150, 200, 250, 300 and 700 (metres per line segment). There are 792 walks whose lengths are less than that generated by a λ of 700 (metres). For two identical cases, 0-similarity is possible but is not achieved by a value of λ as high as 900 (metres). Figure 3 shows these walks, the shortest route and the 1-similar route which was attained for a λ value of 26 (metres). The 0-similar routes for these two walks *could* be reached with a sufficiently high value of λ but would have lengths exceeding that of the walks. The route taken in this case has 1 line segment in common with the shortest route.

Table 3.2: table of number of walks for each λ of the κ-dissimilar strategy.

| λ (metres per line segment) | walks with length recovered at 0-similarity | walks with length not recovered: achieved 0-similarity | walks with length not recovered: 0- similarity not achieved |
|---|---|---|---|
| 50 | 752 | 19 | 49 |
| 100 | 762 | 24 | 34 |
| 150 | 769 | 26 | 25 |
| 200 | 780 | 26 | 14 |
| 250 | 780 | 26 | 14 |
| 300 | 780 | 26 | 14 |
| 700 | 792 | 26 | 2 |

Other possible strategies for generation of choice sets include that of Antonisse, Daly and Ben-Akiva (1989) who proposed the calculation of least-cost paths according to each of a range of generalised cost criteria such as time, distance, road crossings and linear combinations of them.

## 4. COMPARISON OF WALKS AND SHORTEST ROUTES

This section briefly compares the length of the walks and the corresponding shortest routes. As mentioned in section 3, 616 of the 820 sample walks took the shortest route. We now look at the other 204 walks and consider the excess length over and above that of the length of the shortest available route. Figure 4 plots the excess length of each of these 204 walks against the shortest route. As can be seen, the excess length is fairly well distributed. The least-squares straight line fit is rather poor, with coefficients 94 (metres) and 0.059 which have standard errors 16 (metres) and 0.015, respectively; the $r^2$ value is only 0.068. The mean of these strictly positive excess lengths is 159 metres and the standard deviation 196 metres, further emphasising the scattered nature of the plot. This shows that about three quarters of the walks were on the shortest available routes and that for the remainder, the mean excess length is highly variable between walks and is only weakly correlated with the shortest length.

## 5. CONCLUSION

This paper has investigated the route choice of pedestrians by using sample data collected in Northampton. This entailed the identification of the actual routes of the walks undertaken using a Geographical Information System. Then route choice sets were generated by implementing the *k*-shortest and κ-dissimilar strategies. It was seen that about three quarters of the walks in the sample were on the shortest available route and that, of the other quarter, the excess length was highly variable and fairly evenly distributed over the shortest route length, increasing weakly with it. At the time of writing, it is intended that discrete choice modelling techniques will be applied to investigate the route choice process of pedestrians and in particular to inform on the characteristics that influence them in this.

## ACKNOWLEDGEMENTS

The authors wish to thank Kay Axhausen of Innsbruck University and Heather Ward of University College London for their contributions to this work. We would also like to thank Ordnance Survey for the free use of their Northampton Landline digital data, without which this work would not have been possible. This work was funded by the Engineering and Physical Sciences Research Council. The pedestrian survey data was provided by the AA Foundation for Road Safety Research.

## REFERENCES

Antonisse, RW, Daly, AJ and Ben-Akiva, M (1989) Highway assignment methods based on behavioural models of car drivers' route choice, *Transportation Research Record*, **1220**, 1-11.

*AZ Street Plan Northampton* (1995) Geographers A-Z Map Co Ltd, Sevenoaks, UK.

Environmental Systems Research Institute (1992b) *ArcInfo: Network Analysis*, ESRI Inc: Redlands, California, USA.

Environmental Systems Research Institute (1992a) *ArcInfo: Understanding GIS the ARCInfo Method*, ESRI Inc: Redlands, California, USA.

*Northampton Streetfinder Atlas* (1993) Bartholomew of Harper Collins Publishers, Hong Kong, China.

Phillips, DT and Garcia-Diaz, A (1981) *Fundamentals of Network Analysis*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, USA.

