INTRODUCING GEOMETRY WITH SPATIAL SCAN STATISTICS

JENNIFER M. FRANKO, RACHEL SCHWELL, AND WILLIAM E. WOOD

Abstract. This module explores the role of non-Euclidean metrics in spatial scan statistics, a tool in biosurveillance. The problem is framed in a biological context while being formulated in mathematical terms. This includes a description of the algorithm and appropriate formulas, followed by both computational and philosophical questions. The formal definition and properties of metrics are also given and illustrated through discussions and exercises. The ultimate goal is highlighting the role and potential importance of geometry in the non-traditional setting of biosurveillance.

Target Audience: The intended target audience is junior-level students whose majors include mathematics or related fields. However, leaving out the specifics of metrics makes this module appropriate for any math or critical thinking course, for example, Discrete Math.

Prerequisites: This module assumes some level of mathematical maturity. Some exercises and projects may require some more mathematical sophistication, programming ability, or the ability to download and install software, but these can be easily skipped.

1. Introduction

1.1. Biosurveillance is the monitoring of biological conditions, for example disease spread, for the purpose of detection of outbreak. We shall explore in this module some accepted techniques for outbreak detection that also invite an opportunity to introduce a more flexible way of thinking about distance.

Consider the following problem: we measure the quantities of infected people in a divided region and compute the percentage we would expect to be "normal". We then want a technique that will check for pockets of "abnormal" amounts of infected people, as this will indicate an outbreak. Spatial scan statistics is one method that searches for

The authors would like to thank the Reconnect Program at the DIMACS institute at Rutgers University and Nina Fefferman in particular for support and encouragement in preparing this manuscript.

these pockets.

DISCUSSION 1. What could be appropriate definitions for the word "outbreak"? What could be appropriate definitions of "normal" and "abnormal"? What are some of the problems that could arise with these definitions? Discuss different ways to divide the region.

1.2. As in all modeling techniques, we must make assumptions about our data to reduce the complexity. First, we ignore any demographic bias. For our purposes, we begin with a very basic geographic model in which all regions are imposed on a regular grid and uniformly populated. Each node represents a region and all individuals in the region are assumed to reside at the node. Lastly, our input variable X is Boolean, i.e. each individual is assumed to be either sick or not sick.

DISCUSSION 2. Is it problematic to ignore demographic bias? Why or why not? Since it is extremely unlikely that a population would be distributed equally among evenly-spaced nodes, discuss grid systems that would more accurately reflect an actual population distribution (could be based on either population density or distance). What kind of issues are being ignored by considering the variable to be Boolean?

1.3. The basic procedure given by spatial scan statistics is as follows: we pick what will be considered "centers", in our case the nodes, and construct discs about these nodes of varying radii. We consider every disc of every radius possible. We then calculate a quantity called the *likelihood ratio* for each of the discs. To find the *most likely cluster*, we search through all possible candidate clusters in our space (discs), calculate the likelihood ratio for each cluster, and report the largest. This is what could indicate an outbreak.

DISCUSSION 3. What do we mean by "all discs of all radii" (is this possible)? How many is this? Does increasing the grid size by a small amount alter the number of possible discs by a comparably small amount?

2. Mathematical Formulation of the Problem

2.1. Imagine you are flipping a fair coin n times. Recall that a simple decision tree yields the formula probability of the coin's landing on heads or tails k times:

$$P(n,k) = \binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k}$$

where $P(n,k)$ is the probability of obtaining k heads (or tails) in n tosses of the coin. This is an example of a *Bernoulli process*, where some random quantity that takes on either of two values (heads/tails) is examined. Disease spread can also be modeled in this way (sick/not sick). A similar calculation to that above leads to the formula for the *likelihood function* on the disc $D(n,r)$:

$$L(n,r) = \left(\frac{c}{p}\right)^c\left(\frac{p-c}{p}\right)^{p-c}\left(\frac{C-c}{P-p}\right)^{C-c}\left(\frac{(P-p)-(C-c)}{P-p}\right)^{(P-p)-(C-c)}$$

where P is our total population, C is the total number of cases, p is the population in the disc $D(n,r)$, and c is the number of cases in $D(n,r)$. Note the similarity of this formula to the fractional terms in the formula for the Bernoulli process. The binomial coefficient that does not appear in the likelihood function exists in the Bernoulli formula because it counts the number of possibilities of having k heads in n tosses (i.e. the number of applicable branches in the decision tree). This coefficient is not included in the likelihood function since we already have our data and are thus considering one particular outcome (branch), rather than counting all the possible outcomes. Observe that the multiplied ratios in the likelihood function represent, in order, the proportion of cases within the disc, the proportion of non-cases within the disc, the proportion of cases outside the disc, and the proportion of non-cases outside the disc.

