$$s = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Divide through by the larger of the *x*- or *y*- differences, i.e. let

$$x=(x_1-x_2),\ y=(y_1-y_2),\ v=min\ \{x, y\},\ w=max\ \{x, y\},\ z=v/w\ (w\neq 0),\ z=0\ (w=0)$$

then $0 \leq z \leq 1$, and we can write:

$$s = w\sqrt{z^2 + 1}$$

If we only consider very small distances, *ds*, and without loss of generality we assume $w=x$ then this expression can be written as:

$$ds = dx\sqrt{(dy/dx)^2 + 1}$$

In this formulation it can be seen that the length of a differentiable plane curve, $y(x)$, can be found by differentiating the expression for the curve and integrating both sides:

$$s = \int ds = \int \sqrt{((y'(x)^2 + 1)}dx$$

This formulation is used in the next subsection (and Annex 2) to compute path costs where incremental distance is of the form $ds^*=Fds$ and *F* is a simple non-uniform positive cost function defined across the plane. If the curve, or path, $y(x)$, is either not differentiable everywhere or not a plane curve (e.g. it is a path across a surface such as a physical landscape) this formulation no longer holds and alternative expressions and methods must be sought. This case is discussed in several of the Sections that follow.

A second incremental reformulation of the Cartesian formula using the same model as before, is:

$$ds = dx\sqrt{\tan^2\theta + 1} = \sec\theta dx$$

where $\theta$ is the angle made by the line between the two coordinate points and the *x*-axis. This provides the basis for an alternative method of path length estimation, as described in Section 5.2.7, below.

Another useful reformulation we consider is to expand the expression as a power series[13] in *z* about *z=0*:

$$s = w(1 + z^2/2 - z^4/8 + z^6/16 - 5z^8/128 + \varepsilon)$$

The value of the error term, $\varepsilon$, is $\varepsilon \leq 0.00002$ when $z \leq 0.5$, and remains small (well under 1%) for values of $z \leq 0.9999$. If required, precision can be improved further by adding the term $+7z^{10}/256$ or can be reduced for the sake of increased analytic or computational convenience by removing one or more of the higher terms in *z*. For example, limiting the series to only 3 terms still provides accuracy of better than 3% at *z=0.5*. Furthermore, letting *b=z*z/2* the power series can be written without the need for any explicit power calculations thus providing improved computational accuracy, as:

$$s = w(1 + b(1 - 0.5b(1 - b(1 - 0.625b))))$$

This estimate has a mean accuracy of under 0.05% and a maximum error of 0.27%.

Without loss of generality we can assume *w=x*, multiply through by *x*, to obtain the revised series:

$$s = x + y^2/2x - y^4/8x^3 + y^6/16x^5 - 5y^8/128x^7 + \varepsilon$$

We can now differentiate this function with respect to *x* and *y* to produce two further power series, first for *x* (re-introducing the expression for *z* and dropping the highest order term, we have):

$$\frac{\partial s}{\partial x} = 1 - z^2/2 + 3z^4/8 - 5z^6/16$$

Similarly for *y*:

$$\frac{\partial s}{\partial y} = z(1 - z^2/2 + 3z^4/8 - 5z^6/16)$$

from which we see that, as before:

$$\frac{\partial s}{\partial y} = z\frac{\partial s}{\partial x}$$

Reducing the number of terms in a power series results in loss of precision, although very accurate approximations to circular *arcs* can be achieved with polynomials of order five[14]. These remain analytically awkward to use and other options warrant investigation.

One alternative approach is to seek an optimal linear approximation. A remarkably accurate linear estimate can be derived from recent research into the use of *chamfer* metrics[15] in image processing (see further, Section 5.2.8). The expression:

$$s = 0.36930v + 0.95509w$$

where *v* and *w* are defined as above, estimates *s* to within 2% on average and within 6.35% for all *v*, *w*. Summations which utilise this expression include positive and negative errors of estimate and this tends to reduce errors further - errors in the sum of three or more figures rarely exceeding 3%. Furthermore, minimisation of a sum of distances from a set of points to an unknown or 'query' point using this model simply involves minimisation of the sums of the maximum differences and minimum differences. This can be achieved very rapidly by simple binary chop iteration or similar procedures, where the search space is restricted to the max and min values of the coordinates in the set (i.e. differences in either x or y, whichever is the greater).

Note that the expression

$$s = \left(-1 + \sqrt{2}\right) * v + 1.0 * w$$

provides a less accurate approximation in most cases (it provides a better approximation only along the diagonals and N-S-E-W radii from a sample point). The former expression for $s$ may be re-written as:

$$s = 0.36930\,|\,x_1 - x_2\,| + 0.95509\,|\,y_1 - y_2\,| \qquad \text{if } |\,x_1 - x_2\,| < |\,y_1 - y_2\,|$$

$$s = 0.36930\,|\,y_1 - y_2\,| + 0.95509\,|\,x_1 - x_2\,| \qquad \text{if } |\,x_1 - x_2\,| \geq |\,y_1 - y_2\,|$$

The two inequalities, coupled with evaluating the expression over positive and negative values for $x$ and $y$ (4 quadrants) results in an octagonal locus as an approximation to the true circular form of the Euclidean metric. Closer approximations can be achieved by the introduction of additional inequalities, but the improvement gained may be outweighed by the solution complexity. If distances are selected using polar rather than Cartesian coordinates in the plane it is possible to obtain a very slight improvement to the linear approximation above (this topic is discussed in more detail in the next Chapter, Section 5.2.8, in the context of lattice approximations and distance transforms).

The Euclidean metric (and approximations to it) provides a distance value for the separation of a single pair of points. It is possible for this distance to be computed on the same basis but via an intermediary point, $P(p,q)$, such that if $d(x,y)$ represents the Euclidean distance between the points $x$ and $y$, then a new metric can be defined as:

$$d_P(x,y) = d(x, P) + d(P, y) \;\text{ if } x \neq y, \qquad d_P(x,y) = 0 \;\text{ if } x = y$$

This is an example of a *hybrid* Euclidean metric, since the formulation involves a distance calculation using the Euclidean measures but taken via an intermediate point. It is sometimes described as the Post Office metric, and applies to situations in which all traffic (physical, electronic, data etc) must be routed via a specified location. A variety of other hybrid metrics are discussed later in this Chapter and in Chapter 6, Distance Statistics.

#### 4.2.4.2 $L_p$ or Minkowski inequality metrics and related forms

Type 3 in Table 4-1 is the same as Type 2 if $n=2$ and $p=2$. However, if $0 < p < 1$ then the measure is a semi-metric since the triangle inequality no longer holds[16]. Shreider has discussed this case for points in the plane in some detail and has produced a diagrammatic interpretation of the measure for various values of $p$ (Figure 4-2). In this case we use the form:

$$d_p = \left( | x_1 - x_2 |^p + | y_1 - y_2 |^p \right)^{1/p} \qquad ; \; p \geq 1$$

which is often referred to as $L_p$ distance[17]. A similar approach to that described for Euclidean distance can be used to obtain a power series representation for this metric, expanded about $z=1$, given $p$, although the series is more complex when $p$ is not an integer. In this case the expansion is of

$$s = w\left(\sqrt[p]{z^p + 1}\right)$$

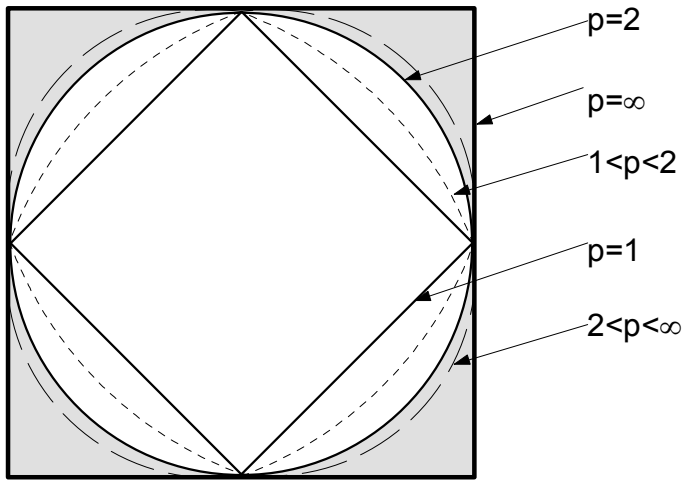
Of the set of all $L_p$ norms, only the case $p=2$ (the Euclidean metric) is invariant with respect to (orthogonal) rotation – all others are invariant with respect to translation and central reflections of the coordinate axes, but not rotation. This latter observation is clarified by examination of Figure 4-2 and observing that:

$$d_p = d_2 \left( | \cos\theta |^p + | \sin\theta |^p \right)^{1/p} \qquad ; \; 0 \leq \theta \leq \pi/2$$

The directional bias of the $L_p$ norm increases as $p$ deviates from 2 and is a maximum for all $p$ at $\theta=\pi/4$ where the bias is proportional to $2^{1/p-1/2}$. One result of this observation is that rotation of the axes will alter the estimated value of $p$ and therefore any use of the $L_p$ norm or metrics based on this norm (e.g. weighted and/or hybrids of this metric) should either use a range of axis rotations from $[0, \pi/4]$ during sampling and/or parameter fitting or state that the results are not independent of coordinate system rotation. Furthermore, it is clear that all values of $p>2$ (including $p = \infty$) are related to $p<2$ by a simple rotation. In geographical literature the case $p=1$ is variously called the Manhattan, Taxicab, Rectilinear or City block metric, $L_1$ – this example is discussed

further below and in Section 8.3. The case $p = \infty$ is called the maximum, supremum or Chebyshev norm, $L_\infty$, and measures the maximum deviation in $x$ or $y$, i.e. $d=max\{|x_1-x_2|, |y_1-y_2|\}$

Figure 4-2 $L_p$ space circles




*redrawn, after Y U Shreider*

The above observation on rotational relationships can be used in problem solving – in some cases it may be possible to solve selected problems by applying the sequence of actions:

(i) rotate the problem coordinate frame
(ii) solve the problem using the metric appropriate to the rotated frame, and then
(iii) rotate the frame back to its original position.

For example, in order to determine the Voronoi diagram for a set of planar points under the $L_\infty$ metric, the set of points may be rotated by an angle of $\pi/4$ and the regions computed using the $L_1$ metric, with the resulting diagram rotated back by $\pi/4$ degrees to produce the final solution[18].

The use of the term *Manhattan metric* in connection with the $L_1$ metric is somewhat misleading, in that it does not correspond to the way in which traffic moves in Manhattan at all – the Manhattan road system is a specific, physical network, with a finite number of paths, almost all of which are directed (i.e. one way). This is not a

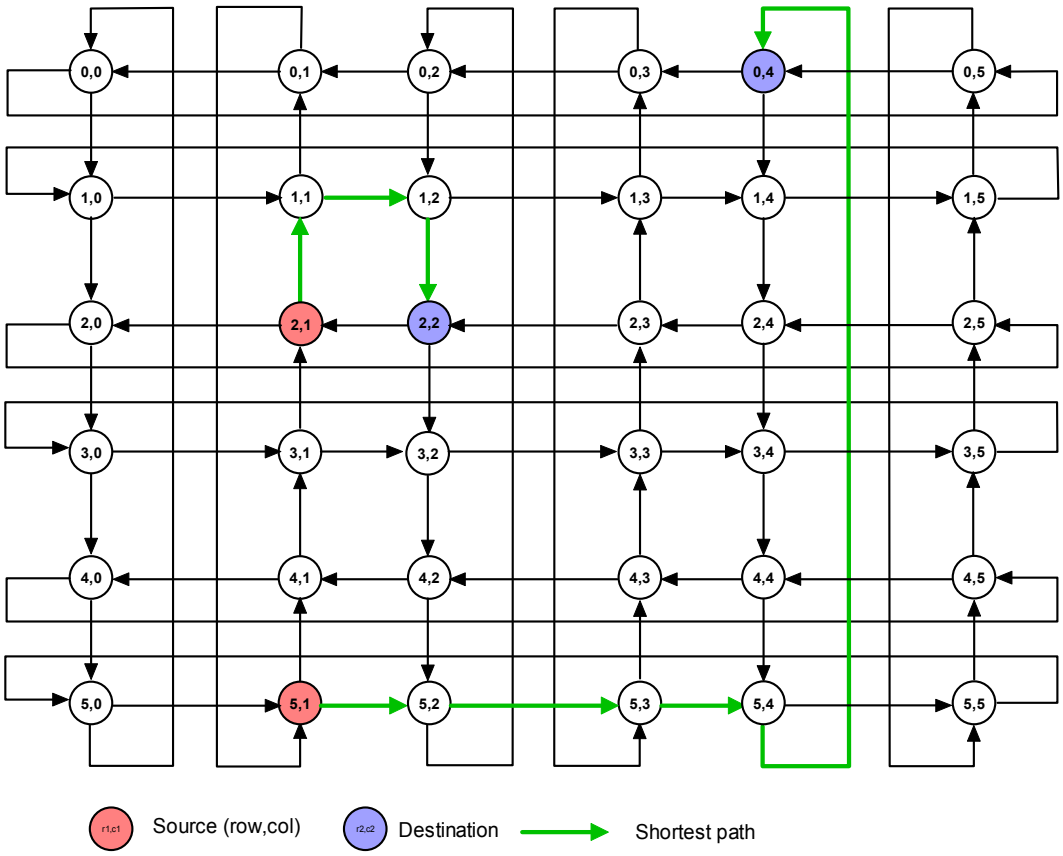
metric space, since symmetry and triangularity requirements are not satisfied. However, the (idealised) finite Manhattan Street Network (MSN) can be viewed as a two-dimensional, unidirectional wire-frame torus network or digraph (an example of which is shown in Figure 4-3). Pairs of east-west and north-south routes are effectively rings (which is, of course, not quite like Manhattan streets at the mesh edges), with shortest routes involving paths that move around the rings. The MSN has been extensively studied and applied in telecommunications network design[19], particularly in packet- or message-structured systems where the number of hops (junctions passed through) is the most important measure of distance, rather than the specific path length. In the case of a 6x6 network, as illustrated, the average shortest path distance (i.e. the sum of all possible shortest paths divided by the number of nodes) is 3.89 hops. An analytic expression for the average number of hops in an *NxM* MSN network is[20] :

$$d_{ave} = (N + M + 4)/4 - 4/MN$$

The *diameter* of the MSN (defined as the maximum distance from a node to another node, and an important parameter in network design) is either *max{N,M}* or *max{N,M}+1* depending on whether *max{N,M}/2* is even or odd.

