Tile Intersection Probabilities for Lines and Areas in GIS

Ashton M. Shortridge and Michael F. Goodchild
National Center for Geographic Information & Analysis, and
Department of Geography
University of California, Santa Barbara, CA, 93106-4060, USA
Phone: (805) 893-8652
Email: ashton@geog.ucsb.edu, & good@geog.ucsb.edu

Keywords: classification, resolution, uncertainty analysis

Abstract
This paper identifies analytical and empirical methods for determining the probability that lines and areas intersect tiles in a regular tessellation. Such intersections are common in geographic information systems (GIS) applications. Knowledge of intersection probabilities is valuable in many instances, including estimating complexity and time required to process a distance operation, developing optimal tiling schemes for national georeferencing systems, precalculating the number of map sheets a spatial feature may occupy, and identifying appropriate cell resolutions for vector-to-raster conversions. Buffon's Needle-type solutions from the field of geometric probability provide the framework for deriving probabilities for lines. Probabilities for simple areas like rectangles and circles are derived using geometric techniques. Employing such probabilities may yield more rigorous and theoretically informed results from GIS analysis, leading to better decisions and greater insight into spatial phenomena.

1. Introduction
The intersection of one spatial feature with another is a fundamental geographic operation. A particular, though still quite general, instance of intersection occurs when one of the features is a tile in a regular tessellation, and the other is a straight line or area. Several examples follow: calculation of distance between two points referenced in the United Kingdom National Grid; determination of the number of United States Geological Survey 7.5' quadrangle map sheets required to cover a specific watershed; rasterization of a vector landcover map. In such instances, the probability that a feature will extend beyond a single tile in the tessellation is of interest. This paper develops a general framework for determining this probability.

Many applications in geographic information science involve the analysis of data that are stored or measured in regular tessellations on some projected portion of the earth's surface (Boots, 1999). In such instances the earth's surface is modeled as a planar region R, where R is completely covered by a repeatable pattern of regular tiles T, such that all locations in R fall in one and only one T. These tiles might be quadrats in an ecological study region, square cells in a raster map, rectangular sheets in a map series, or zones in a georeferencing system. Grid reference systems are ubiquitous in GIS for spatial data indexing; national grid reference systems like that used in the United Kingdom have been proposed for the United States, and non-rectangular global tessellations have also been developed (Goodchild & Yang, 1987).

Solutions for intersection probabilities on these sorts of tesselations appear to be underreported in recent GIS literature, though their mathematical underpinnings are identifiable in texts on geometric probability (see for example Klain and Rota, 1997,

Solomon, 1978). This is unfortunate, given the power of such solutions to characterize computational complexity of GIS algorithms and thereby estimate their efficiency. Several works in previous decades have looked at related problems. Switzer (1975) examined the probability that a randomly chosen point within a cell was misclassified. He determined this probability as a function the classification of nearby cells. Goodchild (1980) summarized and extended research on area estimation accuracy. This approach was primarily concerned with the relationship between grid resolution and map boundary complexity, which is beyond the scope of the present work. Bregt and others (1991) explored the relationship between cell size, polygon complexity, and rasterization error for a particular map series. Maling (1989) reported on several applications of geometric probability for manual calculation of line length and area measurement in cartometry.

The next section presents solutions for the intersection of straight lines on tiles in a tessellation. These lines could be actual features or could represent the relationship between two points. Section Three covers intersection probabilities for areas. Probabilities of simple areas to intersect tile boundaries are presented first. Then, probabilities for simple areas to intersect tile central points are covered. Section Four discusses the utility of such solutions for GIS, and proposes extensions.

2. Buffon's Needle-type solutions for lines

Geometric probability is a branch of mathematics that is concerned with the probabilities associated with geometric configurations of objects. Among the most famous of these applications is the Buffon's Needle problem. The 18th century French naturalist Buffon conceived of the problem while considering a popular game of chance in which a stick was thrown upon a tile floor. The problem continues to appeal to students of mathematics for its variety of elegant solutions. The classic problem and its solution are as follows. Parallel, equidistant lines are spaced s units apart in the plane. A straight needle of length l, where $l < s$, is dropped onto this plane at random. The probability $P(i)$ that the needle will intersect one of the lines is:

$$P(i) = \frac{2l}{\pi s} \qquad (1)$$

Many clever proofs for this solution and for extensions to the classic problem have been developed over the years and may be found in several sources, including Klain & Rota (1997), Solomon (1978), and Uspensky (1937).