We now define the likelihood ratio for the disc $D(n,r)$ as:

$$\Lambda(n,r) = \frac{L(n,r)}{\left(\frac{C}{P}\right)^C\left(\frac{P-C}{P}\right)^{P-C}}$$

where P, C are as above. Observe that the average number of sick people, or the expected value of X, is $\frac{C}{P}$. We only consider discs for which the likelihood ratio is greater than the expected value by setting $\Lambda(n,r)$ equal to 0 otherwise. Finally, we identify the most likely cluster by choosing the disc $D(n_0,r_0)$ such that $\Lambda(n_0,r_0) =$

$$\max_{n\in\{\text{nodes}\},r\in\mathbf{R}^+} \Lambda(n,r).$$

Some statistical observations about the likelihood ratio:
In terms of hypotheses, the likelihood function (the numerator of the likelihood ratio) represents the maximum probability of the observed event (the data collected) under the alternative hypothesis (that the proportion of diseased in the cluster is higher than the expected value over the entire region). The denominator remains the same for each disc, and represents the probability of the observed result under the null hypothesis (that the disease incidence in the cluster is equal to the expected value). Thus, the likelihood ratio will be high when the probability of a person's having the disease is much higher within the cluster than it is in general.

Note: In all of the examples and assignments below, we compute the likelihood ratio and then take its natural log to obtain what we will simply call its *likelihood*. We do this not only because it yields "nicer" numbers, but because it is actually the case that the distribution of the random variable $-2\log\Lambda$ approaches a chi-square distribution.

Discussion 4. The likelihood function we picked was based on the premise that a person either had the disease or did not have the disease. We entirely ignored the possibility that a person could carry the disease but not be infected. Consider how the likelihood function for different models might change. For example, you may wish to direct your students to the SatScan manual [K2] to view other likelihood functions.

Having defined likelihood ratio, our problem now reduces to straightforward optimization. In practice, of course, we cannot search through every possible disc we could draw in our space. For our initial examples, recall we make the simplifying assumption that our clusters are discs centered at the population nodes.

We begin with a ludicrously simple example. The fictitious state of North Rectangula has twenty-five counties arranged in a perfect 5×5 grid, each with a population of 500. See Figure 1. Rather than track the exact location of every citizen, our model assumes that everybody in each county is located at a single point (the *centroid*). In this example the centroids of neighboring counties are spaced 100 miles apart. Suppose County 13 (in the middle of the state) has 50 infected people, and exactly one person in every other county is infected. Unsurprisingly, the most likely cluster is the point (or, "disc of radius zero") at the centroid of county 13. If we set our coordinates so that County 1

