OFF THE ROAD: FROM DATAPOINTS TO THE NETWORK IN GIS-BASED NETWORK ANALYSIS

J R RITSEMA VAN ECK and T DE JONG

Urban Research Centre Utrecht, Faculty of Geographical Sciences,
P O Box 80115, 3504 TC Utrecht, The Netherlands

1 INTRODUCTION

In one of the first geographical applications of the analysis of networks, Garrison defined the Interstate Highway System as a graph, connecting the metropolitan areas of continental USA. In his paper (Garrison 1960, pg. 243) he noted a problem with respect to the definition of the nodes in this graph: the metropolitan areas must be nodes, because they are the objects that are connected by the segments (highways); the intersections must also be nodes, because otherwise the topology of the graph would be incorrect. His definition of nodes was partly topological, partly based on urban size criteria, but in the graph-theoretical analysis, both types of nodes had to be treated in the same way. On the other hand, 23 out of the 166 Standard Metropolitan Areas were not located on the IHS, and therefore had to be excluded from the analysis altogether.

This problem, the fact that network nodes and geographically interesting locations are not the same, is relevant to many applications of network analysis in GIS.

GIS in general has done a wonderful job liberating us from restrictions that geographical computer models used to impose on the spatial representation of our data. We can define the objects we model to be points, lines or areas of any shape and flavour, scatter them over the map as we encounter them in the field, and GIS allows us to analyse their spatial relations using buffers, overlays, point-in-polygon-tests et cetera.

But if we want to use network analysis, we have to be much more careful. The typical network analysis routines, based as they are on algorithms from graph-theory, calculate distances between nodes in the network, and nodes only. The geographically interesting locations must be nodes. If they are not, new nodes, and in many cases new segments, must be added to the network file to avoid that locations would be excluded from the analysis. This introduces an additional step in the data preparation for network analysis. Furthermore, it creates the necessity to maintain a separate copy of the network file for each combination of origins and destinations that you might want to analyse: applying the same network for analysing many different sets of geographical locations would successively lead to many added segments, most of them no longer necessary and violating the integrity of the network.

We therefore want to argue for a more flexible approach to the relation between data points and transport networks. The basis of this approach is the selection of a suitable connection point for each data point to the network. In the next section, we will illustrate this approach with some practical applications.

2 SOME EXAMPLES FROM PRACTICAL APPLICATIONS

Most GIS-based network analysis routines operate in the same way (Lupien e.a. 1987). The user can specify an origin and (if applicable) a destination, by pointing with a cursor, entering co-ordinates or indicating a selected point entity from the network or another file. The nearest network node is selected and supposed to be the correct location for the data point. The distance from the data point to the network node is ignored, i.e. the data point is effectively moved to the node.

In many practical applications however, this model appears to be incorrect. We will use some examples to show that there are several other models for the connection between data points and digital transport networks. An appropriate model must be selected for each new research project.

In a study of commuting, involving 2000 employees of Utrecht University and University Hospital, network distances from home to work had to be calculated (reported in Ritsema van Eck 1993, pp. 122-140). On the basis of the full postcodes of the employees' addresses, their homes were located with a positional error of less than 200 meters. Using a fairly complete network data set it turned out that all homes were located near the network (within 200 meters), but not necessarily near a node. We decided to connect each data point to the nearest point on a network segment, because the locations of the employees' homes and the locations where they park their cars are not restricted to nodes in the network. As a consequence, the network analysis routine had to be able to calculate the distance between any two points on the network, not just between nodes. We ignored the distance from the data points to the network, because this distance was less than the positional error of the data points.

A second example is described by De Jong e.a. (1991). In this study a planning support system was developed to optimise the spatial distribution of service stations for trucks. The location of service centres was evaluated using the travel-time to registered customers. Registration data were available at the level of postcode areas. As trucks should travel preferably on major roads a relatively sparse road network could be used to save computing time. This meant that not every postcode area contained a node, but there were at least some segments near each area. Travel time from the centroid to the nearest major road was in some cases a significant part of total travel time to the service station; therefore the distance from the data points to the network could not be ignored.

We multiplied the airline distance by a crow fly conversion coefficient (Bonsall 1975, pg. 10) of 1.25 to estimate the road distance and calculated travel time assuming an average speed of 30 kmh and added these travel times to the total trip time. As in the preceding example, there was no reason to restrict the connection points to nodes only, so we connected the data points to the nearest point on any segment. One interesting problem came up in this context: although it is reasonable to assume that a car can enter an ordinary road at any location along that road, for motorways this is obviously not the case. Therefore the network module was adapted in such a way, that the data points could not be connected to a point anywhere on a motorway segment.

The last example is a pilot study for the Dutch Ministry of Transport, Public Works and Water Management (Ritsema van Eck & De Jong 1994). We built a GIS-based system for analysis of a public transportation system. Transport performance was defined by frequency and shortest travel time, including walking time to or from a bus stop or railway station, between centroids of postcode areas.