Labelling of nodes in this example is in integer row/column format – such labelling can be used to provide addressing schemes that facilitate very efficient (i.e. fast, low overhead) deterministic shortest path routing.

Figure 4-3 36-node Manhattan Street Network




*after N F Maxemchuk (1987)*

Key advantages of this kind of regular mesh network in telecommunications include the simplicity of routing, robustness of the structure to link and node failures, containment of localised traffic within distinct sub-regions of the overall network, and ease of extension. Disadvantages include the number of nodes required, the volume of processing per node, and the amount of cabling involved. The structure has been found to be especially effective in the design of very high-speed networks in metropolitan regions and within some local area network (LAN) designs.

It should be noted that the $L_p$ metric for general values of $p>1$ has been used extensively in psychometrics and by a number of geographers in the context of multi-dimensional scaling[21] and has been widely considered in other disciplines such as road traffic studies and operations research[22].

One, commonly used, multi-parameter variant of this metric is:

$$d = a\left(|x_1 - x_2|^p + |y_1 - y_2|^p\right)^{1/s} \qquad \text{where } a, p, s \geq 1$$

This form has been used in modelling road networks with improved accuracy over the simple $L_p$ norm. For example, using the two-parameter model (i.e. with $s=p$) Love and Morris[23] sampled inter-city distances in Wisconsin and across the USA and found estimates of $a=1.11$ and $p=1.69$ in the first case, and $a=1.15$, $p=1.78$ in the second. If $p/s>1$ in the model above then it becomes a semi-metric, and since the assumption of $s=p$ retains most of the explanatory power of the measure (at least in road network modelling) it is preferable to adopt the two-parameter metric model where possible.

Similar research in Sweden has shown that road distances in towns is on average 1.21 times the Euclidean distance. Tobler[24] has suggested that such findings could be used as a simple 'fudge factor', *a* (some authors use the term 'road coefficient' or 'route factor' – see further, Section 6.5) retaining conventional metrics in these cases and adjusting the results by a constant factor. This suggestion retains the desirable objective of rotational invariance with the possibility of utilising a metric with improved explanatory power.

Pursuant to this suggestion we may define a new metric, which we shall call the *Modified Euclidean* metric, as:

$$d = a\left(|x_1 - x_2|^2 + |y_1 - y_2|^2\right)^{s/2} \qquad \text{where } a, s \geq 1$$

This metric provides a formulation with many of the characteristics desired, although not rotational invariance, and warrants examination of its utility in GIS-T and related applications. Alternatively the $L_p$ metric may be used, subject to the precautionary notes made earlier, or a more complex estimation process used, such as vector quantisation[25].

Setting $K=p/s$ in the model of Love and Morris described above, and letting $a=1$, it can be seen that their metric and our Modified Euclidean metric are essentially of the form $L_p^K$, i.e. the $K^{th}$ power of the standard $L_p$ norm. Morris[26] has shown that when this metric is used as the distance measure in distance minimisation models of location (Weber problems), the objective function it is convex if $K \geq 1$ but not generally if $0<K<1$.

Furthermore, if $0<K<1$ the measure is not a full metric since the triangle inequality does not hold.

Yet another variant of the $L_p$ metric has been utilised by Muller[27] in the context of mapping functional distances. He takes the multi-parameter model:

$$d = \left(a \mid x_1 - x_2 \mid^p + b \mid y_1 - y_2 \mid^p\right)^s$$

where *p, a, b* and *s* are parameters to be estimated

This model is metric if $p>1$, $s = 1/p$ and $a=b=1$. Muller found that this model works well for road distances, actual travel times by car and estimated travel times by car, but not for air travel times (all cases based on selected German cities). His estimates for the three road travel cases found $s \approx 1/p$ and $a/b \approx 1.25$ (*1.1 - 1.4*), suggesting that simpler models would yield good results in many cases. Two-dimensional maps produced using these metrics led to more meaningful and 'accurate' (lower stress) results than those obtained by using conventional metric scaling with a Euclidean metric.

The observation that $a/b \neq 1$ (together with the fact that Muller found values of $p\neq2$), suggests that (a) directional (e.g. North-South) bias exists in the datasets (notably in the actual and estimated travel time analyses) and (b) rotation of the reference frame would alter the findings (as noted above). This is confirmed by observing that Muller's metric, and the $L_p$-based metrics in general, are related to the shape known as a *super ellipse*. A super ellipse (in 2-space) has the form:

$$\left|\frac{x}{a}\right|^p + \left|\frac{y}{b}\right|^p = 1$$

where $p\geq1$. Clearly if $a=b=1$ and $p=2$ we have a circle. If $a=b$ and $p\neq2$ we have $L_p$ space circles, and if $a\neq b$, $p\neq2$ we have super-ellipses. The latter look just like the space circles in Figure 4-2 above, but stretched East-West or North-South.

The term "super ellipse" is sometimes reserved for the case $a=b$, $p=5/2$ which was studied and developed by the Danish architect and designer, Piet Hein, in the 1960s.

This formulation has been used in designs ranging from furniture to road traffic islands (super-elliptical roundabouts). The family of curves for general $a$, $b$ and $p$ was first studied by the French mathematician, Gabriel Lamé, in 1818, i.e. long before Minkowski produced the inequality discussed earlier. An extension of the super-ellipse, with close similarity to the metric generalisations discussed above, is known as the *generalised super-ellipse* and is defined by:

$$\left|\frac{x}{a}\right|^p + \left|\frac{y}{b}\right|^q = 1$$

The general (and unsurprising) consensus across many studies is that distance functions that contain more parameters (2, 3 or more) tend to provide better estimates of road distance, but at a cost of increased complexity – frequently the improvements seen are marginal and models with lower numbers of parameters to be estimated are to be preferred. It is also important to note that in some cases (as discussed further below) it is simply not possible to use a single model approach (e.g. radial/ring hybrids are a better reflection of the actual road pattern than norm-like formulations). The principal advantage of such measures is their ability to provide accurate, analytic, functionally relevant distance estimates between any two points in the sample space, whether continuous or discrete space (e.g. lattice). These can then be used in applications such as route planning and traffic forecasting, location-allocation modelling and space-partitioning/zoning problems.

The $L_p$ metric does not determine unique paths between point pairs[28]. For example, with $p=1$ there is an unlimited number of possible (shortest) paths if the street network is sufficiently fine. However, suppose that one is examining an existing complex (but otherwise relatively homogeneous) street network in a city. By measuring a sample of inter-location shortest distances, $L^n$, a set of values (estimates) for $p$ may be obtained, the (weighted) mean of which may then be used in the formula above without the need for calculation of all possible shortest paths[29]. Separate samples could be used for distinct areas – for example, in Edinburgh, each main district of the City could be treated as distinct. This raises the question of which streets to include - all streets, allowing for one-way rules, or just streets marked on the map as ‘main’ routes?

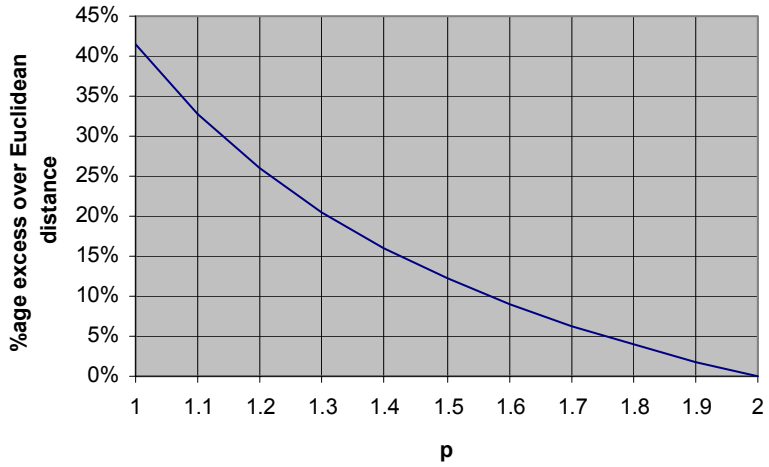
Keeping with the simple, single parameter norm, if $p$ is a parameter estimated for all or part of a network, a summary of interpretations of its value is provided below:

| | |
|---|---|
| $0 < p < 1$ | The $L_p$ formula does not yield a metric because the triangle inequality no longer holds – for example, construction of minimal spanning trees on the basis of progressively connecting nearest neighbours is not valid. Routes in cities may have values of $p<1$ indicating that routes are complex – a value of $p=2/3$ indicates that distances are effectively twice as far as direct line estimates would suggest. The locus of points in this case is concave. |
| $1 < p < 2$ | The distance, $d$, is a metric which is intermediate between the Manhattan grid-like street pattern and the normal Euclidean metric, and is thus applicable to more complex city street networks. |
| $p > 2$ | As $p$ increases greater weight is given to the largest components of distance in the network. Consider the distance from $(0,0)$ to $(1,1-E)$ where $0<E<1$. For $p=1$ $d=2-E$; for $p=2$ and $E<<1$ $d \approx \sqrt{2}\sqrt{(1-E)}$ and as $p$ increases, $d \rightarrow 1$. In all cases as $p \rightarrow \infty$ then $d = max(\lvert x_1-x_2 \rvert, \lvert y_1-y_2 \rvert)$ |

Note that if either the $x$ or $y$ coordinate differences ($dx$ or $dy$) are zero the measure simply provides the absolute difference in the remaining coordinates and is thus independent of the value assigned to $p$. In fact, the *maximum* difference between the Euclidean and $L_p$ metrics occurs when the $x$-increment and $y$-increment are equal, as shown in the graph below (Figure 4-4). If $dx>>dy$ or $dy>>dx$ the percentage excess tends to zero, since this result is equivalent to $dx$ or $dy$ tending to 0.

Increasing $p$ may be used in minimax procedures whereby the maximum horizontal or vertical distance from a vertex to a facility is minimised – such a metric might therefore be appropriate for the selection of strategic locations or satisfying certain network performance criteria.

Figure 4-4 $L_p$ metric versus Euclidean metric for (0,0) to (1,1), *p = 1 to 2*




With *p=1* all shortest paths in the plane are rectilinear 'staircase-like' forms with a minimum of two steps (one in the *y*-direction and one in the *x*-direction).

It is possible to define a metric in which the staircase-like paths are not rectilinear, but set at an angle to each other. In general it is sufficient to assume that one direction is parallel to the *x*-axis and the other is at some angle <90 degrees to this axis. Such paths have lengths that are the sum of the Euclidean distances in each direction and these are known as *gauge distances*[30]. As such they are a generalisation of the rectilinear distance model and have been the subject of detailed study in recent years by specialists in location theory (see further, Section 8.1).

#### *4.2.4.3 Normalised metrics*

Type 4, the Canberra metric, is an example of a normalised metric (it may usefully be compared with the pseudometric measure of similarity discussed earlier). Such measures are used in a wide range of multidimensional data analysis techniques. The Canberra metric normalises the differences between each data pair by their component magnitudes. This removes the effects of extreme values upon the totals and helps ensure that the scales used in each direction or dimension may be readily compared (are *commensurate*).

A variety of other normalised distance measures are used in fields such as multi-dimensional scaling, factor analysis, cluster analysis and indexing problems, in many

cases using the range or standard deviation of each dimension as the normalisation factor. As an example, the normalised Euclidean metric may be written as:

$$d(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left( \sum_{k=1}^{n} \frac{(x_{ik} - x_{jk})^2}{\sigma_k^{\,2}} \right)^{1/2}$$

where $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ represent *n*-tuples or points in *n*-space (2+ dimensions) and are samples from a set of such points with variance in the $k^{th}$ dimension of $\sigma_k^2$. These measures, many of which satisfy the metric criteria stated earlier, are only discussed briefly in this study as their primary function is not directly related to spatial analysis.