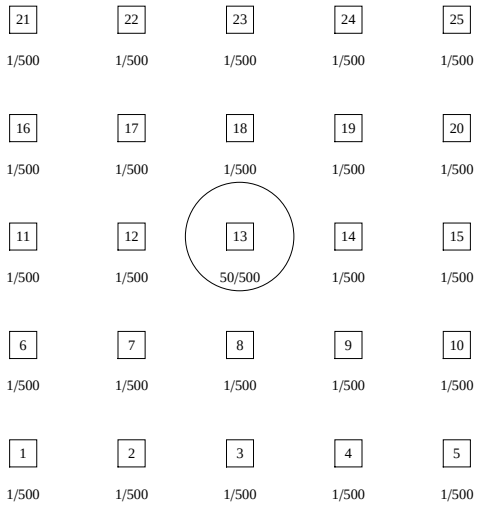


FIGURE 1. A simple cluster

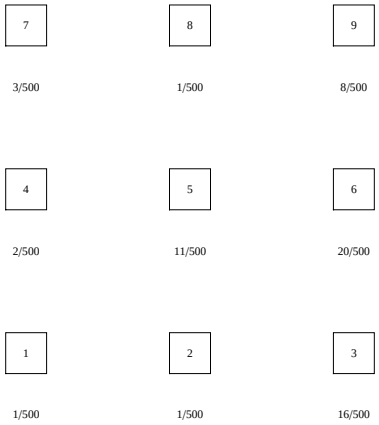


FIGURE 2. Find the most likely cluster

is at $(100, 100)$, then the centroid of County 13 is at $(300, 300)$. The likelihood for this cluster is 117.7. If we increase the number of infected in county 13 from 50 to 100, the likelihood increases to 272.1. For comparison purposes, it turns out that the cluster centered at $(200, 200)$ of radius 100 miles (thus including counties 2, 6, 7, 8, and 12) has likelihood 5.2.

EXERCISE 1. In Figure 2, what is the likelihood for the cluster centered at $(100, 100)$ with radius 50 miles? Assume that the coordinates

are as in Figure 1, with distance between each node being 100 miles.

EXERCISE 2. How many different likelihood values are there in Figure 2?

EXERCISE 3. Find the most likely cluster in the configuration shown in Figure 2 (recall that we assume the clusters are centered at the centroid nodes). You may want to implement a spreadsheet to aid in the calculations. Also, observe that it may not be necessary to compute the likelihood for every cluster.

EXERCISE 4. (programming). Implement a naïve version of this algorithm that exhaustively searches for the highest likelihood among all clusters in a 3×3 grid. Assume all clusters are centered at the nodes. You should choose an appropriate step size and maximal radius in running your search.

3. CHANGING THE GEOMETRY

3.1. Metrics. In our examples up until this point, we organized our clusters in terms of the distances between the nodes. Specifically, we used the *Euclidean distance*, which defines the distance between two ordered pairs (x_1, y_1) and (x_2, y_2) as $d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. Besides familiarity, there is no *a priori* reason to use this particular function – is there something to be gained from trying something else?

Suppose u, v, and w are points in a set (in our case, coordinates on a map). A *metric* is a function that reports the "distance" between two such points, where "distance" is not necessarily the Euclidean distance with which we are familiar, but rather can be any function that satisfies the following properties:

(1) (identity of indiscernibles) $d(v, w) = 0$ if and only if $v = w$.
(2) (symmetric) $d(v, w) = d(w, v)$.
(3) (triangle inequality) $d(v, w) + d(u, v) \geq d(v, w)$.

EXERCISE 5. Use properties (1)-(3) in the definition of a metric to prove that if d is a metric then $d(u, v) \geq 0$ for all u, v in the set. (Combined with the identity of indiscernibles, this property is called *positive definite*.)

EXERCISE 6. Consider the function $d_1((x_1, y_1), (x_2, y_2)) = 10\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. Verify that this is also a metric. Look at the previous examples, and interpret d_1 as the usual Euclidean metric with a change of units.

Now we will introduce something a bit more exotic. Consider the set of ordered pairs (x, y) where x and y are integers. Instead of the Euclidean metric, we define $d_t((x_1, y_1), (x_2, y_2)) = |x_1 - x_2| + |y_1 - y_2|$. For example, $d_t((1, -2), (-5, 3)) = |1 - (-5)| + |-5 - 3| = 14$. This is called the *taxicab metric* because you may think of this as the metric a taxi driver uses when driving around a perfectly gridded city. The Euclidean metric is not quite appropriate because the driver cannot cut diagonally through a city block.

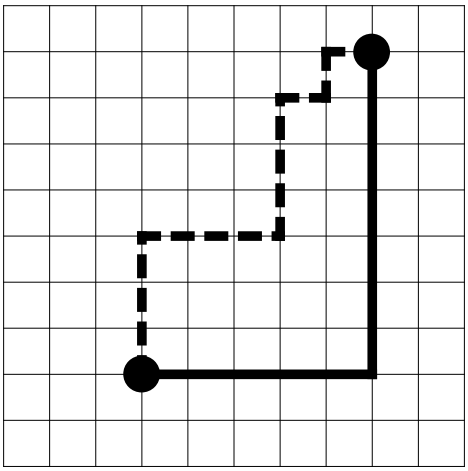


FIGURE 3. Two shortest paths of length 12 in the taxicab metric

One feature of this metric that distinguishes it from Euclidean geometry is that shortest paths are not unique. To get from $(0, 0)$ to $(2, 2)$, one could move 2 units up the y-axis and then two units along the x-axis, or instead one could go 2 units along the x-axis first and then move two units up the x-axis. You could also go up one, right one, up one, right one. All of these give paths that are four units long, and there is no shorter route. See Figure 3.