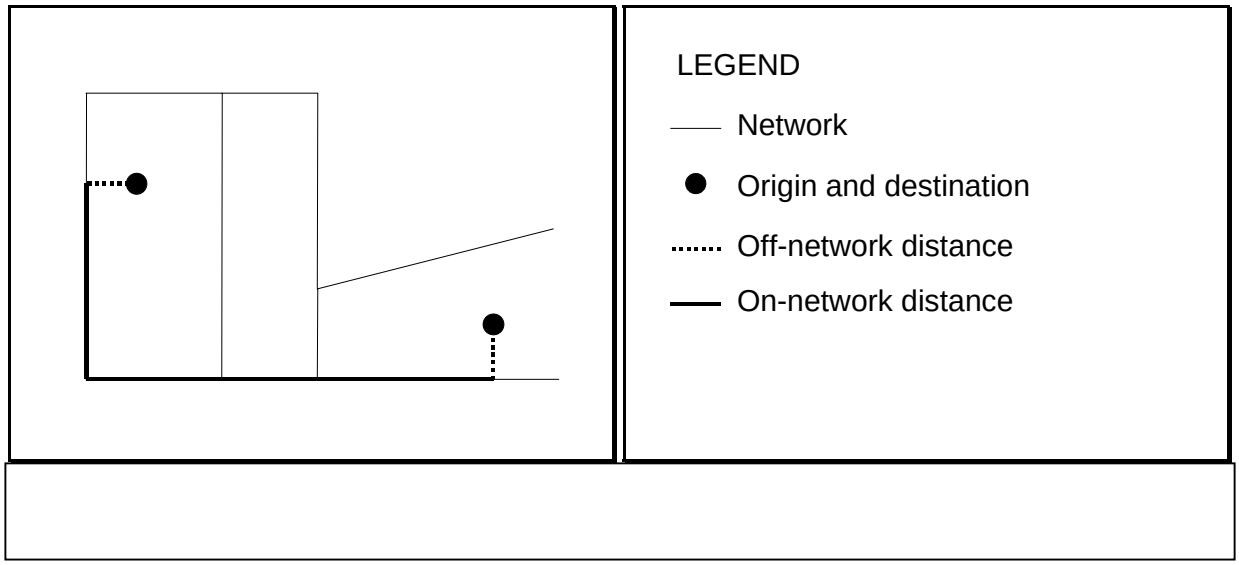
In this case, the network was defined as an overlay of the complete rail networks for railroad and tramway, and the routes of the local and regional bus services. The nodes were defined to be the bus/tramway stops and railway stations.

In this case it was obvious that the data points (postcode area centroids) should be connected to nodes (bus stops). We decided to estimate the walking time between centroids and bus stops, assuming an average speed of 5 kms/h and a path length of 1.25 times the distance "as the crow flies".

In this project, a new problem came up: there are a number of bus stops in any postcode area, and since most bus stops are serviced by just one or two bus lines, people walk to different bus stops depending on their destination. In the GIS analysis, this must be reflected in the selection of the node: for each pair of data points select those nodes in the respective areas that result in the shortest total travel time.

In these examples, we have seen that there are a number of different aspects that are of importance for the best way to connect the data points to a network. In the following section we will discuss these aspects one by one.

3 THE CONNECTION PROCEDURE: A SYSTEMATIC OVERVIEW



The distance (travel time or any impedance) between an origin and destination can be thought of as consisting of three parts (see figure 1): first an "off-network" distance from the origin location to a connection point on the network, second a "on-network" distance over the network from the origin's connection point to the destination's connection point and finally an "off-network" distance from the second connection point to the destination location.

The "off-network" distances and the selection of the connection points are the subject of this section.

First of all we should decide whether the off-network distances should be taken into account at all, and if so, how. Then we should decide what type of network locations can be used to connect to the network: nodes only, or any point on a segment. A further question is, whether for a given data point, there is only one possible connection point, or a number of them. In the last part of this section, we will discuss some other aspects that may be important.

3.1 The off-network distance

The off-network distance is the distance from a location to the point (node or otherwise) on the network where the network route begins or ends (see also Donnay 1995, p. 492). This distance can be ignored if the network is reasonably complete and origin and destination locations are practically located on the network. In that case, the off-network distances are the result of inaccuracies (origins and destinations located at some distance from the network although in reality, they are quite close to it) or scale effects (the distance from the centre of a house to the heart line of the road may be 25 or 30 metres, although the house is located directly on the street).

In other cases, the off-network distance should not be ignored. Two different categories can be distinguished.