#### 4.2.4.4 Functional metrics

Type 5 is perhaps the least recognisable as a metric. Although it has some similarity to the least squares metric described earlier, it is not obviously applicable to geographic problems. The expression does satisfy the requirements for a metric and an example shows its utility in the geographic field. Suppose that *x* represents the distance in kilometres along a road of length *L* upon which an organisation wishes to locate a depot so as to minimise its delivery costs. In addition, suppose that the demand to be served can be represented by a positive function of position along the road, $D(x)$, and the depot location is taken as $x_0$, where $x$ and $x_0 \in [0,L]$. Then the function, $z$, to be minimised is given by:

$$z = \left| \int_0^L f(x)dx \right| = \left| \int_0^L d(x, x_0) D(x)dx \right|$$

i.e. $$z = \int_0^L | \, d(0,x) - d(0,x_0) \, | \, D(x)dx$$

or $$z = \int_0^{x_0} (d(0,x_0) - d(0,x)) D(x)dx \; + \; \int_{x_0}^{L} (d(0,x) - d(0,x_0)) D(x)dx$$

partially differentiating wrt $x_0$:

$$\partial z / \partial x_0 = d'(0, x_0)\left[\int_0^{x_0} D(x)dx - \int_{x_0}^{L} D(x)dx\right]$$

The optimum depot location, $x^*$, is found where the partial differential equates to 0, i.e. where

$$\int_0^{x^*} D(x)dx = \int_{x^*}^{L} D(x)dx$$

So the optimum location is at the median of demand. This minimum point is unique[31] and independent of the particular demand function, $D(x)$, chosen. Since $D(x)$ can be a very general cost function, such as delivery cost, taking into account congestion etc., the result is particularly striking. An obvious extension of the above case, for example, is to examine the result when two or more depots are required, or when demand patterns are uncertain. These two cases are examined briefly below.

By a similar argument to that provided above, the optimum location for two depots ($x_0$ , $x_1$) may be found. In this case each depot occupies the median of the partitioned demand distribution, $D(x) = D_0(x) + D_1(x)$, where the point of partition is $x^*$. To find $x_0$ and $x_1$ a location-allocation algorithm is required[32], whereby $x^*$ is selected (say at the median of $D(x)$ or at $L/2$), then the locations of $x_0$ and $x_1$ computed with respect to the two initial sections of $D(x)$ and then $x^*$ (and $x_0$ and $x_1$) recomputed iteratively until no discernable improvement is found.

Generalised to two dimensions, $D(x)$ becomes $D(x,y)$ and $d(x, x_0)$ becomes $d(\boldsymbol{x},\boldsymbol{x}_0)$ where $\boldsymbol{x}=(x,y)$ and $\boldsymbol{x}_0=(x_0,y_0)$. The task of minimising the total transport effort, $z$, is then known as Weber's problem, generalised to permit metrics other than Euclidean. With more than one depot to be located Weber's problem is known as a generalised unconstrained location-allocation problem.

In order to analyse the question of depot location when the demand function, $D(x,y)$ is probabilistic, two observations must be made. The first concerns construction of two-dimensional demand functions. Several authors[33] have considered generalisations of Weber's problem in which either there are:

(i) a finite set of $n$ demand points, with weights $w_i$, are each considered to have probabilistically defined locations, e.g. location $i$ has mean location $(x_i,y_i)$, weight $w_i$, and is randomly located in a circle of radius $r_i$; or

(ii) a finite set of regions is defined in the plane, each of which is assigned a weight and within each of which the demand is assumed to have a particular distribution, $p_i(x,y)$.

For the purpose of analysis we regard both of these forms of demand function as replaceable by the composite demand function:

$$D(x,y) = \bigcup_i w_i^* p_i(x,y)$$

i.e. the demand function is the union of all the sub-demand functions $w_i{*}p_i(x,y)$; $w_i{*}$ is the (composite) weight associated with the $i^{th}$ point or area in the region, $R_i$, of interest, and $p_i(x,y)$ is the (composite) distribution of demand around point $i$ or in region $R_i$.

The second observation to be made is that minimising the expected distance to a depot located in the sample region $R$ is equivalent to minimising a two-dimensional version of the line-based problem described earlier, viz:

$$\min\ z = \int\int_R d(\boldsymbol{x},\boldsymbol{x}_0)D(x,y)dydx$$

Furthermore, if $d$ is the $L_1$ metric, the two dimensions $x$ and $y$ may be considered independently and the coordinates of the optimum locations are simply the medians of the marginal distributions $D_x(x,y)$ and $D_y(x,y)$, where

$$D_\xi(x,y) = \int_{-\infty}^{\infty} D(x,y)d\xi$$

This conclusion avoids the need for complex optimisation procedures such as those discussed by Wesolowsky[34] and Drezner[35]. To illustrate this result we may take Wesolowsky's example of a 5-point problem in which each point, $(x_i,y_i)$ is specified by a bivariate uniform probability distribution:

$$p_i(x_i, y_i) = \frac{1}{(f_i - g_i)} \cdot \frac{1}{(h_i - g_i)}$$

Table 4-2 shows Wesolowsky's data and the pattern of overlap of probability regions in the *x*-dimension. To compute the composite marginal demand function, $D_x(x,y)$ we simply calculate the demand for each sub-interval, $[s_i, s_{i+1}]$ which is the product of the interval length and the weighted probability functions found in that interval. The total demand is simply the sum of the destination weights, i.e. 8, so the solution point is to be found at the point where the sum of the probable demand = 4 in both *x* and *y* directions. Taking the *x*-direction, the solution required for $x_0$ must therefore satisfy the equation:

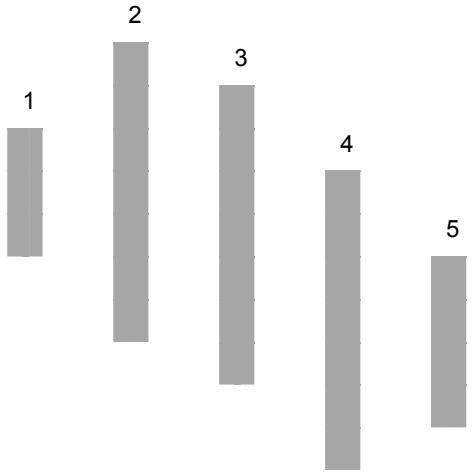
$$\sum_{-17.71}^{x_0} D_x(x, y) = 4$$

from which $x_0$ is found to be 20.17, in accordance with Wesolowsky's result.

Table 4-2 A five point problem, due to Wesolowsky

| *Destination* | $w_i$ | $e_i$ | $f_i$ | $g_i$ | $h_i$ |
|---|---|---|---|---|---|
| 1 | 1 | -7.12 | 17.12 | 5.12 | 12.31 |
| 2 | 2 | -17.71 | 37.71 | 3.15 | 14.51 |
| 3 | 1 | -16.18 | 46.18 | -5.12 | 15.12 |
| 4 | 3 | -1.18 | 61.18 | -12.12 | 42.12 |
| 5 | 1 | 19.22 | 60.78 | 5.12 | 12.12 |

| *Intervals* |
|---|
| $s_1$ = -17.71 |
| $s_2$ = -16.18 |
| $s_3$ = -7.12 |
| $s_4$ = -1.18 |
| $s_5$ = 17.12 |
| $s_6$ = 19.22 |
| $s_7$ = 37.71 |
| $s_8$ = 46.18 |
| $s_9$= 60.78 |
| $s_{10}$ = 61.18 |



The same process may be repeated for the *y*-direction. Calculations can be readily performed using a spreadsheet. Unfortunately it is not possible to extend this methodology to Euclidean or generalised $L_p$ metrics and for such problems specialised optimisation routines and location-allocation algorithms will continue to be required. The notion of constructing composite demand functions is, however, of considerable interest – it introduces the concept of selecting locations (and choosing routes) in regions where conditions such as demand, land costs and traffic density vary in space, which is far closer to real-world problems. Such problems are addressed in subsequent Chapters.

#### 4.2.4.5 Riemannian metrics

Type 6 in the table of metrics (Table 4-1) is perhaps more familiar to physicists, cosmologists and multivariate statisticians than to mainstream social scientists. It is based on the work of Riemann, which we outlined in Section 3.4. The key element of the expression for incremental distance is $g_{ij}dx_i dx_j$ (sometimes written as $g_{ij}dx^i dx^j$). By convention the summation symbol may be omitted and summation is assumed over repeated subscripts. This summation is the general *quadratic form* for incremental distance[36] in a curved *n*-space. When $n=2$ the space is a 2-dimensional surface and we may write the expression in full (using the notation $dx = dx^1$ and $dy = dx^2$, and reverting to the normal use of superscripts for powers) as[37]:

$$ds^2 = g_{11}dx^2 + g_{12}dxdy + g_{21}dydx + g_{22}dy^2$$

The $g_{ij}$ terms define the curvature of the space at every point. For the Euclidean plane, $g_{12}=g_{21}=0$ and $g_{11}=g_{22}=1$, giving

$$ds^2 = dx^2 + dy^2$$

Note that within the formulation Euclidean space is unique in that the incremental distance metric is separable in *x* and *y*, i.e. there are no combined terms.

The integral in the general formula is to be taken along the shortest path, *C*, between the two specified end points (and is a straight line in the Euclidean plane in the commonly understood sense of 'straight'). Since there may be more than one shortest path

(especially globally) the shortest of all such paths (or *infimum*) is taken and the distance value ascribed to this path.

For a general two dimensional surface $z = f(x,y)$ we have:

$$g_{11}=1 + f_x^2,\ g_{22} = 1+f_y^2 \text{ and } g_{12}=g_{21}=f_x f_y$$

where subscripts denote partial differentiation. Substituting in the expression above and simplifying, we have:

$$ds^2 = dx^2 + dy^2 + (f_x dx + f_y dy)^2$$

Thus incremental distance over a surface is simply the incremental Cartesian distance plus an element that depends upon the incremental slope of the cost function in the *x*- and *y*-directions – this is the expression for the path gradient across the surface *z*. With this measure, for every surface the incremental distance is greater than or equal to the plane distance.

Whilst the Riemannian metric provides us with information concerning the lengths of very short paths, the calculation of distance between points further apart requires information about the path, *C*, along which the integral is to be evaluated. For very simple surfaces analytic formulae can be derived, but for generalised surfaces the equations to be solved are typically intractable.

The Riemannian formulation allows us to examine a number of general principles that in turn, assist in the solution of more general network problems, including those involving variable cost fields. If we assume that *F* is a continuously differentiable[38] generalised cost function and *ds* is incremental distance, then *Fds* is incremental cost or time, and we can write:

$$D = \int_C Fds$$ where $D$ is a measure of cost or time distance over path $C$.

The minimum value of this integral is found by seeking a particular path, *C,* for which the expression is a minimum. If *F* is a constant function then *C* will be a geodesic path for the surface over which *ds* is incremental distance (e.g. a straight line in the plane).

Now let $F = e^{\sigma}$ and let $g^{*}_{ij} = e^{2\sigma} g_{ij}$ , then the equation for incremental distance above may be re-written as:

$$(ds^{*})^2 = e^{2\sigma} g_{ij} dx^i dx^j = g^{*}_{ij} dx^i dx^j$$

thus

$$D = \int_{C^*} ds^{*}$$

The transformation $g^{*}_{ij} = e^{2\sigma} g_{ij}$ is conformal[39] and thus locally shape will be preserved by the mapping. In particular angular relations between network lines will be preserved. This conformality applies to corresponding directions at corresponding points, but does not imply preservation of geodesics - the geodesic path *C** above will not normally be the image of *C* under the transformation given.

This transformation shows that least cost paths in a space *S* with metric *ds* map to shortest paths in a space *S** with metric *ds*.* Thus the problem of constructing least cost paths across a surface such as in the plane is equivalent to that of finding shortest paths (geodesics) on a related cost surface and constructing networks from these.

The problem of finding shortest paths on general surfaces, amounts to one of finding solutions to a second order (boundary value) problem of differential equations. The solution path may be expressed as a function of *x, y*(*x*) say[40], and the boundary conditions are $y(x_0)=y_0$, $y(x_n)=y_n$, where $(x_0,y_0)$ and $(x_n,y_n)$ are the initial and final points on the path. Now, let the differential equation to be solved be of the form $G = 0$, where *G* is a function of *x, y, dy/dx* and $d^2y/dx^2$. For the plane $G = d^2y/dx^2$ which is directly integrable to give:

$$(x_0-x_n)y = (y_0-y_n)x + (x_0y_n-y_0x_n)$$

i.e. a straight line. We now introduce a theorem regarding geodesic equations (a proof is provided in Annex 2), and use this to provide the basis for solution of generalised path finding and network problems:

> ***Theorem: Geodesic paths on generalised surfaces***
> *If $G = 0$ represents the differential equation for geodesics on any analytic surface, S, with metric, ds, and F is any positive-valued analytic function defined over S, then geodesics on the surface $S^*$ with metric $ds^*$ are solutions to a differential equation of the form $G + R = 0$.*

Of particular interest is the case where $S$ is a plane. In this case we find:

$$R = (1 + [y'(x)]^2)(F_x y'(x) - F_y)/F$$

where subscripts denote partial differentiation and $y'(x) = dy/dx$.