A variation of the classic Buffon's Needle is of particular interest here. In this variation, the parallel lines of the original problem are replaced by a rectangular matrix of square cells s units on a side. The "needle", a straight line segment of length l, may also be regarded as two point locations separated by distance l. When the distance between two points is smaller than the grid cell resolution, and the grid is oriented independently of the location of the points, the case is known as the Laplace extension of the Buffon problem (Solomon, 1978). The converse problem of the "long needle", when $l > s$, has also been studied but will not be covered in this paper.

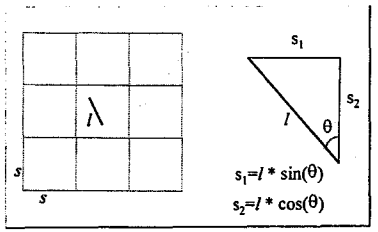

Figure 1, Geometric representation of the short needle (l<s) problem.

Figure 1 illustrates the Laplace extension. The position of the needle is determined by the two coordinates x, y of its midpoint within a grid cell and its angle θ relative to the horizontal axis of the grid. The domain comprised of all possible values of x, y, and θ is a parallelepiped volume with sides: $0 < x < s$; $0 < y < s$; $-\pi/2 < \theta < \pi/2$

The volume of this domain is πs^2. The volume of the domain representing needle positions entirely within the grid cell is:

$$v = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} (s - l\cos\theta)(s - l\sin\theta)\,d\vartheta \qquad (2)$$

Solving the integration and dividing by πs^2 gives the probability $1\text{-}P(i)$ of both points falling in the same cell:

$$1 - P(i) = 1 - \frac{4ls - l^2}{\pi s^2} \qquad (3)$$

So the probability of the points falling in separate cells is:

$$P(i) = \frac{4ls - l^2}{\pi s^2} \qquad (4)$$

The solution for the more general case of a rectangular tiling with sides a and b, provided in detail in Uspensky (1937), is:

$$P(i) = \frac{2l(a+b) - l^2}{\pi ab} \qquad (5)$$

where $P(i)$ is the probability of a needle of length l ($l < a <= b$) intersecting a tile boundary.

Equations (4) and (5) are of relevance for that set of geographic problems that involve the overlay of grids or rasters on point and line data. These problems include developing quadrat grids for plant species surveys, trip length studies (Kirby, 1997), point interpolation to raster grids, and any spatial modeling effort using GIS operations to convolve raster and vector data. In some cases, a concern may be that point to point interactions will be missed due to the coarseness of the quadrat coverage. Kirby (1997) investigated the specific application of this case to transportation surveys, but the problem is more general.

In other instances, the main concern may be that the line does not cross a tile boundary. Consider for example the Ordnance Survey's National Grid for the United Kingdom. The

United Kingdom, projected in transverse Mercator, is overlain with tiles 100 km on a side to enable quick reference of feature locations. Calculation of distance between an origin and destination is trivial if both are in the same tile. However, the calculation becomes more complex if boundaries must be crossed (Ordnance Survey, 1998). Figure 2 plots trip distance against the probability that the trip crosses a tile boundary, using values from equation (1). From this graph it can be seen that the probability of crossing a boundary increases beyond 0.5 when trip distance reaches 45 km.

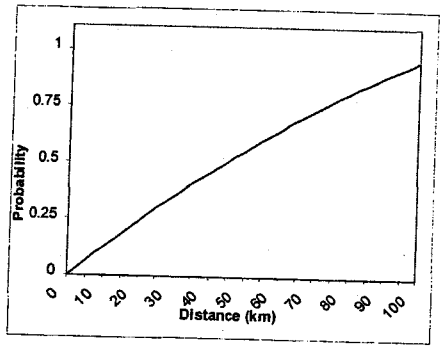

Figure 2. Probability of 2 points a given distance apart falling in different UK National Grid zones.

Eq.s (1) and (2) may also be employed for the inverse problem, that of estimating computing time for determining distance. Let t_1 equal computational time when both points are in the same cell, and t_2 (where $t_2 > t_1$) equal computational time when the points are in different cells, and P = Equation (1). Then the expected time to solve the problem, E(t), is:

$$E(t) = t_1(1-P) + t_2(P) \qquad (6)$$

This sort of calculation is useful not only for determining probabilities of intersections in an existing system, but also for developing appropriate grid or cell resolutions when planning specific geographic applications.

3. Intersection probabilities for simple area features

This section covers boundary intersection probabilities for rectangles and circles. Such problems are of interest in the calculation of the number of map sheets a feature of a given size is likely to occupy. Alternatively, the probabilities could be used to identify an optimal tile size into which spatial data can be divided without increasing the chance that features of interest would be split onto multiple tiles.