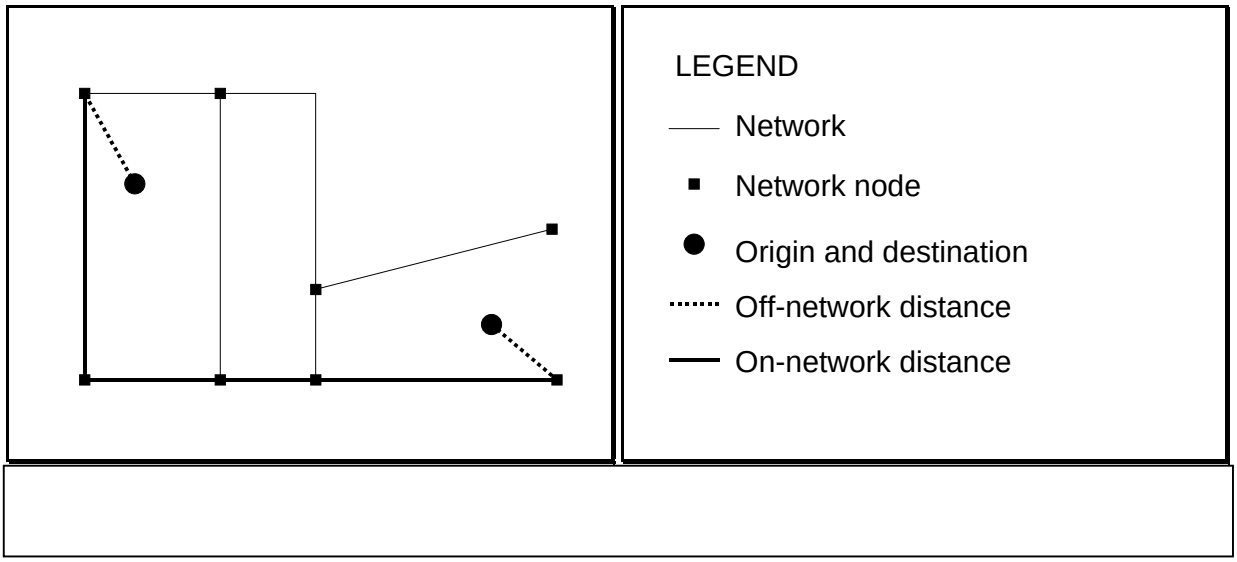
In the first category, the network database is reasonably complete, but terrain conditions make off-the-road movement a conceivable option. We could assume straight-line movement from the origin location to the nearest network point if movement is not constrained by the terrain. In case movement is constrained by the terrain, we could either use some model for movement in this terrain or use a sinuosity-factor, the value of which may depend on the topography.

In the second category, the network database is not complete; less important network segments are missing for reasons of data availability, memory limitations or computing time. Also, if the analysis concerns the future situation in a developing area, the location of new local roads may be yet unknown. In such cases, the off-network distance should be used to model the distance over the missing part of the network. Usually the airline distance from origin to network is shorter than the road distance that is modelled by it. The road distance can be estimated by multiplying the airline distance with a crow fly conversion coefficient, reflecting the curvature of the road (Bonsall 1975, pg. 10). The crow fly conversion coefficient varies, partly depending on the scale and topography of the area; research by different authors suggests values from 1.20 to 1.50 (Ritsema van Eck 1993, pp. 143-152).

If the result of the network analysis is not a simple distance, but an impedance measure like travel time or transport costs, the off-network distance must be converted to the appropriate units by dividing it by a reasonable speed, cost per meter or whatever ratio applies. For instance in the evaluation of truck service centres, we assumed that speed on the missing network segments would be relatively slow: 30 kms/hour. Combined with a crow fly conversion coefficient of 1.25 this results in a travel time of 60 (minutes per hour) / 30 * 1.25 = 2.5 minutes for each kilometre of straight-line-distance from the network.

3.2 Connect to nodes or to any network point?

The connection point represents the point of entry into the network as represented by the network file. If the off-network distance is ignored, the connection point effectively represents the beginning or end of the complete route. If the off-network distance is taken into account because real off-network movement is possible, the connection point represents the location where the network is entered. If the off-network distance is taken into account because the network file is incomplete, the connection point represents the location where missing segments are connected to the network as represented in the file.