EXERCISE 7. Verify that d_t satisfies the definition of metric.

EXERCISE 8. Sketch all shortest paths from $(0, 0)$ to $(3, 4)$ in the taxicab metric.

DISCUSSION 5. Compare the taxicab metric on the grid with the comparable metric on the triangular lattice depicted in Figure 4.

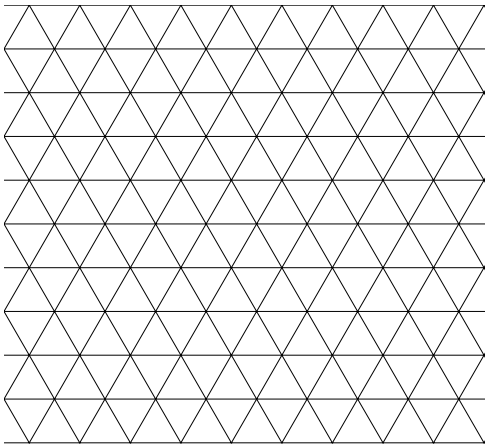


FIGURE 4. A triangular lattice

We would also like to generalize our definition of a circle. If d is any metric, r any non-negative number, and x_0 is a point in our set S, then define the *circle* of radius r and center x_0 to be the set of all points whose distance from x_0 in the metric d is exactly r, or $\{x \in S : d(x, x_0) = r\}$. Similarly, a *disc* is the set of points whose distance from x is less than or equal to r. (Technically, we should call this the *closed* disc, to distinguish it from the *open* disc which includes only points whose distance from x is strictly less than r.)

EXERCISE 9. Describe circles of radius zero in general. Prove your answer.

EXERCISE 10. Describe circles and discs in taxicab geometry.

EXERCISE 11. How would you define an *ellipse* in the taxicab metric? Sketch some examples. (For more fun with taxicab geometry, see [Kr])

EXERCISE 12. Define d_D on *any* set so that $d_D(x, y) = 0$ if $x = y$ and $d_D(x, y) = 1$ otherwise. Verify that this is a metric. What are the

circles of different radii?

EXERCISE 13. Now suppose that d_D is again defined on any set, but that $d_D(x, y) = 1$ if $x = y$ and $d_D(x, y) = 0$ otherwise. Is this also a metric? If so, prove it. If not, which property(s) does it fail to satisfy and why (give a counterexample)?

EXERCISE 14. Plot the function $f(r) =$ the number of points in a taxicab disc of radius r.

There is a qualitative difference between the Euclidean and taxicab metrics in that the Euclidean metric works on a continuous space (the plane) whereas the taxicab metric works on a discrete space (the integer points). For example, consider how discs change in the two geometries as the radius increases. In Euclidean geometry, the discs get bigger and bigger continuously as the radius increases. In taxicab geometry, however, we notice that the disc of radius zero is just a point, but so are the discs of radius $\frac{1}{4}, \frac{1}{2}, \frac{3}{4}$, and so on until the radius is 1, when suddenly four new points appear. The disc does not change again until we hit the next integer.

It is also possible to interpret the Euclidean metric as discrete by simply restricting the space to include only the integer points. That is, the distance from $(0, 0)$ to $(2, 2)$ is still $2\sqrt{2}$, but there are no points a distance of exactly, say, 3.5 from $(0, 0)$. Just as in the taxicab metric, discs grow in sudden jumps rather than smoothly. If this feels familiar, it may be because this exactly what we were doing when we explored most likely clusters!