Scott, K, Pabon-Jimenez, G and Bernstein, D (1997 preprint) Finding alternatives to the best path. *Transportation Research Board 76$^{th}$ Annual Meeting*, Washington DC, USA.

Ward, H, Cave, J, Morrison, A, Allsop, RE, Evans, A, Kuiper, C and Willumsen, LG (1994) *Pedestrian activity and accident risk.* AA Foundation for Road Safety Research, Basingstoke, UK.

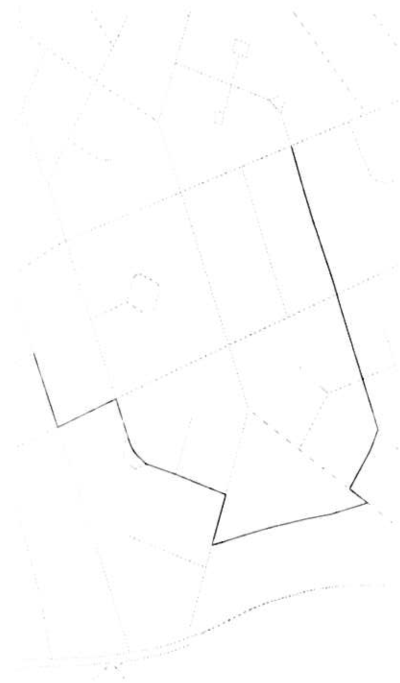

Figure 2.1: a sample walk that required manual reconstruction.

Figure 2.2: Northampton walk network.

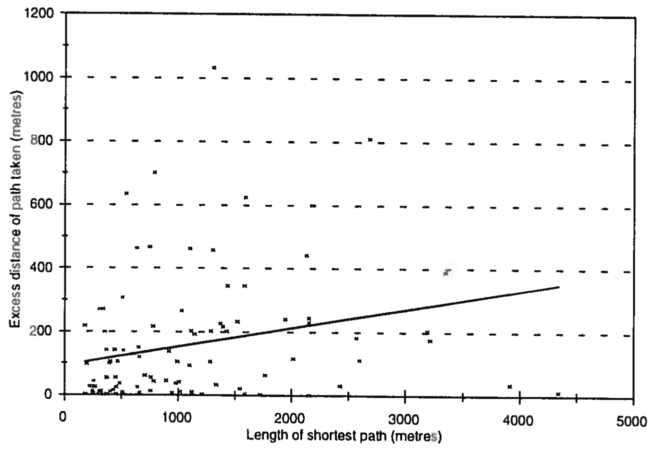


Figure 4: distribution of excess path lengths greater than 0.

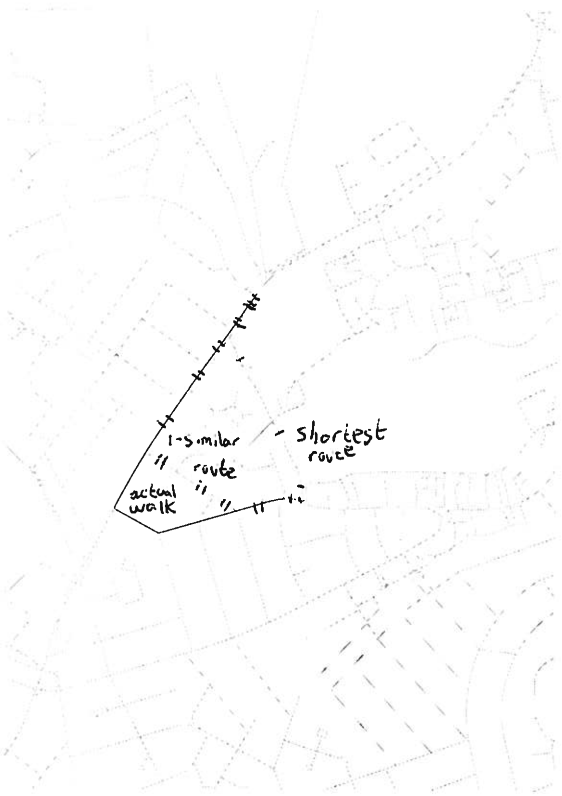


Figure 3: A circuitous walk undertaken, the shortest available route and the shortest 1-similar route.