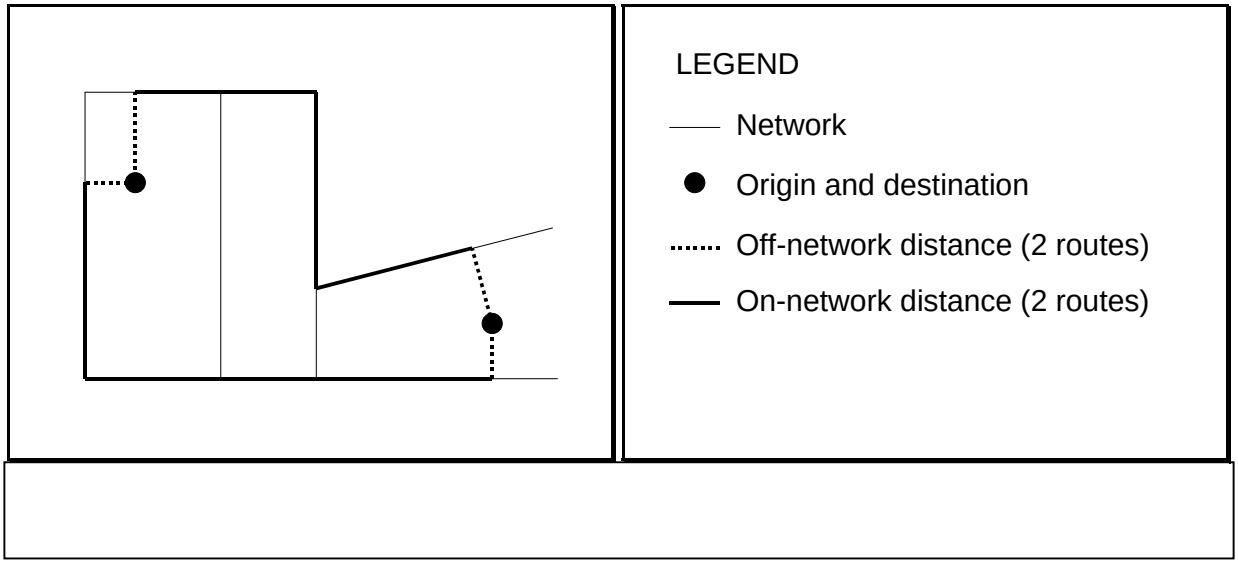


Evidently, if all the points of entry as defined above are represented in the network file as nodes, data points should be connected to nodes only (see figure 2). For example in our public transport study, we made all bus stops nodes by definition. The question is, what to do if this is not the case.

It is of course possible to manually add a node for each point of entry, but this is only practically feasible if the number of entry points is not too large. Alternatively, we could use a "densify"-type of operation to generate a large number of nodes, some of which would be near the entry points. This would however have a huge impact on the computing time for the network analysis.

In most cases it appears to be unnecessarily restrictive to connect to network nodes only. If there are a large number of entry points to the network, many of which do not coincide with a network node, the network routine should connect to any network point, that is: to the nearest point on a network segment. As was mentioned in the example of the commuting study, this means that the network routine must be able to calculate distances between any pair of points on the network, not just nodes. This can be achieved by making use of dynamic segmentation or a similar technique (Uiterwijk & Holsmüller 1991).

3.3 Only one connection point or more?



Once we've admitted the possibility of transport outside the digitised network model, the question must be asked: which other possibilities are there? Is it reasonable to assume, that there is one and only one transport link from each data point to the network, reaching that network at the nearest location? Or are there links in the opposite direction also, to another network segment located further away? Figure 3 illustrates a situation with two links from each data point.

In some cases, it is evident that we have only one link between a data point and the network, for instance each house has only one path from the garage to the road. Usually, in such cases, the data points are much closer to one segment than to any other segment in the network file.

In many cases however, we must assume there are a number of links between each data point and the network. In our evaluation of truck service centres, we connected each postcode area centroid to the nearest point on the main road network. Admittedly, it is possible to enter or leave a postcode area from different directions, so it would have been more accurate to connect each centroid to a number of nearby segments.

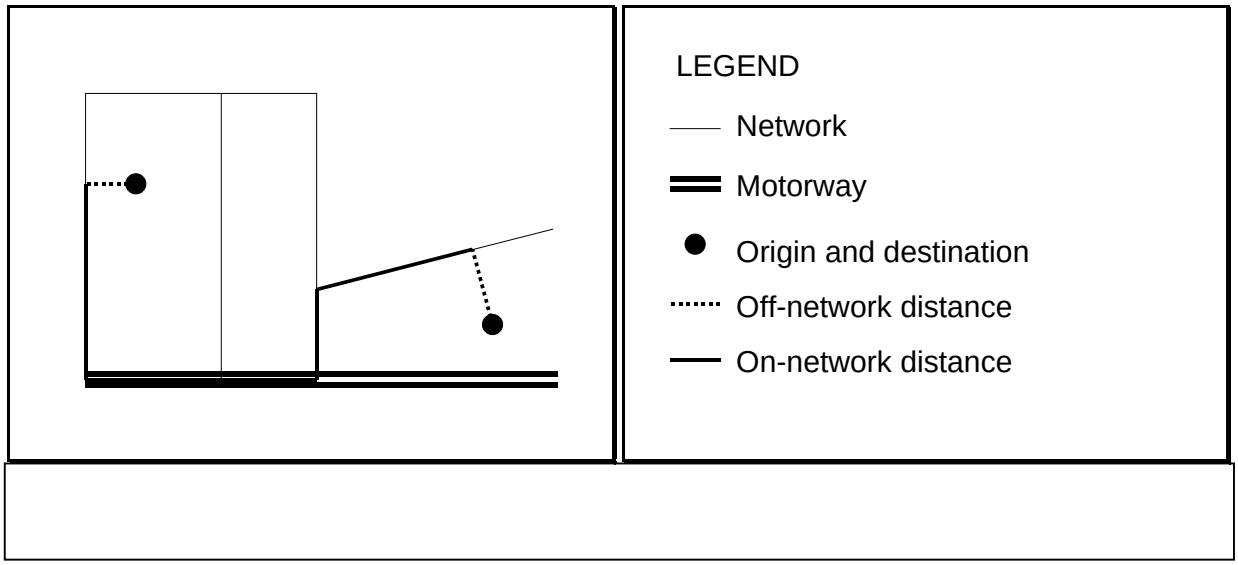
This is exactly what we did in our public transport pilot study. In the case of public transport, it is especially important to define a number of connection points for each data point, because bus stops are typically visited by only a small number of bus lines, so people from the same area walk to different bus stops depending on their destination. We decided that all bus stops within a postcode area could be used as connection point for that area; the network algorithm automatically selects the one that results in the shortest travel-time to the destination. Note, that this does not always result in the selection of the bus stop with the shortest travel time to the destination bus stop: the walking time from the centroid to the bus stop must also be taken into consideration. This favours bus stops near the centroid of the area.