3.2. Airborne Hazard. On April 26, 1986, there was an explosion at the Chernobyl nuclear power plant in the Ukraine. Radioactive materials were blown northward into Sweden and Finland. There was an uneven dispersal of the fallout as winds and weather conditions initially kept the dangerous materials away from Kiev. However, changing wind and weather patterns spread clouds of radioactive substances throughout Europe. Due to lack of information regarding the accident, panic spread as certain foods were banned and there was a run on iodine medicine used to combat the radioactivity [Gr]. Continued contamination in foods and forests were natural concerns as approximately 80% of reindeer Sweden were destroyed due to high levels of caesium [R]. Also, incidences of thyroid cancer in Belarus rose dramatically with over 4000 cases among those 0-18 years at the time of the accident, having been diagnosed by 2002 [B].

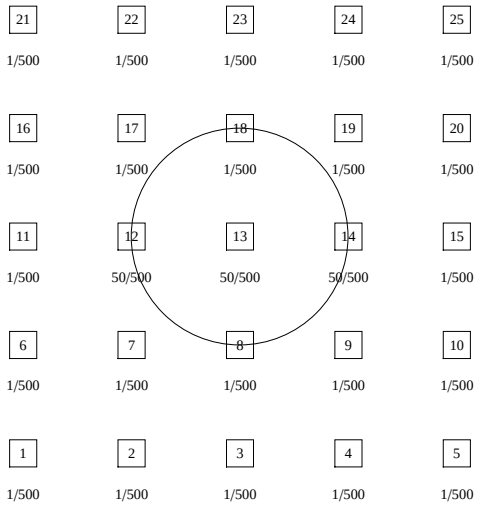


FIGURE 5. A Euclidean circle of radius 100

Suppose the disease now in question is airborne, and the winds of North Rectangula tend to blow east-west. We would like to include this information in our analysis. Consider the scenario pictured below in which there are 50 cases in three horizontally connected counties (counties 12, 13, and 14) and negligible illness elsewhere. The most likely cluster using the Euclidean metric includes our three high-incidence counties, but the use of Euclidean circles precludes catching all three without also including the two unwanted counties 8 and 18. See Figures 5 and 6.

We might have expected this sort of arrangement if we had thought about wind tendencies. Anticipating east-west distributions, we may adjust our metric as follows:

$$d_{ew}((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + 9(y_1 - y_2)^2}$$

For example, $d_{ew}((0,0),(100,0)) = 100$ whereas $d_{ew}((0,0),(0,100)) = 300$. The metric weights vertical distance more heavily than horizontal distance. In this metric, the most likely cluster contains precisely the three high-incidence counties, giving a higher likelihood than the Euclidean cluster containing the two unwanted counties (likelihood 206.2 versus 149.8).

EXERCISE 15. Verify that d_{ew} is a metric in which the circles are Euclidean ellipses whose vertical radii are one third of their horizontal

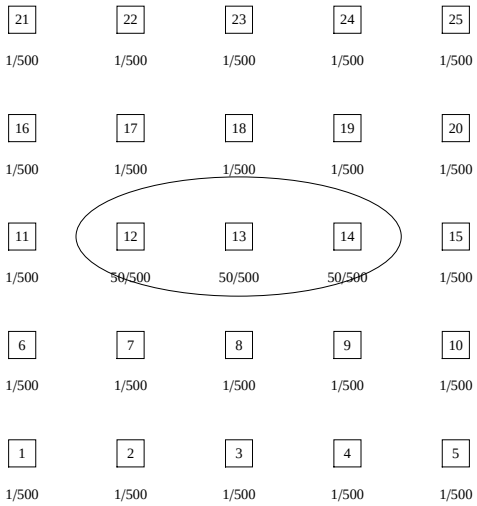


FIGURE 6. In $d_e w$, the circle looks like a Euclidean ellipse.

radii.

EXERCISE 16. This example imposed the requirement that the horizontal distance be scaled by a factor of three, but what if we wanted more flexibility in the shapes of our ellipses? How would the result change if the factor were four instead of three? How could we adjust the spatial scanning algorithm to allow *any* ellipse as a cluster? Can you anticipate any disadvantages to doing this?

3.3. Waterborne Disease. In 1849, John Snow began to look at the geographical distribution of incidences of cholera to determine the origin and spread of the disease [S]. He observed that doctors viewing the bodies of victims after death did not seem to become infected with cholera and he noted that they would carefully wash their hands after such inspections. However, those in contact with water from washing soiled linens seemed more likely to become infected [H]. He used this and other arguments to hypothesize that the spread of the disease was caused by bacteria in the water supply, i.e. that cholera was a waterborne disease [S]. During the 1854 London cholera epidemic, he was able to test this theory when he suggested removing a water pump handle on Broad Street [T]. After this, deaths did decline, but it was not clear that the epidemic subsided because of the removal of the pump [Ko]. Nevertheless, by using geographical information, he was able to make a connection between the water supply and cholera.