A certain "law of geography" states that everything of geographic interest lies on the border of two or four map sheets (Clarke, 1995). This truism is all too familiar to anyone who has used multiple map sheets or images to identify a lake, watershed, or other area of interest. In fact, given the dimensions of a rectangular tile in a regular tessellation, the probability that a rectangular or circular feature will intersect a tile boundary is straightforward to calculate. This simple feature might be the area of interest itself, or it might represent a bounding box or radius about the area's central point. The probability of intersecting exactly two or exactly four separate tiles may also be derived.

First consider a rectangular tiling of the plane with tile dimensions a and b. Positioned within this plane is a rectangular area with sides s and t oriented to the tile, where $s \le t < a \le b$. It is clear that position of the rectangular area is solely a function of the placement of the center of the rectangle within a single tile.

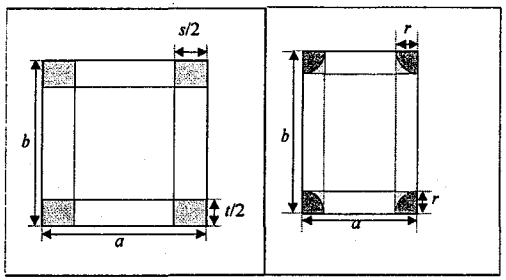

Figure 3. a. If the center of a rectangle st falls in the darkest zone it intersects four tiles. If it falls in the lighter gray zone it intersects two tiles. b). A circle with radius r intersects 4, 3, 2, or 1 tiles, if its center falls in the zones shaded from darkest to lightest, respectively.

The shaded area in Figure 3a represents that portion of the tile that the rectangle center must fall in for it to intersect a tile boundary. The probability P(i) for this to occur is equal to the ratio of the area of the shaded region to the total area of the tile, that is:

$$P(i) = \frac{at + bs - st}{ab} \qquad (7)$$

It is also possible to calculate the probability that the rectangle intersects exactly two or exactly four tiles. For this to occur, the center point of the rectangle must fall in the lightly shaded region or the darkest shaded region, respectively, of Figure 3a. These probabilities are:

$$P(i2) = \frac{sb + at - 2st}{ab} \qquad (8)$$

$$P(i4) = \frac{st}{ab} \qquad (9)$$

Now consider similar problems for a circular area instead of a rectangle. As with the rectangle, the circle's location is determined solely by the location of its center, which must fall within a single tile. The probability of intersecting is a function of tile size and the radius of the circle. Figure 3b shows the "small circle" case, in which the diameter is smaller than the minor side of the rectangular tile.

Consider first the probability $P(i)$that a "small" circle with radius r intersects at least two tiles. The shaded region of Figure 3b represents that portion of the tile that the circle center must fall within for it to intersect more than a single tile. The proportion of this shaded area to the total area of the tile total shaded region of the tile equals this probability:

$$P(i) = \frac{2r(a + b - 2r)}{ab} \qquad (10)$$

where a and b are the sides of the tile. The probabilities of intersecting exactly two, three, or four tiles for a small circle are:

$$P(i2) = \frac{2r(a+b-4r)}{ab} \quad (11)$$

$$P(i3) = \frac{r^2(4-\pi)}{ab} \quad (12)$$

$$P(i4) = \frac{\pi r^2}{ab} \quad (13)$$

respectively. The numerator in each case represents the corresponding shaded area of the tile in Figure 3b.

These equations can be used to calculate the probability that a particular rectangular or circular feature will overlap two, three, or four map sheets, without the absolute location of the feature being known. They could also be employed to develop spatial data tiling schemes to maximize the probability of capturing specific features within a single tile.

4. Discussion and conclusions

This paper has reported probabilities for the intersection of lines and areas on regular tiles in a tessellation. The probability of an arbitrarily oriented "short" ($l \quad s$) segment intersecting the tile was identified. This case has relevance for a variety of GIS applications including distance calculation in zonal georeferencing systems like the United Kingdom's National Grid. A second set of solutions dealt with probabilities for the intersection of simple areas and regular tiles in a tessellation. These probabilities could be used to develop optimal tile sizes for specific applications or to estimate the number of map sheets required to cover an area of interest.

In the spectrum of activities that comprise the development and use of a GIS, geometric probability appears to have greatest application in data modeling decisions. If critical application specifications are known (e.g. mean trip distance for a proposed routing system, or wetland polygon size for a land information system) then optimal tile sizes can be established. Techniques from geometric probability also have utility for developing accuracy estimates and data quality statements for derived datasets.