Of course, the decision to allow all bus stops within the area is a little bit arbitrary. An alternative possibility was to allow all bus stops within a 1-km zone around the area's centroid, or all bus stops that satisfy some other definition of "nearness". The main point here, is that we allowed a number of bus stops and let the network algorithm select the best one in each case.

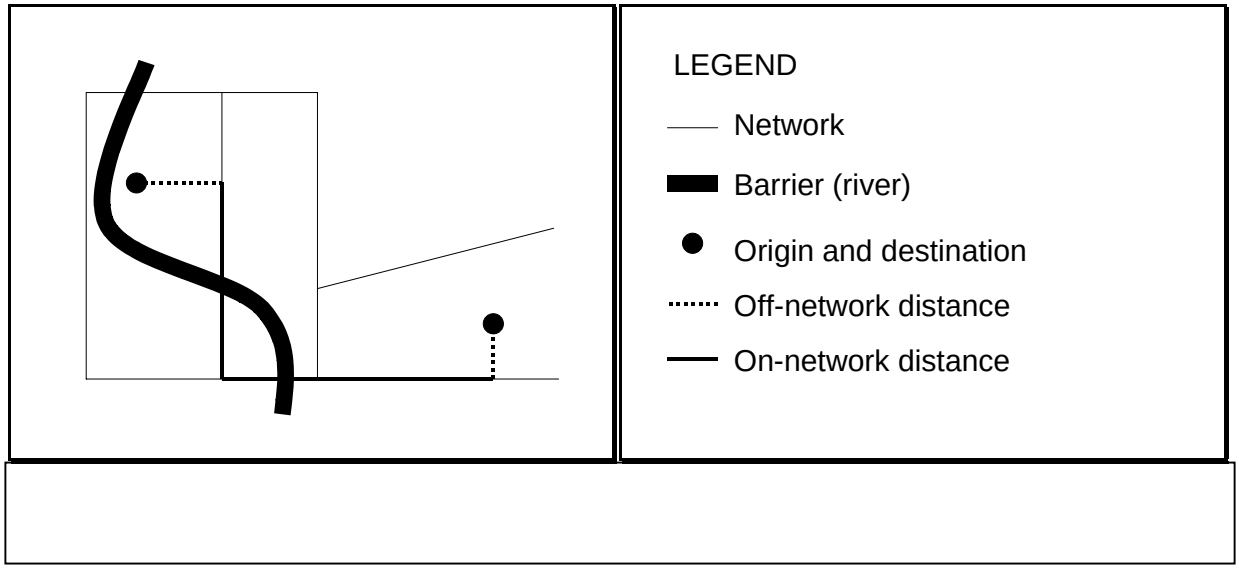
It should be mentioned, that this is somewhat similar to an approach often taken in traffic models, where each zone centroid is connected to the network by a number of "feed links" (Blunden & Black 1984, pp. 179-180, 199-209). In current software packages like TransCad (Caliper 1998) these feed links can be generated automatically. A distance range can be set, the feed links are added to the network and connection points become nodes. As the feed link is now a full part of the network, a corresponding impedance value can be set in the corresponding attribute table.

3.4 Other considerations

Depending on the model and the application, other criteria can be important in the selection of connection points.