Suppose we suspect that a given illness may be linked to a water source such as a river which runs through our region. How could we verify our conjecture? Can we develop a method to do so? We claim that a simple adjustment to our spatial scan statistic procedure will aid in this detection. We begin by re-evaluating our idea of "distance", i.e. we consider a different metric. However, we must be wary when we do so not to induce false positives or false negatives. For example, if the disease is not water-borne but there are cases of infection near the river which cause our metric to include the entire river in a cluster, this is a false positive. A false negative could be produced if the disease is indeed water-borne but our algorithm fails to detect this due to sporadic placement of the cases of infection along the river.

DISCUSSION 6. Suppose South Rectangula has a river flowing through it as in Figure 7. We are concerned about an outbreak of a waterborne disease, and would like to adjust the clustering procedure to account for increased disease spread for counties connected by water. How might we do this?

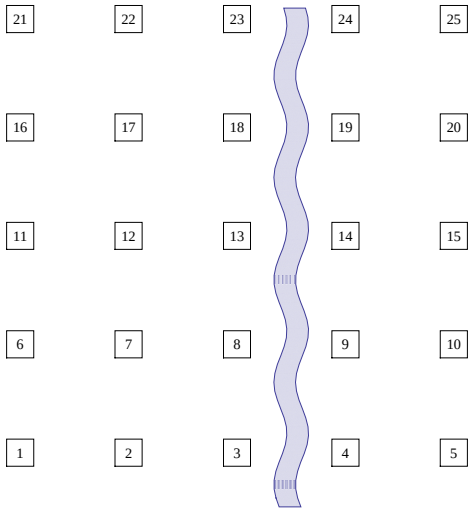


FIGURE 7. South Rectangula

One approach to accounting for the river is to *quotient out* by the river. This means that we consider all counties adjacent to the river to be equivalent, the reasoning being that if an outbreak occurs in one such county, then an outbreak will occur in the others. To define this geometrically, we first define for any county x the function $\rho(x) =$ the shortest distance from x to *any* county on the river. Then for any two counties x and y, define $d_r(x, y)$ to be either the Euclidean

distance between x and y, or the sum $\rho(x)+\rho(y)$, whichever is smaller. For example, referring again to Figure 1, the distance from county 2 to county 3 is still 100, denoted $d_r(2,3)$, but we also have $d_r(2,4) = d_r(2,19) = 100$ because counties 3, 4, and 19 are all on the river and are thus considered equivalent. It also follows that $d_r(2,25) = 200$ because the shortest distance is obtained by traveling 100 miles from counties 2 to 3, then up the river at no cost to county 24, then another 100 miles to county 25.

Notice that d_r is *not* a metric on the set of counties because, for example, $d_r(3,4) = 0$. It is a metric, however, if we consider the river counties to be one large "super-county." One way to picture this is shown in Figure 8.

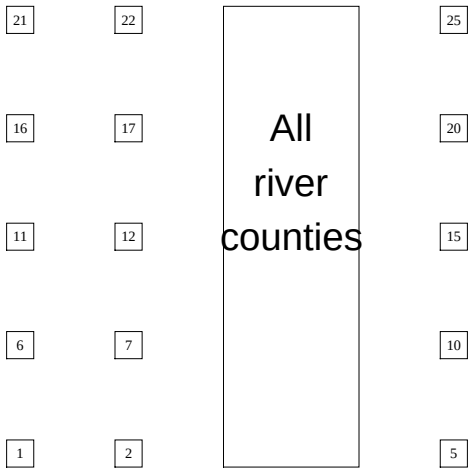


FIGURE 8. By "quotienting out" by the river, our distance measure treats all river-adjacent counties as equivalent.

EXERCISE 17. Sketch the disc of radius one in d_r about the counties 13 and 17.

EXERCISE 18. Determine if the distance function given by

$$d_{ave} = \text{average}\{d(x,y), \rho(x) + \rho(y)\}$$

is a metric.