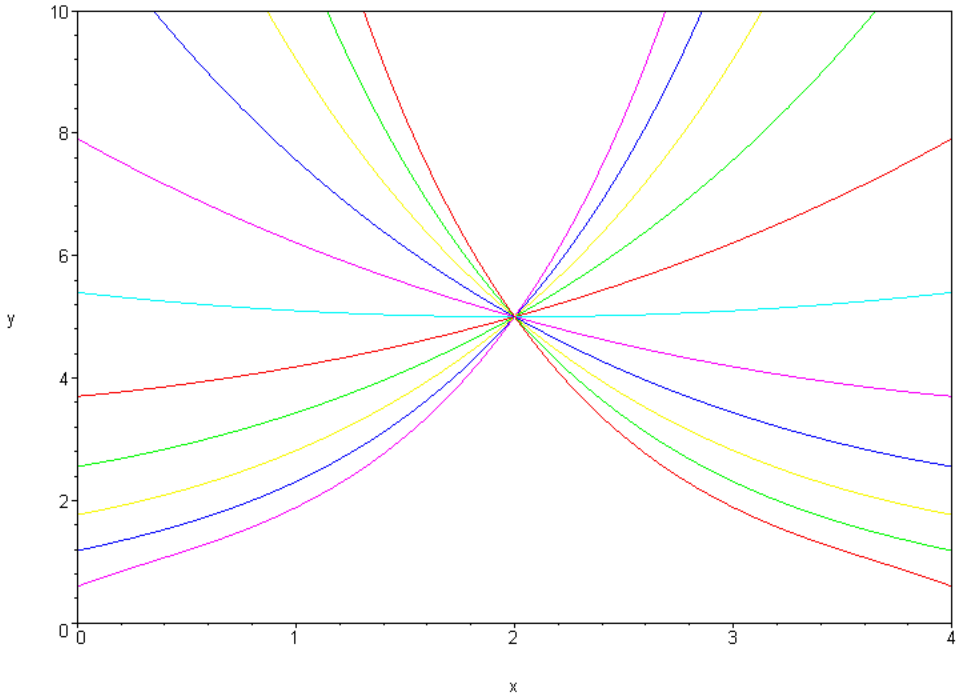
Direct solution of $G + R = 0$ is not possible for general $F$. Solutions where $F$ is either a function of one variable or is radially symmetric are given in Annex 2. For functions of one variable we have:

$$R = (1 + [y'(x)]^2)(-F_y)/F \qquad \text{for } F(x,y) = F(y)$$

$$R = (1 + [y'(x)]^2)F_x y'(x)/F \qquad \text{for } F(x,y) = F(x)$$

If $F(x,y) = k$, a constant then the partial differentials disappear and $R=0$, as is expected. If $F(x,y)$ is the equation of a plane parallel to the original surface, the term $(F_x y'(x) - F_y)/F = 0$ once again and the solution paths will be straight lines on the original surface, as before. In general, if we set $R = 0$ where $F \neq k$ then $y'(x) = F_y/F_x$ , i.e. shortest paths will remain as straight lines if the rate of change of the cost function matches the rate of change of the path. Least cost paths for a simple function of one variable: $F = ay + b$, where $a$=0.25, $b$=1 with $y(2) = 5$ are shown in Figure 4-5.

Figure 4-5 Least cost path in the plane, with cost function *F = y/4 + 1*




This case is analytically tractable with the solution being *catenary* curves, similar to that produced by a uniform chain hanging between two fixed points. Solution paths are of the form:

$$y(x) = \alpha cosh(x/\alpha+\beta/\alpha)-b/a$$

where $\alpha$ and $\beta$ are constants determined by the initial value and either the initial direction or the location of a target end point.

If the cost function is of the form $F = a/y$, i.e. an inverse function of one variable, solution paths are arcs of circles. Analytic solutions of this type do not appear to be available for more complex cost functions, except in the case of radially symmetric cost or velocity fields (for derivations of these results and a fuller discussion, see Annex 2). In the following Chapters and Annex 2 of this study we provide an extensive discussion of this family of metrics in the context of optimum route and depot or facility location problems.

#### 4.2.4.6 Spherical and terrestrial metrics

The sphere is of special significance owing to its close approximation to the shape of the Earth. Over distances of less than 20kms the Earth can be regarded as approximately

flat and the Euclidean measure applied with latitude and longitude coordinates gives results that are accurate to within 30 metres on the (spherical) Earth's surface (except for Arctic and Antarctic regions). Closer to the poles the distortion of lines of latitude and longitude result in increasing errors in this formula and a variant known as the Polar Coordinate Flat-Earth formula may be used as an improved approximation[41]:

$$d = R\sqrt{(\pi/2 - \phi_1)^2 + (\pi/2 - \phi_2)^2 - 2(\pi/2 - \phi_1)(\pi/2 - \phi_2)\cos(\lambda_2 - \lambda_1)}$$

where we are using polar coordinates ($\phi$, $\lambda$) to represent latitude and longitude expressed in radians, with $R$ = the radius of the terrestrial sphere. Using this formula the error is at most 20 metres over 20kms even at 88°N or S.

Shortest paths on a sphere have long been known to be great circles. The great circle distance between two locations is given by the *cosine* formula as:

$$d = R\cos^{-1}\left[\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos(\lambda_1 - \lambda_2)\right]$$

This formula can be subject to rounding error because the cosines of all very small angles (under 1 minute of arc) are 0.9999999 in the first seven digits. A mathematically equivalent formula (the *haversine* formula) is given below, which many authors recommend as better behaved for use over short distances/where very small angular differences are involved[42]:

$$d = 2R\sin^{-1}\left(\sqrt{\sin^2\left(\frac{\phi_1 - \phi_2}{2}\right) + \sin^2\left(\frac{\lambda_1 - \lambda_2}{2}\right)\cos\phi_1 \cos\phi_2}\right)$$

A segment of a great circle from the pole to the equator has length $2\pi R/4$. Using the WGS-84 value for the semi-major axis, $R$ (given earlier, and ignoring the flattening effect of the ellipsoid) we have a circular length of 10,018.75 kms using either of the above formulae. Both formulae yield a figure of 1855.3 metres for one minute of arc, which compares with the average minute of arc given for the WGS-84 ellipsoid of 1852.2 metres (WGS max = 1861.6m, WGS min = 1842.9m).

If a correction for the elliptical path is made an approximate adjustment is to replace $R$ with either a simple average radius or the squared average radius:

$$R_{est}=\sqrt{(R_1^{\,2}+R_2^{\,2})/2}$$

where $R_1$ and $R_2$ are the major and minor semi-axes of the ellipse. Alternatively, one can use the approximation formula for the perimeter length of an ellipse, $P$, due to Ramanujan:

$$P=\pi\left(3(R_1+R_2)-\sqrt{(R_1+3R_2)(3R_1+R_2)}\right)$$

With either of these approximations the segment length is found to be 10,001.97kms - this value matches the ellipsoidal value and is very close to the original 18th century objective discussed in Section 3.3.2, whereby a metre was defined as 1/10,000,000th of the earth segment from pole to equator.

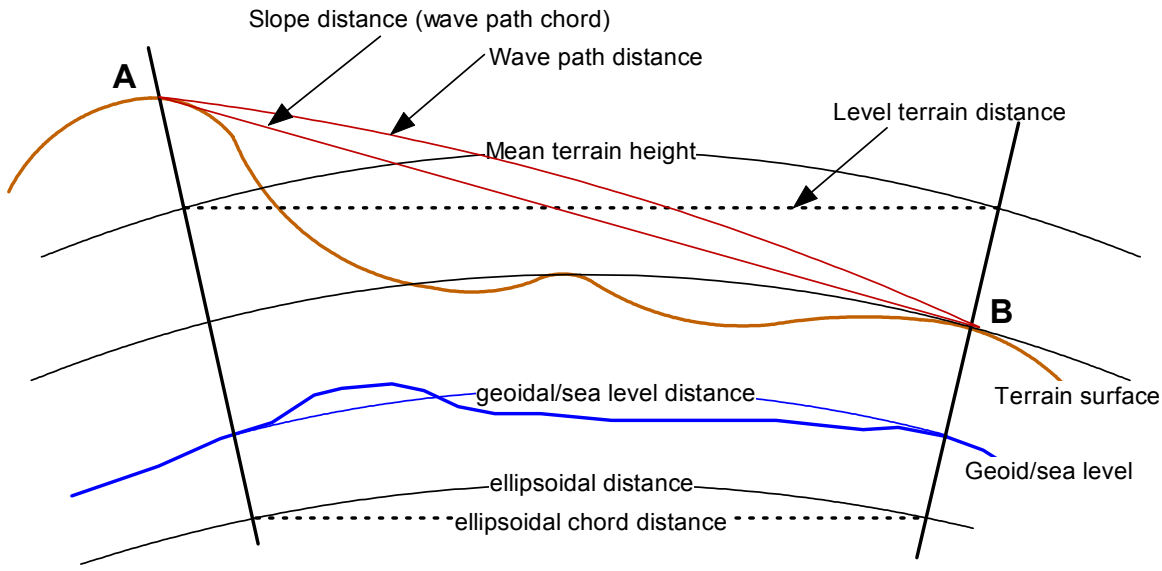
A distance calculation based on the explicit ellipsoidal model for arbitrary latitude and longitude pairs (sometimes called *terrestrial geodesic distance*) requires calculation of a series expansion and/or use of iterative procedure and must be determined numerically. A numerical algorithm due to Vincenty[43] is frequently used for this purpose. For example, the great circle distance between (30°S, 150°E) and a point 50° due South with mean ellipsoidal radius is 9kms less than the Vincenty estimate of 5565kms.

A related issue is the question of reduction of *measured* terrestrial distances to a common base, such as the reference ellipsoid. Modern electronic distance measurement (EDM) devices, which are based on light-waves or microwaves, provide very precise figures for distances between measurement stations. But these measurements are subject to a number of well-defined errors, principally the result of atmospheric refraction, which in turn is dependent on the temperature, pressure and humidity along the inter-station path (these effects are minor below 15kms in normal conditions)[44]. It is impractical in general to measure conditions along the path itself, but measurements at each end are sufficient for most situations.

Atmospheric refraction has the effect of increasing the apparent height of objects observed from a distance by a factor related to the square of the intervening distance. At a distance of 10 miles, assuming the observations are made at sea level and 65°F, the refraction equates to approximately 8.8 feet increase in height and increases to c.200 feet at 50 miles. This effect diminishes with increasing altitude and temperature, and may even reverse if there is a temperature inversion. In the Great Trigonometrical Survey of India in the 1800s, George Everest noted this effect and used it coupled with observations of specially designed bright torches at night to help increase the size of triangulation steps to 50+ miles.

The curved wave path distance can be adjusted (*reduced*) to take account of this path curvature and the relative heights of the station points[45] (Figure 4-6). The distance can then be adjusted to the reference ellipsoid to provide ellipsoidal distance, or to the modern equivalent of mean sea level (geoidal/sea level). Note that the reference ellipsoid may be above or below the geoidal surface, and neither distance corresponds to distance across the landscape.

Figure 4-6 Reduction of measured distances




*redrawn, with permission, after: GDA Technical Manual, V2.1 Australian Intergovernmental Committee on Surveying and Mapping (ICSM)*

EDM measurement is rarely used for distances of 100kms or more, and increasingly GPS and DGPS measurements of location followed by direct calculations avoid the

requirement for the wave path adjustments discussed above. EDM and GPS (separately or combined) provide the opportunity to use methods of trilateration (measuring the lengths rather than the angles of triangulations) or the hybrid technique known as triangulateration, to provide very accurate localised surveys. Such methods, whilst known about since at least the 18th century, were impractical to use until very recently. It has been found that the hybrid technique of triangulateration offers little improvement over trilateration, but statistical modelling of sample triangulations has shown that distance measurement errors are halved when trilateration is adopted.

### 4.2.5 Hybrid metrics

In addition to single expressions, which define the distance between pairs of points independent of location, hybrid metrics may be used which combine a number of single expressions according to a variety of rules. Usually these rules involve computing a set of values (in a number of cases maximum or minimum values) and selecting one of the set as the distance to use in particular circumstances.

#### *4.2.5.1 Toroidal distance*

The term toroidal distance in this study refers to the distance between pairs of points on a simple ring torus (a doughnut shape). Its use in spatial analysis has principally been as a remapping of a rectangular region, where both pairs of opposite sides of the rectangle have been joined (physically or logically). The Manhattan Street Network (MSN), described earlier, is a further example of a toroidal space, but in this case with all points constrained to lie on a directed wireframe. Logical mapping examples of toroidal spaces include the path of a point travelling across the boundary of a computer screen and reappearing on the opposite edge; and the treatment of the boundary of a rectangular map as if it were attached to the opposite side, thereby creating a 2-dimensional finite space (without boundaries) which is of use in distance statistics studies (see further, Section 6.2).

If a rectangular region has corners $(0,0)$, $(0,a)$, $(b,a)$ and $(b,0)$ then the Euclidean form for toroidal distance, $t_{ij}$, between two points, $i$ and $j$ within this region is:

$$t_{ij} = (x_{ij}^2 + y_{ij}^2)^{1/2}$$

where: $x_{ij}^2 = min\{(x_i-x_j)^2, (x_i-x_j-a)^2, (x_i-x_j+a)^2\}$, $y_{ij}^2 = min\{(y_i-y_j)^2, (y_i-y_j-b)^2, (y_i-y_j+b)^2\}$

#### 4.2.5.2 Hausdorff distance

Hausdorff [46] distance, which was developed in connection with the analysis of sets and topologies, has wide applicability in fields such as pattern recognition in 2- and 3-space, and deserves greater attention in spatial analysis. It provides a measure of similarity for objects with spatial extent – very similar objects tend to have small values for their Hausdorff distance, even if the position and orientation of the objects relative to each other is altered. Formally, Hausdorff distance, $h(x,y)$, is defined as:

$$h(x,y) = \max_{x \in X}\left\{\min_{y \in Y}\{d(x,y)\}\right\}$$

where $X$ and $Y$ are two sets (typically in spatial analysis sets of points, e.g. polygons) and $d(x,y)$ is any metric, such as Euclidean distance. This is an asymmetric *maximin* measure, and both values are of use in pattern recognition. To impose symmetry, this definition may be generalised as:

$$H(x,y) = max\{h(x,y),h(y,x)\}$$

This generalisation guarantees true metric behaviour and provides a measure of the maximum distance of a point in one set to the nearest point in the second set. It offers a way of comparing objects such as polygons – for example, comparing polygonal zones or a template (model) zone with sampled zones.