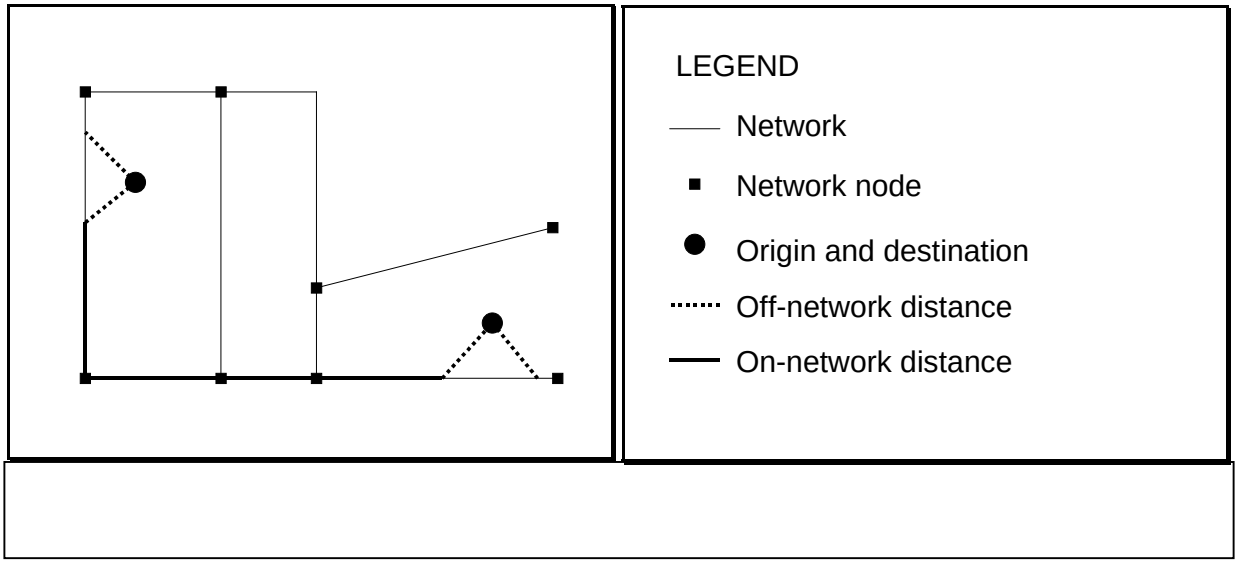


For instance in the case of a road network, and connecting to any point on the segment, it is probably not a good idea to connect to a location somewhere on a segment of a motorway (see figure 4). It is not possible to solve this problem by simply selecting a transport network without the motorways. The motorways should be part of the network during the calculation of the “shortest” route; they just should not be allowed as connection points. The same type of problem occurs when connecting to nodes: a node between a number of motorways (not connected to any other road) is not a plausible start- or endpoint for any trip. Also in public transport, where we want to consider only nodes as connection points, but not all topological nodes in the network represent bus stops or railway stations.

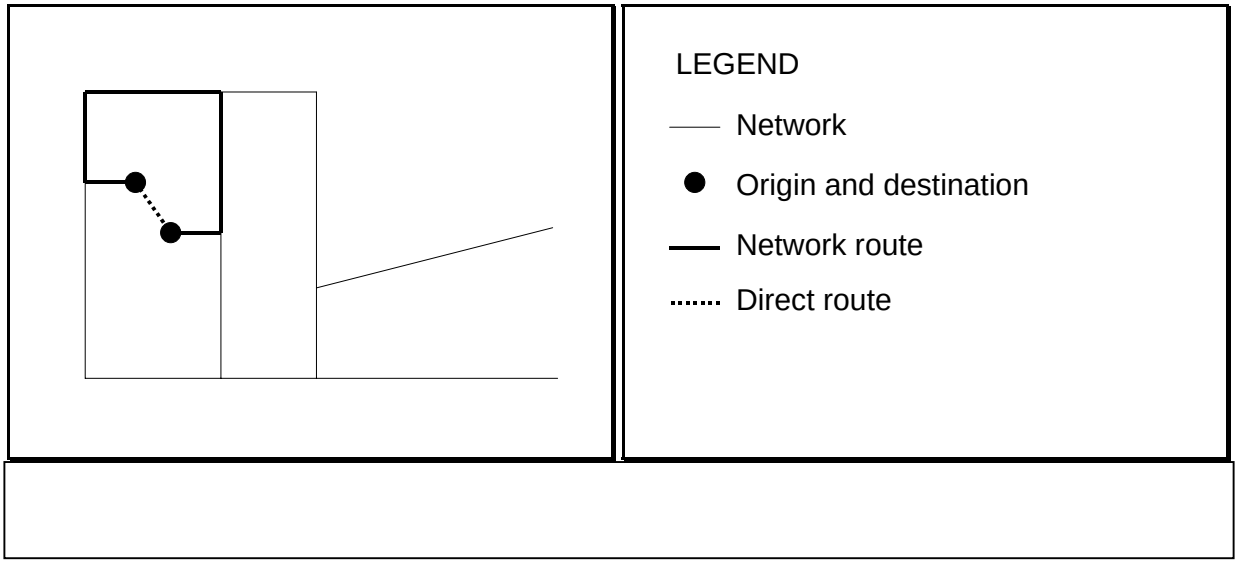


A second case in point are barriers between a data point and the network (see figure 5). A farm or a village may be at one side of a river, and the nearest main road at the other side of the river, without a bridge to get there. In such a case, we can expect our network routine to select the main road as the connection point. If we could make it select the nearest road on the same side of the river, of course this would be much preferable. It could for instance ask the user for the name of a map layer that contains all barriers.