DISCUSSION 7. What disadvantages might there be to using this approach in a real example?

EXERCISE 19. Let X be a non-empty finite set with metric d_X, let S be any subset of X, and denote by $X - S$ the set of all elements of X that are not in S. Define the *quotient metric* on $X - S$ by

$$d_q(u, v) = \min\{d_X(u, v), \min_{s,t \in S}\{d_X(u, s) + d_X(v, t)\}\}.$$

Show that d_q is a metric for $X - S$. What happens if $S = X$? What if S is empty? Explore some intermediate cases.

EXERCISE 20. Use the methods in this section to analyze a grid with the curved river depicted in Figure 9. Define the metric and sketch all possible shapes of clusters.

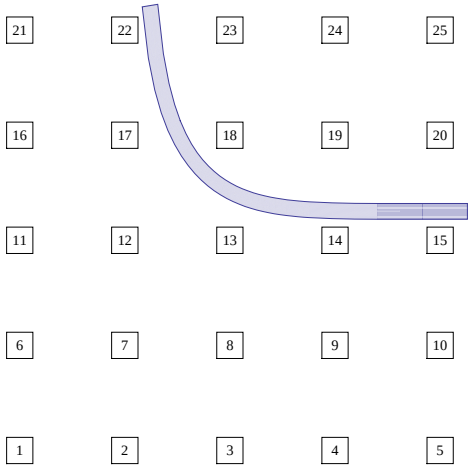


FIGURE 9. How can we account for a river with this shape?

PROJECT 1. Recall that we made many unrealistic assumptions in this model, for example, that population is uniformly distributed. Explore how the model will change with the following variations:

- Vary the distance between nodes.
- Allow discs to be centered at locations other than nodes.
- Consider models with multiple water sources.

PROJECT 2. (programming) Expand Exercise 1 to a more complete spatial scanner. You might try including some of the following features:

- Allow larger, arbitrarily shaped grids.
- Use search algorithms to speed the cluster search.

- Report the locations and likelihoods of several of the most likely clusters.
- Allow the user to choose from a variety of metrics, or even input his or her own.

PROJECT 3. Use publicly available software such as SatScan (see [K],[K2]) to reproduce as many of the examples in this module as you can. Then see what you can do on some real biosurveillance data!

PROJECT 4. In Section 3.3, we developed a metric that makes all river-adjacent counties equivalent. How would we approach the problem if we want to reflect the effect of the river on disease spread, but we do not want to go as far as to say all counties on the river are equivalent. Begin by considering a model in which the distances between two counties that are both on the river are halved (as opposed to zeroed).

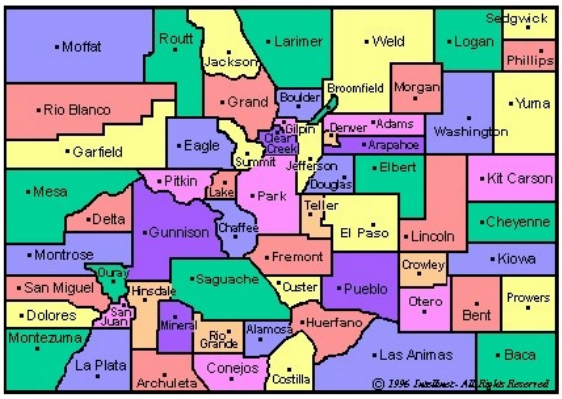


FIGURE 10. The counties of Colorado

PROJECT 5. Figure 10 a map of Colorado counties. Sketch the graph corresponding to this map. It is no longer a nice, regular grid. How does this impact the techniques developed in this module? Try sketching clusters in this grid, and then try on other states. (Optional: incorporate these grids into a spatial scanner, either one downloaded from the internet or one developed in the exercises in this module.)

PROJECT 6. Another approach is to allow the data to shape the clusters. Explore [TT], [TYT] for a discussion and implementation of this idea.

PROJECT 7. Have students choose an epidemic or outbreak and research it from identification to discovery of cause(s). Discuss other types of disease transmission and how identifying the mode of transport, i.e. water, air, etc., may help curb the outbreak.