Prior to calculation it is usual to apply simple transforms (e.g. scaling, rotation) of either the template or one of the two zones such that it matches the target zone as closely as possible. However, the metric is susceptible to outliers, i.e. extreme values, which might, for example, be errors in the data. For this reason, if there are a finite number ($m$) of values, it can be more effective to generate an ordered set of values, $H_k(x,y)$, and use the $k^{th}$-largest ($k<m$) rather than the $m^{th}$.

#### 4.2.5.3 Variable routing metrics

If a space consists of a network infrastructure (e.g. a variable-speed road network, a multi-modal transport system, or a switched telecommunications network) the path selected between two points typically will be the result of some decision or choice process (which could be random, but is more frequently non-random). Non-random choices might be based on a wide range of criteria, e.g. shortest distance, minimum time, least congestion, most scenic, least visible, minimum intersections or fewest mode changes. The resulting routing may be very complex and dynamic or fairly simple and essentially static. A useful example of the latter is a development of the so-called Karlsruhe metric by Hyman and Mayhew[47].

The Karlsruhe metric is based on the assumption that the city is a circular region with all routes determined by either radial paths from the centre or ring (orbital) paths around the centre, or a combination of both, such that distance is minimised between the origin and destination points. An arbitrarily fine road network is assumed, allowing for any selection of origin and destination within the city. The metric, $K_D$, is then:

$$K_D = min\{radial\ route,\ min\{radial\ plus\ orbital\ route\}\}$$

The first term may include a single or double radial (the latter being a pair of radial paths via the city centre) and the second term is the smaller of the two 'shortest' radial plus orbital routes. Lines of equal distance from a point (the locus or distance isolines) for this metric are convex and either circular (for the city centre) or leaf-shaped (for points some distance from the centre).

Hyman and Mayhew have modified this metric by removing the assumption that orbitals exist everywhere, replacing this with a single orbital at a radius, R, from the city centre. They also modify the model to be time-based rather than distance based, in order to allow for variable speeds on different road types (radials have speed $V_r$ and orbitals have speed $V_o$). The revised hybrid metric, in polar coordinates, is then:

$$K_{T1} = \min\left\{\frac{r_1 + r}{V_r}, \left[\frac{R\,|\,\theta\,|}{V_o} + \frac{|\,R - r_1\,| + |\,R - r\,|}{V_r}\right]\right\}$$

where the start point is $(r,0)$ and the end point is $(r_1,\theta)$. The travel times generated by this metric depend upon the relative speeds of the radial and orbital routes, the relative positions of the start and end points with respect to the orbital (i.e. inside or outside the orbital) and their angle of separation, $\theta$.

Extensions of this model to handle variable radial road speeds (simple power function) and multiple orbital routes are relatively straightforward. Further extensions, for example to incorporate only a limited number of radials, variations in speed by flow direction and by time of day, are possible but make the model increasingly complex.

Another form of hybrid routing metric is one that is calibrated for multiple zones. For example, the family of metrics based on the Minkowski inequality may be calibrated on a zonal basis for improved accuracy of modelling. Zones might be urban/rural or sections of a city (e.g., the separate Boroughs of New York City). Zoning does, however, raise the issue of calculating distances or times between points in different zones. The simplest approach in this case is to assume a linear path or path of multiple linear segments, and apply each zonal metric in proportion to the linear segment within the zone in question. In other applications it may be more appropriate to consider point-centroid distances within zones and centroid-centroid distances between zones. Some statistical measures of use in such applications are provided in Chapter 6.

## 4.3 Paths and metrics

The preceding Sections introduced the terms *path* and *shortest path* without defining these terms precisely. Smith[48] provides formal definitions of path, path networks, and path length and from these, a clarification of the notion of the shortest path. We utilise his formulation in the following three paragraphs.

A path can be thought of as a set of points in a sample space, such that as one travels along the path there is an ordering of the points with respect to the start. In simple terms, if the points along the path are parameterised by some variable, $t \in [0,1]$, then the path defined by $p(t)$ satisfies the condition that for all values $a$, $b$ where $a < b$, then $p(a)$ precedes $p(b)$. The length of the path, $l(p)$, is then a function which yields a real-

numbered value for the path, such that this number is 0 for null paths, additive (i.e. the length of bits of a path adds up to the length of the concatenated bits), and finally, changing the parameter, $t$, does not affect the length. Clearly $l(p(a)) < l(p(b))$. Within the scope of this definition, a *path* is seen as a one-dimensional object that has finite extent between finite (distinct) points of the path.

A shortest path (with respect to $l$) is one for which $l$ is minimised. If this is a global minimum (i.e. is the minimum for all possible paths, $p(t)$) then it implies that every sub-section of the path, or every partition, is also a minimum between the start and end points of the partition. The conclusion one can draw from this observation is that if we generate any path between two points, $a_0$ and $a_n$, it is possible to determine whether the path is a globally shortest path (or global *geodesic*) simply by examining any sub-section of the path, say $p(a_0, a_1)$. Furthermore, if the length of this sub-section turns out not to be the shortest possible, then the subsection can be replaced with an improved path, e.g. $q(a_0, a_1)$ and the next sub-section examined. The resulting revised path will *not*, in general, be the shortest path between $a_0$ and $a_n$ since it still includes some intermediate points that may be sub-optimal. To obtain a shortest path each intermediate point must be moved (varied) to identify whether the overall path length can be reduced further. This requires an iterative process.

These definitions of path and path length correspond to our intuitive notions, but are generalised. An important result that can be drawn from this axiomatisation is that a distance function $d(x,y)$ is a shortest path distance iff $d$ is a quasi-metric, i.e. shortest path distances can only be found if all the requirements for a metric hold except *strong* symmetry. This is not especially surprising, but the converse result states that the study of quasi-metrics is equivalent to the study of shortest path distances, and thus results proven for (uniform) quasi-metric spaces also apply to shortest paths.

A further conclusion is that if a distance metric and associated space does not satisfy triangularity the concept of shortest paths may not be meaningful. This becomes especially apparent when the distance measure is some form of generalised 'cost distance'[49]. In many such cases it is not safe to assume that the sum of least cost sub-sections of a route will cost the same as the entire route – cheap day returns, penalties for stop-offs and inter-connection costs are all examples of factors affecting such

notions. Such cases are principally found in measures based upon existing infrastructure, such as physical networks and associated services. However, if one can describe costs in the sample space as some form of continuous cost surface or field, and cost distances are monotonically related to path length, then cost distance measures can be used in a similar manner to physical distance.

For example, if the cost surface is described by some function $F(x,y)$ and the distance measure by $ds = d(x,y)$ where the points $(x,y)$ are sufficiently close, then the sum or integral of $Fds$ along any path will yield a cost distance. Likewise, if the cost function is of the form $G[d(x,y)]$ where $G[.]$ meets the criterion stated earlier, then one can apply the same methods for path optimisation as are available for metric and uniform quasi-metric spaces. In practice, however, we may find that additional constraints and assumptions may be required in order to identify globally optimal paths.

## 4.4 Conclusions

In this Chapter we have seen that the notion of distance, and thus of space and spatial analysis, is more complex that it appears at first sight. We have shown that there is a wide variety of possible formulations for distance, many of which satisfy a set of rules that ensure simple operations work in a consistent manner. However, we have also seen that the conventional formulations have many weaknesses, both in terms of their inability to provide meaningful measures of separation in real-world situations, and as a result of their inconvenient analytical form.

In order to address these issues we have described a range of alternative distance formulations, in particular examining incremental and hybrid measures and the associated notion of paths. We have also investigated approximations, which may be coupled with incremental formulations to provide a powerful set of tools with which practical, real-world problems in spatial analysis may be addressed.

There are many real-world examples for which the assumptions of symmetry and/or triangularity do not hold. It is therefore appropriate to investigate many problems and applications from a perspective that does not assume Euclidean space as the *a priori* first or only choice. The use of a range of alternative metrics, semi-metrics and

generalised or quasi-metrics all have validity in the context of specific problems, although for certain types of problem (e.g. shortest path problems) metric or near metric distance measures are required.

An extension to these observations is a consideration of spatial homogeneity and continuity. There are many instances (most real-world situations) in which these fundamental assumptions are open to question. Some of these issues are discussed in the following Chapter on measurement, but some simple examples serve to highlight these issues at this point: firstly, barriers, such as walls or fences, no-go areas, one-way streets, natural features of the landscape, variations in traffic, etc. may all serve to distort any analytic measures and make them inapplicable or of limited applicability; secondly, transport infrastructures, operational rules, safety or security considerations may all require the use of different measures at different scales or in different zones – for example, one model might apply within a given distance from the city centre (e.g. within the rapid transit zone) whilst another may apply for longer journeys – there are also many issues relating to the calculation of distances and related measurements when zone sizes, shapes or arrangements are changed; and thirdly in some cases questions of continuity and dimensionality is uppermost, as for example in the measurement of boundaries and in assuming uniform availability of and access to transport capacity.

Despite the undoubted value of analytic metric space models of distance, there is the scope, and much evidence, to warrant re-examination of traditional geographic theories and models with the Euclidean metric assumption weakened, generalised, altered or discarded entirely. It is notable, however, that *incremental* metric formulations provide the foundation for much of the analysis that follows.

## Notes and References:

[1] ***Dodge M, Kitchen R*** (***2001***) *Atlas of Cyberspace, Addison Wesley, New York*

***Tobler W R*** (***1961***) *Map transformations of geographic space. Unpub. PhD thesis, Univ. of Washington, Seattle*

***Watson J W*** (***1955***) *Geography – a discipline in distance, Scottish Geog. Mag., LXXI, 1, 1-12*

[2] ***Blaut J M*** (***1961***) *Space and process, Prof. Geog., 13, 1-7*

***Harvey D*** (***1969***) *Explanation in geography, E Arnold, London, p.210*

***Beer S*** (***1970***) *Questions of metric, Oper. Res. Quart., 22* (*Conference issue*)*, 133-144*

[3] *in this study we do not consider distance measures which are binary, ordinal or some mix of such data types with fully quantitative measures, although there are many spatial and non-spatial applications for which such measures are applicable*

[4] ***Metric****: the term' metric' was introduced by* ***Felix Hausdorff*** (***1868-1942***) *in 1914 in connection with his study of set theory, topology and the notion of neighbourhoods. The so-called Hausdorff metric* (*or semi-metric*)*, discussed further below, is readily seen to be a more general, set-theoretic form than the metrics of classical geometry*

[5] ***Postulates****: (i) we postulate that every function that satisfies the standard metric criteria over a continuous space must exhibit a symmetric convex locus about the origin; (ii) conversely we postulate that if a function is defined and its locus plotted, it cannot be a metric if this locus is not convex and symmetric. No separate proof of these postulates is offered here, but proofs would appear to exist in the mathematical literature of convexity theory and topological spaces. The postulates are related to Hilbert's 4$^{th}$ problem in two dimensions. For example, it is known that any closed convex curve in the plane that is symmetric about the origin is a norm and a projective metric. The first postulate is readily seen to be true for $L_p$ metrics* (*$p \geq 1$, Figure 4-2*) *and chamfer metrics* (*Figure 5-22*)

[6] ***Jeffrey-Burroughs W, Sadalla E K*** (***1979***) *Asymmetries in distance cognition, Geog Anal., 11. 414-421*

***Beals R, Krantz D H*** (***1967***) *Metrics and geodesics induced by order relations, Mathematische Zeitschrift, 101, 285-298*

***Gould P R*** (***1966***) *On mental maps, Mich. Inter-Univ. Community of Math. Geog., Disc. Paper 9, Michigan*

[7] ***Shreider Yu A*** (***1974***) *What is distance? Univ. of Chicago Press, Chicago*

***Zaustinsky E M*** (***1959***) *Spaces with non-symmetric distance, Memoir 34, Amer. Math. Soc., Providence, R.I.; Zaustinstky provides a slightly more precise definition which includes conditions of convergence of the set of points in the sample space*

[8] ***ADSL:*** *Asynchronous digital subscriber line*

[9] ***Smith T E*** (***1989***) *Shortest-path distances: An axiomatic approach, Geog. Anal., 21, 1, 1-31, Sections 6 and 7*

[10] ***Blumenthal L M*** (***1970***) *Distance geometry, Chelsea, New York*

[11] *although it may be possible to produce a 2-dimensional representation of the data using multi-dimensional scaling techniques, albeit with a degree of 'stress'*

[12] ***Beguine H, Thisse J-F*** (***1979***) *An axiomatic approach to geographic space. Geog. Anal., 11, 325-341. In the text we have changed their notation slightly for consistency*