It should be noted that the connection procedure in our public transportation application did essentially this: it selected all bus stops at any distance from the post zone centroid, provided they were on the same side of the "barrier" formed by the borders of the post code area. In the case of automatically generated feed links it is of course possible to select and remove border crossing feed links afterwards.



Thirdly we should consider the need to create more than one connection point per segment (see figure 6). In case a connection point is used the shortest path will always use only a part of the segment; either the part from start node to connection point or the part from end node to connection point. Especially when a road segment is curved the optimum connection point can differ dramatically between these two cases.



Finally we should consider if we need to use the network at all. It two data point are situated relatively close, it the optimal route might be a simple and straight “off road” connection between the two points (see figure 7). It can be argued that the direct route should at least be used, in case this connection is shorter than the two “off-road” parts of the normal route taken together.

4 THE RELATIVE SCALE OF THE NETWORK

With all the different procedures we have described to connect data points to networks, the question may arise: "does it make any difference"? The answer depends on the scale of the network, or rather on the difference of scale between distribution of the data points and the level of detail of the network. To estimate this difference in scale one can convert the set of data points to Thiessen polygons and compare the average size of the resulting polygons to the average size of the circuits in the network. In principle three cases can be distinguished:

1. The level of detail of the network is much finer than that of the data points. In this case the off-network distance makes up only minor part of the shortest path between any two data points and therefore the error made by ignoring it is usually acceptable. On the down side it must be noted that shortest path calculation will be relatively slow as the network contains a great many segments that will never to be part of a shortest path between two data points.

2. The level of detail of the network is in the same order as that of the data points. In theory, all data points can be close enough to the network to make it acceptable to ignore the off-network distance. However, more often than not the network will tend to bisect the distances between adjacent data points (for instance if the data points are centroids of administrative areas, and major roads form the border of these areas). In this case the off-network distance can make up a considerable part of the total distance between two data points especially on the shorter distances.

3. The level of detail of the network is much coarser than that of the data points. In this case (which may occur quite legitimately when analysing future scenarios) the off-network distance usually makes up a considerable part of the short and intermediate distances. Special consideration should be given to the situation where both origin and destination location are located within the same network circuit; in this case none of the connection methods is likely to come up with a reasonable solution. It would be better to use airline distance in combination with a crow fly conversion coefficient.

From the above discussion, we conclude that the selection of a connection procedure is largely irrelevant in case the detail of the network exceeds that of the distribution of the data points by far. In all other cases, the connection procedure may have a considerable influence on the resulting network distances and it is important to select the correct one for the application at hand.

Of course, this presupposes that tools are available to implement this procedure. This is subject of the next section.

5 THE NEED FOR FLEXIBLY SNAPPING NETWORK MODULES IN GIS

In section 3, we discussed a number of possibilities for the connection between data points and network in real-world applications of network analysis. Most of the functionalities that were described there, are not present in the typical network analysis module in an ordinary off-the-shelve GIS program.

For the applications, described in section 2, we programmed our own specific network analysis routines, that had exactly the functionality for connecting data points to the network that was required in each particular application.

In the end of course, it is not efficient, if at all feasible, to program a new network analysis module for each application. A more general type of network analysis module seems to be called for, with options and parameters to control the way in which data points are connected to the network. Five options/parameters follow from the discussion in the preceding section:

a A multiplication factor for the off-network distance; this can be used for both a crow fly conversion coefficient and for a speed or another conversion factor from straight-line distance to the appropriate units; it can also be set to 0 to ignore off-network distance.
b A switch to allow connections to nodes only or to anywhere on network segment.
c A switch to select only one connection point (the nearest one), or all connection points within a user-specified distance (using the best one for each individual distance).
d An optional selection of network segments or nodes that are not allowed as connection points to the network.
e An optional geographic file containing line elements that represent barriers that may not be crossed by the connection between data point and network. This can also be used to define a zone within which the connection point for each data point must be located.

It should be noted that in packages that automatically generate feed links, such as TransCAD, the combination of "all network points" for b with "all points within user-specified distance" for c is disallowed, since the number of points anywhere on a network segment is infinite. In our approach, this combination is possible because the selection of the best points on the network is straightforward for each pair of origin-destination data points.