Lack of space proscribed discussion of the relationship of geometric probability to cell classification error in vector-to-raster conversion. When converting vector class data to raster, a classification rule must be used to assign a value to each raster cell. One possibility is to use the class of the polygon occupying the central point of the cell, a point-counting method of area measurement (Maling, 1989). A concern with this approach is that small polygons may "disappear" in the conversion process because they fail to overlap any cell center. The probability that a polygon is missed during conversion to raster is clearly a function of raster cell size and polygon shape and area. In fact for simple polygons, analytical solutions similar to those developed in Section three exist that calculate the probability that a rasterization algorithm classifies a polygon correctly for a given cell size. Simulation provides an alternative for more complex, "real world" shapes. These probabilities are useful for determining the probability of "missing" small objects when converting vector data to raster grids, or of misreporting the area occupied by a particular class.

This paper has concentrated on intersection problems involving regular square and rectangular tiles. A great many alternative tessellations are employed in GIS, however; other regular tessellations employ triangles and hexagons, while irregular (triangular and otherwise) tessellations are common in surface models and area class maps, respectively (Boots, 1999). Do analytical methods exist for some of these tessellations? Can geometric probability inform the selection of tessellation form and size for more cases than those covered here?

A related problem is the location of the needle in the tessellation. In Section 2 of this paper the location of the needle was drawn from a uniform probability distribution. In many instances needle location is not uniformly distributed, but subject to a more complicated probability density function. Consider for example a classic monocentric urban area with radially symmetric decline of traffic density from the core (Angel and Hyman, 1976). Trip origin and destination in this environment are driven by population density and their frequencies are exponential functions of distance from the core. This distribution has important implications for trip-length survey design, in which maximizing the probability of an intersection is desirable (Kirby, 1997). For line problems generally, alternatives to the uniform distribution for location are important for developing appropriate tessellation strategies.

Much GIS functionality is descriptive and atheoretical, with little or no notion of whether observed quantities or relationships are significant. The work in geometric probability described here provides solid mathematical underpinnings for some fundamental GIS operations. Solutions for such problems may exist in other literatures, but they may not always have been communicated to the GIS community, or applied to geographic problems. The introduction of approaches like these may yield more rigorous and theoretically informed results from GIS analysis, leading to better decisions and greater insight into spatial phenomena.

Acknowledgements

Kevin Curtin's interest and assistance contributed to the research; any errors are solely the responsibility of the authors. This work was financially supported by the National Center for Geographic Information and Analysis and the National Imagery and Mapping Agency (grant #NMA 202-97-1-1021).

References

- *Angel, S. and Hyman, G.M., 1976*, Urban Fields , Pion.
- *Boots, B., 1998*, Spatial Tessellations, in: Geographical Information Systems: Principles and Technical Issues, 2nd Edition, Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W., Wiley & Sons, pp. 503-526.
- *Bregt, A.K., Denneboom, J., Gesink, H.J., Van Randen, Y., 1991*, Determination of rasterizing error: a case study with the soil map of the Netherlands. Int. J. Geographic Information Systems 5(3) pp. 361-367.
- *Clarke, K.C., 1995*, Analytical and Computer Cartography, 2nd Edition, Prentice Hall.
- *Goodchild, M.F., 1980*, Fractals and the accuracy of geographical measures, Mathematical Geology, 12(2), pp. 85-98.
- *Goodchild, M.F., Yang, S., 1992*, A hierarchical spatial data structure for global geographic information systems, CVGIP: Graphical Models and Image Processing 54(1), pp. 31-44.
- *Kirby, H.R., 1997*, Buffon's Needle and the probability of intercepting short-distance trips by multiple screen-line surveys. Geographical Analysis, 29(1), pp. 64-71.

- *Klain, D.A., Rota, G.-C., 1997*, Introduction to Geometric Probability, Cambridge University Press.
- *Maling, D.H., 1989, Measurements from Maps: Principles and Methods of Cartometry*, Pergamon Press.
- *Ordnance Survey, 1998*, Calculating the distance between two points using their National Grid references, Accessed 4 March, 2000,
- http://www.ordsvy.gov.uk/literatu/info/cr212.html
- *Solomon, H., 1978*, Geometric Probability. CBMS-NSF Regional Conference Series in Applied Mathematics 28, Society for Industrial and Applied Mathematics.
- *Switzer, P., 1975*, Estimation of the accuracy of qualitative maps, in: Display and analysis of spatial data, Davis, J.C., McCullagh, M.J., Eds., Wiley.
- *Uspensky, J.V., 1937*, Introduction to Mathematical Probability, McGraw-Hill.

Proceedings Accuracy 2000, Amsterdam, July 2000