[13] *in this case, as a Taylor series*

[14] ***Fang L*** (***1998***) *Circular arcs approximation by quintic polynomial curves, Computer aided geometric design, 15, 843-861*

[15] ***Butt M A, Maragos P*** (***1998***) *Optimum design of chamfer distance transforms, IEEE Transactions on Image processing, 7, 1477-1484*

[16] ***Hardy G H, Littlewood J E, Polya G*** (***1934***) *Inequalities, The University Press, Cambridge*

[17] *The* ***distance*** *between P and Q is defined in terms of norms as* $D(P,Q) = D(Q,P) = N(P - Q)$. *A* ***norm****,* $N(*)$ *on* $\Re^2$ *is a function* $\Re^2 \rightarrow \Re^1$ *such that:* $N(P) > 0$ *if* $P > \underline{0}$ (*the null vector*), $N(P) = 0$ *iff* $P = \underline{0}$, $N(P+Q) \leq N(P) + N(Q)$, $N(aP) = |a|N(P)$. $N(P) = D(0,P)$, *i.e. the distance from the origin to a point P. The Euclidean norm,* $N_E(P)$ *is simply the square root of the inner* (*or dot*) *product of P with itself, i.e.* $\sqrt{P \bullet P}$. *The Euclidean norm is used as a foundation stone of Factor Analysis*

[18] ***Okabe A, Boots B, Sugihara K, Chiu S N*** (***2000***) *Spatial tesselations: Concepts and applications of Voronoi diagrams, 2nd ed., John Wiley, Chichester, England. In Section 3.7 the authors discuss the production of Voronoi diagrams under a range of metrics, including several of those discussed in this Chapter, together with one or two examples of semi-metrics*

[19] ***Maxemchuk F N*** (***1987***) *Routing in the Manhattan Street Network, IEEE Trans. on Communications, COM-35, 5, 503-512*

[20] ***Chung T Y, Agrawal D P*** (***1990***) *On network characterization of and optimal broadcasting in the Manhattan Street Network, Proc. IEEE INFOCOM '90, San Francisco, CA, pp. 465-472*

[21] ***Kruskal J B*** (***1964***) *Multidimensional scaling by optimising goodness of fit to a non-metric hypothesis, Psychometrika, 29, 1-28*

***Beals R, Krantz D H, Tversky A*** (***1968***) *Foundations of multi-dimensional scaling, Psych. Rev., 75, 127-142*

***Sherman C R*** (***1970***) *Non-metric multi-dimensional scaling: the role of the Minkowski metric, L L Thurston Psychometric Lab, Rsch. Rpt. 82, Chapel Hill, N Carolina*

***Jeffrey-Burroughs W, Sadalla E K*** (***1979***) *Asymmetries in distance cognition, Geog Anal., 11. 414-421*

***Everitt B S, Rabe-Hesketh S*** (***1997***) *The analysis of proximity data, Vol. 4 of Kendall's Library of Statistics series, Arnold, London*

[22] ***Love R F, Morris J G, Wesolowsky G O*** (***1988***) *Facilities location, Models and methods, North-Holland, New York*

[23] ***Love R F, Morris J G*** (***1972***) *Modelling intercity road distances by mathematical functions, Oper. Res. Quart., 23, 61-72*

[24] ***Tobler W R*** (***1993***) *Speculations on the geometry of geography, in Three presentations on geographical analysis and modelling, National Center for Geographic Information and Analysis* (*NCGIA*) *Technical Rpt 93-1, Univ. of California, Santa Barbara*

[25] ***Altinel I K, Oommen J, Aras N*** (***1995***) *Vector quantization for arbitrary distance function estimation, Research Rpt TR-95-19, Carleton Univ. Computer Science Dept, Ontario, Canada*

[26] ***Morris J G*** (***1981***) *Convergence of the Weiszfeld algorithm for Weber problems using a generalised "distance" function, Oper. Res., 29, 37-48*

[27] ***Muller J-C*** (***1982***) *Non-Euclidean geographic spaces: Mapping functional distances, Geog. Anal., 14, 3, 189-207*

[28] *although measurements made between any pair of points along a Euclidean straight line in the space will provide a true value for shortest path distance using the chosen metric*

[29] *See for example:* ***Love R F, Morris J G*** (***1972***) *Modelling inter-city road distances by mathematical functions, Oper. Res. Quart., 23, 1, 61-71, and* ***Vaughan R J*** (***1987***) *Urban spatial traffic patterns, Pion, London*

[30] ***Gauge distance****: the term appears to originate with Minkowski and is an extension of the study of so-called block or polyhedral norms. See for example****: Hamacher H W, Klamroth K*** (***2000***) *Planar Weber location problems with barriers and block norms, Annals of Operations Research, 96,191-208*

[31] ***Snyder R*** (***1971***) *A note on the principle of median location, J. Reg. Sci., 11, 391-394*

[32] ***Scott A J*** (***1970***) *Combinatorial programming, spatial analysis and planning, Methuen, London*

[33] ***Love R F*** (***1972***) *A computational procedure for optimally locating a facility with respect to several rectangular regions, J. Reg. Sci. , 12, 233-242;*
***Cooper L*** (***1974***) *A random location equilibrium problem, J. Reg. Sci., 14, 131-136;*
***Bennett C D, Mirathor A*** (***1974***) *Optimum facility location with respect to several regions, J. Reg. Sci., 14, 131-136;*
***Goodchild M F*** (***1977***) *The aggregation problem in location-allocation, Geog. Anal., 11, 240-255*

[34] ***Wesolovsky G O*** (***1977***) *The Weber problem with rectangular distances and randomly distributed destinations, J. Reg. Sci., 17, 53-60*
*subsequently* ***Love R F, Morris J G, Wesolowsky G O*** (***1988***) *Facilities location, models and methods, Ch., 2.2, North-Holland, New York recognised this simpler solution method.*

[35] ***Drezner Z*** (***1979***) *Bounds on the optimal location to the Weber problem under conditions of uncertainty, J. Oper. Res. Soc., 30, 923-931*

[36] ***Quadratic form****: Riemann introduced the quadratic form for incremental distance as an assumption, based upon the Euclidean model, accepting that other forms* (*such as expressions in the fourth degree*) *might be assumed as an alternative. These ideas have been explored in the context of Finsler spaces, but the tensor calculus involved is extremely complex. Gauss* (*in Theoria Motus Corporum Coelestium, 1809, section 186*) *had argued in favour of the quadratic form in connection with the development of*

his method of least squares, partly on the grounds of simplicity - he too considered that sums of $4^{th}$ or $6^{th}$ powers could be equally valid, but more complex to handle

[37] note that if the middle two terms are identical the expression can be written as $ds^2 = Edx^2 + 2Fdxdy + Gdy^2$, which is Gauss' first fundamental form as introduced in Chapter 3

[38] strictly speaking, piecewise differentiability is sufficient, with extension to functions which possess a finite set of well-defined discontinuities being possible (though complex in some cases) by applying Snell's Law of refraction to paths crossing these discontinuities

[39] ***Conformal***: shape-preserving transforms. See further, **Eisenhart L P (1925)** Riemannian Geometry, Princeton Univ. Press, Princeton, New Jersey, Section 28 "Conformal spaces. Spaces conformal to a flat space"

[40] although the parameterisation $y(t)$, $x(t)$ may be more appropriate

[41] for an excellent discussion of these issues, including analysis of alternative spherical geometry formulae see: **Chamberlain R G (last updated, Feb 2001)** What is the best way to calculate the distance between two points? at http://www.census.gov/cgi-bin/geo/gisfaq?Q5.1 Tobler in an online paper, Spherical measures without spherical trigonometry, Solstice, Vol 12, 2 has also pointed out that standard plane trigonometry may be used to compute spherical distances, since great circle arcs are simply the intesection of a plane through the sphere's centre with the surface. Distance is computed using the standard 3-D Cartesian formula with $x_1=R\cos\phi_1\cos\lambda_1$, $x_2=R\cos\phi_2\cos\lambda_2$, $y_1=R\cos\phi_1\cos\lambda_1$, $y_2=R\cos\phi_2\cos\lambda_2$ and with $z_1=R\sin\phi_1$, $z_2=R\sin\phi_2$; the Cartesian distance, C, provides the chord distance, from which the spherical distance can be determined as $S=2\pi R\alpha/180$ where $\alpha = \sin^{-1}(C/2)$ degrees

[42] this is the so-called Haversine formula, as described in Sinnott R W (1984) Virtues of the Haversine, Sky and Telescope, 68, 2, p.159. The Versine (or versed sine) of angle A is defined as 1-cos(A). The Haversine is half the versine, or (1-cos(A))/2, thus hav(A) = (1-cos(A))/2 = sin^2(A/2); many GIS, spatial analysis and mathematical programs now utilise this version of the formula. With modern computers rounding errors are now small enough to be ignored for most terrestrial applications since they will be accurate to within 0.05 seconds of arc

[43] **Vincenty T (1975)** Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations, Survey Review XXII, 176, 88-93. A spreadsheet implementation of this algorithm can be found at: http://www.auslig.gov.au/geodesy/datums/calcs.htm

[44] this problem was well-known as early as the $18^{th}$ century and was one of a number of factors which prevented early surveyors from estimating the height of very large mountains with any accuracy. Refraction varies substantially by time of day

[45] diagram redrawn from Geocentric Datum of Australia Technical Manual,V2.2, Intergovernmental Committee on Surveying and Mapping, Feb 2002, p.2-8

[46] after **Felix Hausdorff (1868-1942)**. Hausdorff appears elsewhere in this study in connection with the formal definitions of metric, metric space and (fractional) dimension. As a Jew in Nazi Germany Hausdorff was scheduled for internment in a concentration camp – instead, in 1942 he, his wife and her sister committed suicide.

[47] ***Hyman G, Mayhew L*** (***2000***) *The properties of route catchments in orbital-radial cities, Environment & Planning B, 27, 843-863*

[48] ***Smith T E*** (***1989***) *Shortest-path distances: An axiomatic approach, Geog. Anal., 21, 1, 1-31, Section 3.*

[49] ***Huriot J-M, Smith T E, Thisse J-F*** (***1989***) *Minimum cost distances in spatial analysis, Geog. Anal., 21, 4, 294-315*

# 5 Distance, Path Measurement and Path Models

*Drawing upon the preceding discussion of distance and metrics, we now examine the practical and theoretical questions raised when attempting to carry out distance measurement. A range of circumstances is investigated in which distances are measurable and meaningful, together with those for which such measurement is problematic or impossible. We then review alternative models of path, including abstract classical models, fractal models, a number of alternative statistical models, and finally lattice models.*

*In connection with statistical models we analyse paths from the perspective of point data recording and associated point-pair uncertainty, providing new error estimates for path lengths and profiles. Statistical methods are also applied in connection with measurement of existing linear forms and in the provision of models for simple, linked and closed paths, and tree networks.*

*In connection with lattice representations, Distance Transforms and their associated local neighbourhood metrics* (*chamfer metrics*)*, familiar to those working in image processing, are discussed and examined in some detail. The first of a series of developments of conventional Distance Transforms is described, in this instance involving obstacle avoidance and associated path diffraction. Further development of this group of Transforms is provided in Chapters 7 and 8.*

## 5.1 Measurability

The present study concentrates upon distance measures, metrics and metric spaces. It is therefore important to identify which areas of interest to spatial analysts lend themselves to such measurement and which do not.

Gatrell[1] has examined the application of alternative formulations to metric space in geographic research in some detail, focusing upon the use of multidimensional scaling (*mds*) and "Q-analysis". He provides a much broader view of geographic spaces as a set of objects together with a *relation* defined between pairs of objects. He argues that we should be less concerned with physical distance or absolute location and more with the relationships between objects, such as places or people. Physical distance is regarded as just one of a large set of possible relations. Gatrell then proceeds to examine time-distance, cost-distance, cognitive distance and social distance relations. The set of such relations and objects may then be subjected to metric or non-metric *mds* techniques to analyse the dimensionality of the data and to generate two-dimensional (i.e. map-like) presentations of the data with a known degree of stress. However, as Cliff and Haggett[2] have observed, the *mds* procedure has distinct limitations:

> *"... we see multidimensional scaling as akin to the problem of erecting a hideously complex frame tent with thousands of tent-poles, each of different lengths. Although the poles may fit perfectly in a high-enough hyperspace, they become ever more stressed and skewed as the dimensions are progressively reduced towards those lower-order spaces which the human brain can comprehend."*

There are many situations in which measurement is problematic, some of which may be amenable to scaling techniques such as *mds*. The various cases may be grouped into a number of *classes*, which are not strictly independent but are separated below in order to highlight a number of key issues.

### 5.1.1 Class 1 - Qualitative

This is the class of (spatial) information that either cannot, or has not, been collected in a quantifiable form because of its descriptive nature. Opinions, emotions and general descriptions are examples. There has been some work to reconcile certain descriptive terms (e.g. lake, mountain, region) with scientific requirements of precision but this

remains far removed from the realms of numerical measurement and mathematical analysis. Ordinal scale measurement is often used to attempt to quantify information this class - e.g. survey data asking people to rank their opinions on the desirability of a new city bypass. Such data generally do not meet metric space requirements although it has been shown that ratio scales (which are suitable) can be derived from ordinal datasets under some circumstances[3].