For the implementation of these options and switches, where possible we make use of the basic functionality present in the GIS:

a A routine to calculate the straight-line distance between data point and connection point is present in the basic GIS; multiplying this by some number and adding the result to the network distance is straightforward.
b/c Routines to select the nearest network node, all network nodes within a specified distance or the nearest network segment from a given location are all present in the basic GIS.
In case the connection point is between two nodes, both these nodes must be selected as entry points. For connection points from the origin, the initial distance must be set to the off-network distance plus the length of the part of the segment between connection point and node (instead of the usual 0). For connection points from the destination, the off-network distance plus the segment length between connection point and node must be added to the final distance, and the smallest final distance is the correct one (Ritsema van Eck & De Jong, 1990, pg. 303). For multiple connection points, the same method can be used.
d) A check if the entry point is located on one of the excluded segments is straightforward.
e) A routine to check if the line between data point and connection point crosses a barrier line element, is present in the basic GIS.

It follows that all these options can be implemented without too much difficulty in a network analysis module, based on Dijkstra's algorithm or another tree building algorithm and part of a vector-based GIS. In fact some of these options were implemented in the network analysis module of the commercial GIS package Genamap and are still used in the freeware GIS package Flowmap.

6 Conclusion

In this paper, we have shown that there are many ways to link the data points to the network, and that there is not any method that is superior in all practical applications.

It might be argued, that there is only one "correct method": to make sure that all data points coincide with network nodes. However, since GIS data are often acquired from different sources, each with their own degree of completeness and accuracy, it is not reasonable to expect the network and data point files to fit onto each other in this way. Nor is it a good practise from a maintenance point of view, to keep adding feed links as an integral part of the network, or to maintain a number of different versions of the same network, one for each combination of origin and destination data sets. Therefore, we think that GIS network analysis routines should offer to the user the possibility to specify the way in which data points and network are connected. We proposed some options and switches that would achieve the required flexibility.

References

BLUNDEN, W.R. & J.A. BLACK (1984): The land-use transport system. Urban and regional planning series vol 2. Sydney etc: Pergamon.

BONSAL, P. (1975): Approaches to the prediction of intrazonal interactions. Leeds: Institute for Transportation Studies (Working Paper 65).

CALIPER (1998): TransCAD user’s guide. Windows version 3.0. Newton Massachusetts: Caliper.

DONNAY, J.-P. & P. LEDENT (1995): Modelling of accessibility fields. In: Proceedings of the Joint European Conference and Exhibition on Geographical Information, pp. 489-494. Basel: AKM.

GARRISON, W.L. (1960): Connectivity of the Interstate Highway System. Reprinted in Berry & Marble (1968): Spatial analysis, a reader in Statistical Geography, pp. 239-249. New Jersey: Prentice Hall.

JONG, T. DE, J.R. RITSEMA VAN ECK & F.TOPPEN (1991): GIS as a tool for locating service centers. In: Proceedings of the Second European GIS Conference, pp. 509-517. Utrecht: EGIS.

LUPIEN, A.E., W.H. MORELAND & J. DANGERMOND (1987): Network analysis in Geographic Information Systems. In: Photogrammetric Engineering & Remote Sensing, vol 53, pp. 1417-1421.

RITSEMA VAN ECK, J.R. & T. DE JONG (1990): Adapting datastructures and algorithms for faster transport network computations. In: Proceedings of the 4th international symposium on Spatial Data Handling, pp. 295-304. Columbus: IGU.

RITSEMA VAN ECK, J.R. (1993): Analyse van transportnetwerken in GIS voor sociaal-geografisch onderzoek (Nederlandse Geografische Studies 164). Utrecht: KNAG/Faculteit Ruimtelijke Wetenschappen (with a summary in English).

RITSEMA VAN ECK, J.R. & T. DE JONG (1994): Een prototype OV-netwerk onder GIS, eindrapport. Unpublished report for the Transportation and Traffic Division of Rijkswaterstaat. Rotterdam: Adviesdienst Verkeer en Vervoer.

UITERWIJK, U. & F. HOLSMÜLLER (1991): GIS as an integrator. In: Proceedings of the second European GIS Conference, pp. 1136-1145. Utrecht: EGIS.

OFF THE ROAD: FROM DATAPOINTS TO THE NETWORK IN GIS-BASED NETWORK ANALYSIS

J R RITSEMA VAN ECK and T DE JONG

Urban Research Centre Utrecht, Faculty of Geographical Sciences,
P O Box 80115, 3504 TC Utrecht, The Netherlands

Dr. **Jan Ritsema van Eck** is Assistent Professor at the Faculty of Geographical Sciences of Utrecht University, the Netherlands. His PhD was about the application of GIS-based network analysis in human geography. His current research interests are land-use transport interaction and accessibility effects of changes of land-use and transportation systems. Jan is co-author of the freeware spatial analysis software package “Flowmap”.

Dr. **Tom de Jong** is Assistent Professor at the Faculty of Geographical Sciences of Utrecht University. His PhD was about the application of gravity models within a planning context. He was one of the first GIS pioneers in the Netherlands. His current research interest is the development of accessibility measurements within GIS and PSS environments.Tom is the designer and main author of the freeware spatial analysis software package “Flowmap”.