### 5.1.2 Class 2 - Discontinuous

There is an argument for stating that much geographic data actually exhibits an infinite number of discontinuities – for example, when seeking to measure the length of a coastline we are aware that both the measurement method and the scale used to compute distances affects the lengths recorded[4]. Mandelbrot treats this problem by examining a variety of deterministic and stochastic processes in the plane, and interpreting the problem in terms of fractional dimension (similar concepts were put forward in earlier work by several authors, including W Bunge and L F Richardson[5]). The fractional dimension viewpoint represents an approach to immeasurable sets which is very appealing and visual, but which strongly conflicts with classical notions:

> *"The question of the number of dimensions is closely linked to the notion of continuity and it would have no meaning for anyone who wished to exclude this notion."*
> *Poincaré* (*1913, p27-8*) [6]

The problem of continuity (and the related problems of scale and uncertainty) arises immediately one takes a closer and closer look at anything in the real world. All coastlines (boundaries, rivers, roads…) have apparently infinite length since one could weave in and out of the finest particles (sand grains, fences, riverbanks, road-stones) forever, and at the finest levels of observation the location of basic subatomic particles is indeterminate. Indeed, in map-making and associated spatial analysis we use a line to *represent* a linear structure, not to be an accurate *rendition* of that structure itself. In the same manner, we use points to *represent* point-like structures, but do not expect these to lie in one-to-one correspondence with the subjects that they are representing. The underlying structures in both cases have complex 3-dimensional spatial extent, which we choose to ignore for selected purposes.

In geographic research (and in mathematics) we avoid such questions by making convenient assumptions and/or basing our analysis on pre-defined scales and methods of measurement and representation. This approach disguises central difficulties of the notion of dimension, continuity, distance and space – one must always be aware of the implicit and explicit decisions made when collecting, representing, storing, using and interpreting spatial data. It should not then come as a surprise when problems arise which are scale-related within the particular scales of interest. Such data are not absolute and their use and interpretation is relative to the underlying data and the purposes for which the data are being used. Data may or may not display consistency (self similarity) at different scales of observation – if self-similarity is observed an indication of fractal-like behaviour over these scales may be suspected (but not assumed). Similar behaviour is found when examining physical or electronic traffic data[7] – patterns may be observed which show self-similarity over a range of scales, but which do not, of themselves, offer additional insight into the processes at work, of a kind needed for predictive modelling.

### 5.1.3 Class 3 - Uncertain[8]

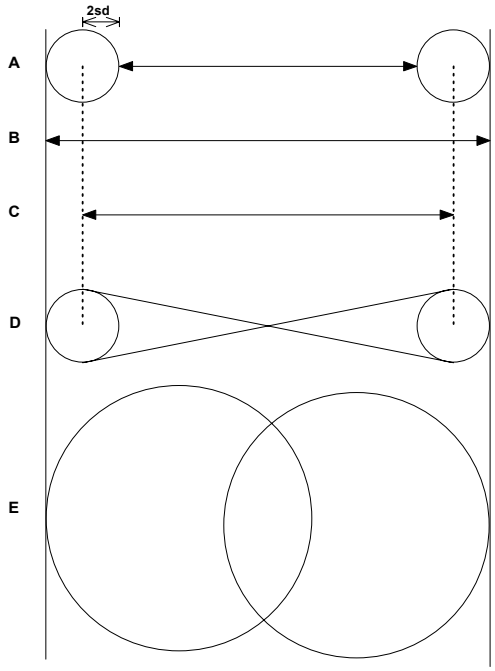
In addition to quantum and scale-related uncertainty issues there are many cases where the data to be collected is known to be an approximation: based on the accuracy of measurement instruments with a known (or unknown) resolution and error pattern (e.g. as discussed in earlier Chapters); the susceptibility of measurement to the skill and behaviour/presence of the observer; the environment in which measurement is taking place; the need to sample data (such as attributes of objects within regions or at sample points); and the processes (systems, software etc.) involved in data conversion and interpretation (e.g. edge detection). Other uncertainty issues include temporal questions (is all the data contemporaneous, is it still correct/is it changing); consistency (was all the data collected in the same way, stored in the same, totally consistent manner, what abstraction processes applied in its recording, storage and retrieval?); and the underlying vagueness or imprecision of the items being measured.

#### *5.1.3.1 Point pair uncertainty*

Positional uncertainty of point and line features have been the subject recently of detailed analysis, particularly in respect of the quality of point and linear data for storage and retrieval[9]. The measured and recorded position of a single point (point-like

object) can be regarded as a sample from a population, with a two-dimensional (circular) or three-dimensional (spherical) statistical distribution (e.g. Normal, Uniform). If a path then is represented by a finite set of such points, the path profile will be subject to error (typically additive errors). If we suppose that there are two points subject to error which are close to each other then the errors will combine to generate errors in the path profile, the distance measurement and the path angle (Figure 5-1).

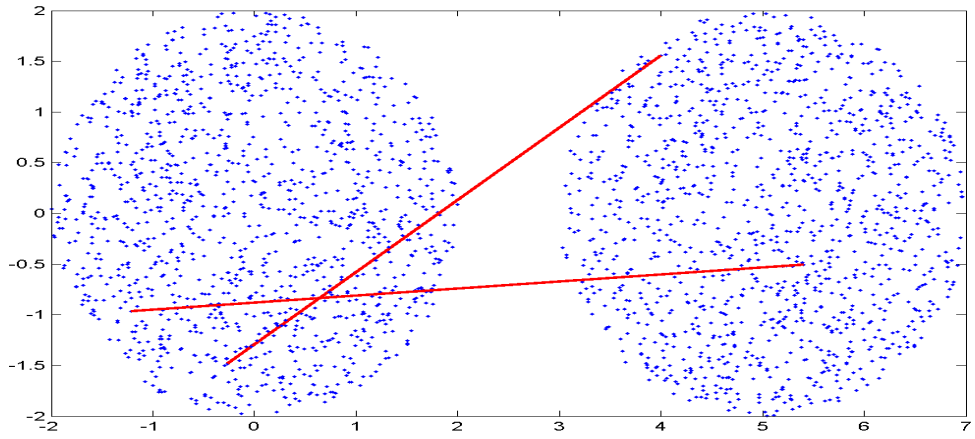
Figure 5-1 Positional errors – 2D model



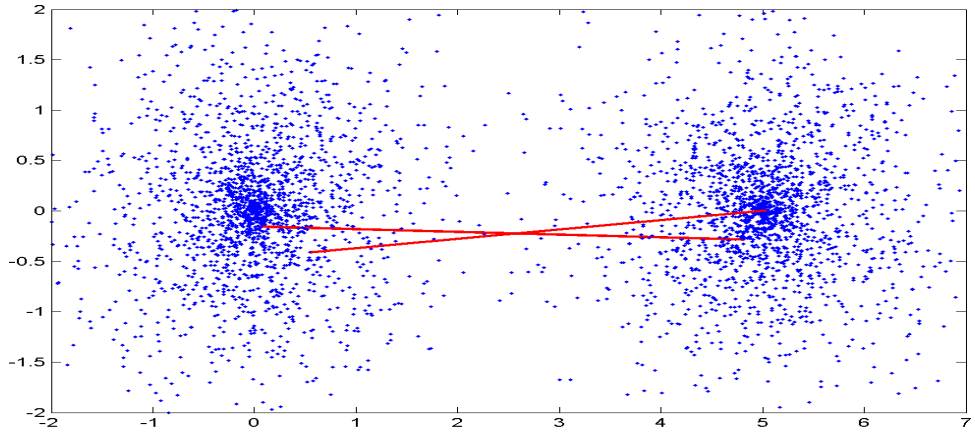

In Cases A-D we assume that the true location of each point is at the centre of the circular regions shown, each of which is 2 standard deviations (*sd*, or $\sigma$) in size and based on an identical, independent bivariate (polar) Normal distribution with no systematically introduced errors. This model may be compared with that of uniform distribution of each point within a circle of radius 2 (Figure 5-2A and B):

Figure 5-2 Uniform and Normal distributions – 2000 sample point pairs

A. Uniform model: random point pairs and sample lines
separation, s = 5.0, mean distance = 5.2



B. Polar Normal model N(0,1): random point pairs and sample lines
separation, s = 5.0, mean distance = 5.1



The mean distance between randomly chosen points in these regions is slightly greater than the true centroid-centroid separation, $s$ , i.e. distance $D_t$, Case C above. A first approximation to this mean distance can be derived from a result provided in Chapter 6 (due to Bouwkamp[10]) for the expected mean distance, $E(r)$, between two points uniformly and randomly located within separate *non-overlapping* circles of radius $R$ and $s$ units apart. In the present case, the approximation for the expected distance in a Uniform model with circle radius $R=2$ and separation $s = 5$ is:

$$E(r) = s(1 + 1/s^2 + 5/12s^4) = 5.2033$$

This approximation provides an upper bound for the equivalent Normal distribution model since almost all sample points will lie within $2\sigma$ of the two centres and on average, sampled points will be closer to the centres than under the Uniform model.

A number of observations should be made regarding these results:

(i) the Normal distribution diagram shown above has been generated used polar coordinates with radius selected from $N(0,1)$ and angle selected from a uniform distribution over the range $(0,2\pi)$. This may be a good model for the problem under consideration, but is not the same as assuming both the $x$ and $y$ components of each point are independently distributed as $N(0,1)$ - the latter distribution is more evenly spread (and has been widely studied in statistics, often known as Circular Error Probability, or CEP). The Bouwkamp model gives a surprisingly good approximation to the mean value for this latter method of random point generation where the separation is taken as $s>=2$ and the radius of the Uniform model circles is $R=2$

(ii) in the model where the $x$ and $y$ coordinates are separately selected from independent $N(0,1)$ distributions, and both pairs of coordinates are selected from distributions with the same centroids (i.e. $s=0$), then the theory of summed squared normal variates can be used ($\chi^2$ distributions) with a change of variables, to produce a simple probability density function (*pdf*) with mean value $p=\sqrt{\pi}$ (essentially this mean value is a simple Gamma integral)[11]. From this result we can derive a good estimator for the mean value in this model for separated samples using the cosine rule for triangles, with separation $s>=0$, as $d=\sqrt{[p^2+s^2-2ps\cos\theta]}$, where $\theta=\pi/2.08$ (derived experimentally). This estimator still appears to be very slightly biased, but has the advantage over the Bouwkamp-based estimator that it applies for all $s>=0$. An explicit pdf for the above case where the centroids are $s$ units apart does not appear to be easily derived

(iii) in the Polar Normal model the same estimator model for $d$ can be used, but with $p=1.204$ and $\theta=\pi/2.05$ (both $p$ and $\theta$ have been derived from experimental evidence). As before, this estimator appears to be very slightly biased.

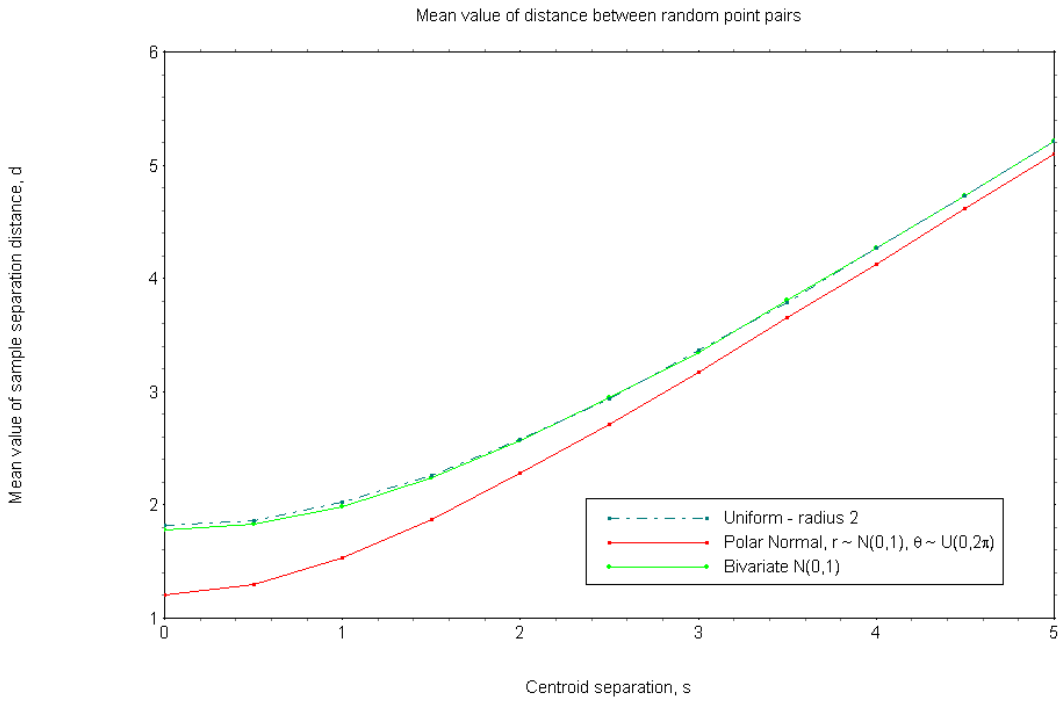
Model and sample values for $d$ when $s$ = 5, 2, 1 and 0 are given in Table 5-1 (figures in brackets are the means computed from a run of 1 million samples using random normal variates generated by the software package, MATLAB). This data is shown in graphical form in Figure 5-3, based on values of $s = 0(0.5)5$:

Table 5-1 Mean distance models for point pairs

| s | Bivariate Normal | Polar Normal | Uniform* |
|---|---|---|---|
| 0 | 1.772 (1.773) | 1.204 (1.201) | 1.811 (1.818) |
| 1 | 1.982 (1.988) | 1.535 (1.534) | - (2.016) |
| 2 | 2.591 (2.565) | 2.295 (2.276) | 2.552 (2.571) |
| 5 | 5.203 (5.206) | 5.098 (5.102) | 5.203 (5.201) |

** for the Uniform case, model values are derived from Bouwkamp for s>=2 and from the exact mean value (256/45π) for the case s=0, R=2 derived by many authors (as discussed in Chapter 6)*

Figure 5-3 Graphs of mean distances – sampled data pairs




What these model results show is that errors in path length are likely to increase as a result of point positional errors when samples are taken at intervals which are less than around 5 times the standard deviation associated with each point; sampling intervals of 10+ times the positional standard deviation provides a safer guide level for sampling - however, of course, sampling frequency needs to be a balance of representational

benefit and the management of positional errors. Using the Polar Normal model with $\sigma=1$ and a set of 100 co-linear points at true intervals of 10 units, the line length will be over estimated by roughly 0.3% (1003 units), compared with 2% (510 units) if the true intervals were 5 units with the same standard deviation.

The maximum measurement (Case B, Figure 5-1) would be $> D_t+4sd$ and the minimum (Case A) would be $< D_t-4sd$. The distance recorded, therefore, could be $D_t+/-4sd$ or more, or an error of $8+sd$. If the points were 10 units apart and $sd=0.5$ units, then a 5% root mean square (*rms*) error in point location would appear as distance measurements varying by up to 40%.

The angular variation in this Case (D) could be 22° or more, assuming the true path is a linear segment over the interval. Case E illustrates the major problems that will occur if the positional error is large in relation to the spacing of adjacent sample points. In this case it is possible that the locations of the two points could be confused and angles and distances would be almost meaningless.

Unlikely as this scenario may appear, situations of this kind do occur - for example, in attempting to locate a point (e.g. an intersection, a vehicle) on a given road, a modest error in point positioning could result in the point being located off the route, on an adjacent route or on the wrong carriageway of the current route. These observations show that the *path* between two adjacent sample points may be subject to error, and hence paths and path lengths are in general subject to such errors, although these are not necessarily cumulative. To minimise such factors (or at least, to account for them explicitly) it has been suggested that paths should be considered as having error bands, rather as per the measurement bands discussed later in this Section (see Section 5.2.6). However, the use of a constant width of band in this case is questionable[12] and variable width and probability profiles have been suggested. It is unlikely that any single model of uncertainty can be ascribed to linear features, since such 'errors' will be dependent upon, and only meaningful in the context of, specific datasets and circumstances – there is no reason to suppose that a model of linear uncertainty which applies to the manual production of polygonal regions abstracted from pre-existing maps at a given scale has close similarity to computerised linear feature detection from high resolution satellite images of the same region. Construction of such models should be based on the

fundamental components and processes by which the observed path has been produced, and the purposes to which such measures might be put, bearing in mind the intrinsic underlying uncertainties highlighted above.

### 5.1.4 Class 4 - Relative

There are at least two areas in which relativistic concepts need to be included in geographic research and analysis. The first involves an extension of the observer influence noted in our discussion of Uncertainty above. The second involves the application of the mathematics of curved spaces (notably curved surfaces), which is widely used in Relativity Theory, to the solution of problems involving the determination of least-cost or least-time paths.

The paper by Roberts and Suppes[13] on the geometry of visual perception is a particularly interesting contribution to the *a priori* nature of space. From a review of controlled experiments with human vision (head and eye movements restricted), such as those of Blank[14] and Platt[15] they conclude that:

> *"those physical curves* (*or loci of light*) *seen as straight or aligned are not the usual Euclidean straight lines, but rather certain hyperbolic curves….the physical curves 'seen as straight' appear to be hyperbolas that are convex to the primary point of fixation."*

This finding, together with their discussion of the laws of Donder and Listing on the movement of the eye, led the authors to conclude that:

> *"… our elementary properties should be divided into two classes, primitive and learned. Those such as Euclidean straightness, parallelism and the like are learned, while such factors as contiguity, boundary and closedness are probably primitive"*

This conclusion concerning the nature of primitive visual space is completely at variance with conventional views and provides further evidence that geometries other than that of Euclid are worthy of attention in the human sciences. More recent published research and limited-scale experiments conducted by the present author upholds these conclusions in broad terms, but has demonstrated that there are large variations in the metrical structure of perceptual space, varying by position, problem and observer[16].

There are many studies that have shown how the *perception* of space differs from conventional notions of physical space – distances are frequently incorrectly estimated by people, resulting in distortions with increasing distance that have been shown to be hyperbolic, logarithmic or similar[17]. Over time, and as individuals move around in space, such distortions alter, making it clear that for certain situations and problems, spatial measures are relative to the observer and the time at which they are taken. The observer (measurer, data user) may be completely unaware of these distortions and so, being as it were 'embedded' in the space, will behave as if the distortions do not exist. An outside observer, however, would see a very different picture.

This may be exemplified by considering a region (a two-dimensional universe) that is 'hottest' at the centre and radially 'cooler' as it extends to the extremes of the region. This temperature-like variation has the effect of systematically deforming everything in the same manner until it disappears completely at the margins. For a person travelling from the centre of this universe to the periphery, the universe would appear infinite, since each step taken would be shorter than the previous one and this change could not be detected. The Dutch artist, Maurits Escher, provided an impression of such a space in his Circle Limit III woodcut (see Figure 5-4A).

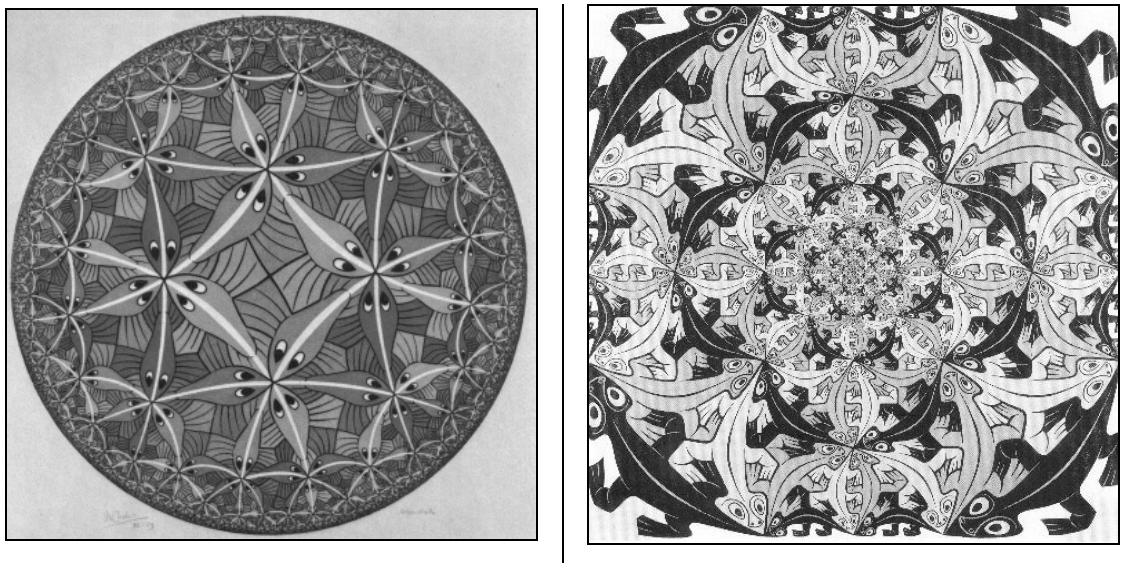
Reichenbach[18], p.32-33, has expressed this phenomenon in the form of a theorem:

> *"Given any universal geometry, G, to which measuring instruments conform, we can imagine a universal force, F, which affects the instruments in such a way that the actual geometry is an arbitrary geometry, G', while the observed deviation from G is due to universal deformation of the measuring instruments."*

Reichenbach's theorem provides a particularly clear description of this aspect of absolute versus relative space and time. Indeed, similar concerns affected the early development of accurate timekeepers: Christian Huygens, in 1690, dismissed the notion that metallic expansion affected pendulum rates based, in part, on contemporary evidence from the measurement of the length of brass pendulums with brass rules in Europe and the tropics – unsurprisingly the studies were unable to detect the true changes that occurred[19]. In fact brass is quite susceptible to temperature changes – as

noted in Table 3-3 a change of 20°C will alter a one second pendulum and bob made of brass by just over 16 seconds/day[20].

Figure 5-4 M C Escher - Woodcuts



A. Circle Limit III

B. Smaller and Smaller

*Escher's woodcuts, Circle Limit I, II, III and IV were all based on Poincaré's planar model of hyperbolic space (compare this with the geometry of Figure 5-5A). Escher did not produce a directly equivalent woodcut for the geometry of Figure 5-5B, but his design "Smaller and Smaller" is similar in concept*

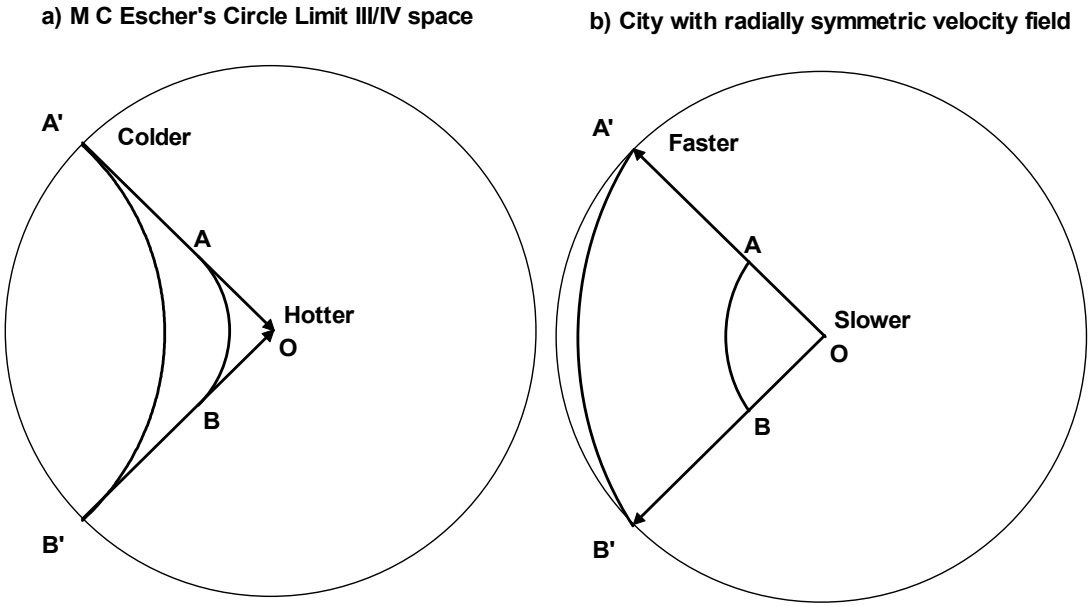

Poincaré (op. cit.), following a similar argument pointed out that the universal force, *F*, would have to represent a continuous transformation. He continues:

> *"Space, when considered independently from our measuring instruments, has therefore neither metric nor projective properties; it has only topological properties."*

The second area in which relativistic concepts have applicability is in the formulation of minimum distance, time or cost paths in the presence of scalar or vector fields. Referring back to the imaginary universe described earlier, we might ask what is the behaviour of a person travelling in a straight line[21] between two arbitrary points, A and B, within this space? Since our space-deforming 'temperature' variation is perhaps not a realistic condition, we might replace this with a radially symmetric scalar cost or travel time field (cf. Section 8.1.3.2 and Section 12.3 Radially symmetric cost functions). In this case a straight line (shortest/ quickest path for an embedded traveller) will appear to an external observer to be a smooth curve convex towards the universe centre, since advantage can be taken of faster journeys nearer to the centre (Figure 5-5A). The

reverse pattern would be seen if travel times/costs were higher in the centre (e.g. an urban area) than at the periphery, as in Figure 5-4B, Figure 5-5B and Figure 5-6A. Note that in both cases radials are always shortest/quickest paths. As was shown in Section 4.2.4.5 the shortest path in such imaginary universes are, in fact, straight lines in an associated curved space – the two views are simply different perspectives on the same phenomenon.

Figure 5-5 Optimal paths in alternative space models




Time is frequently a priority and provides more meaningful units for distance determination. For example, in order to journey from Bromley in South-East London to Sunbury-on-Thames in South West London, the quickest route may well be one that is far longer in terms of physical distance, and commence in a direction almost opposite to the true bearing between the start and end points (Figure 5-6A). The advantage of a (hopefully) high speed ring road more than counter-balances the extra journey length – in fact in this case the journey *time* is expected to be 30% less than the shortest path even though the *distance* travelled is over twice as far.

The theoretical foundations for the design of such radial-ring-radial road networks are due in part to the work of Professor Smeed and his colleagues at University College London and the Transport and Road Research Laboratory in the 1960s and 1970s (see further, Section 6.5).