4. Selected Solutions

Exercise 1: This is the cluster consisting only of the node on in the lower left-hand corner, as the nodes are distance 100 miles from each other. The likelihood for this cluster is 4.41.

Exercise 2: There are 39 possible clusters (including the whole map).

Exercise 3: The most likely cluster is the one that includes counties 3, 5, 6, and 9, so for example, the cluster centered at county 6 with radius 100. This cluster has likelihood 25.66. Observe that although counting only counties 3, 5, and 6 would yield a higher likelihood, using Euclidean discs forces us to pick up county 9 as well.

Exercise 5: We have that $2d(u,v) = d(u,v) + d(u,v)$
$= d(u,v) + d(v,u)$ by the symmetric property
$\geq d(u,u)$ by the triangle inequality
$=0$ by the identity of indiscernibles

Exercise 6: We show that the formula given by $d((x_1,y_1),(x_2,y_2)) = 10\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ forms a metric on $\mathbf{R}^2$.

(1) Identity of indiscernibles: Let $v = (x,y)$. Then $d_1(x,y) = 10\sqrt{(x-x)^2+(y-y)^2} = 0$.
Now suppose that $v = (x_1,y_1)$ and $w = (x_2,y_2)$, and that $d_1(v,w) = 10\sqrt{(x_1-x_2)^2+(y_1-y_2)^2} = 0$. Dividing by 10 and then squaring both sides yields $(x_1-x_2)^2+(y_1-y_2)^2 = 0$, which implies that $(x_1-x_2)^2 = 0$ and $(y_1-y_2)^2 = 0$ since these two terms sum to zero. Thus, $x_1-x_2 = 0$ and $y_1-y_2 = 0$, or $x_1 = x_2$ and $y_1 = y_2$ and thus $v = w$.

(2) Symmetric: Let v, w be as above. We then clearly see that $d_1(v,w) = 10\sqrt{(x_1-x_2)^2+(y_1-y_2)^2} = 10\sqrt{(x_2-x_1)^2+(y_2-y_1)^2} = d_1(w,v)$ and d_1 is symmetric.

(3) Triangle Inequality: Again let v, w be as above and let $u = (x_3, y_3)$. We have that $d_1(u, v) + d_1(v, w) =$
$= 10\sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2} + 10\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$
$= 10(\sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2} + \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2})$
$\geq 10(\sqrt{(x_3 - x_2)^2 + (y_3 - y_2)^2})$ by the triangle inequality on Euclidean distance
$= 10\sqrt{(x_3 - x_2)^2 + 10(y_3 - y_2)^2})$.

Exercise 7: Verify that $d_t((x_1, y_1), (x_2, y_2)) = |x_1 - x_2| + |y_1 - y_2|$ is indeed a metric.

(1) Identity of indiscernibles: Let $v = (x, y)$. Then $d_t(v, v) = |x - x| + |y - y| = 0$. Now let $v = (x_1, y_1)$ and $w = (x_2, y_2)$, and suppose that $d_t(v, w) = |x_1 - x_2| + |y_1 - y_2| = 0$. Since $|x_1 - x_2| \geq 0$ and $|y_1 - y_2| \geq 0$ and the two terms sum to 0, we must have that $|x_1 - x_2| = 0$ and $|y_1 - y_2| = 0$, or $x_1 = x_2$ and $y_1 = y_2$ and thus $v = w$.

(2) Symmetric: Let v, w be as above. Then $d_t(v, w) = |x_1 - x_2| + |y_1 - y_2| = |x_2 - x_1| + |y_2 - y_1| = d_t(w, v)$ and d_t is symmetric.

(3) Triangle Inequality: Again let v, w be as above and also let $u = (x_3, y_3)$. Then $d_t(u, v) + d_t(v, w) =$
$= |x_3 - x_1| + |y_3 - y_1| + |x_1 - x_2| + |y_1 - y_2|$
$\geq |x_3 - x_1 + x_1 - x_2| + |y_3 - y_1 + y_1 - y_2|$
$= |x_3 - x_2| + |y_3 - y_2|$
$= d_t(u, w)$

Exercise 9: Claim: A circle of radius zero is simply a point, specifically the point at which the circle is centered.
Proof: Let $C = \{(x, y) \in \mathbf{R}^2 : d((x, y), (x_0, y_0)) = 0\}$, where d is any metric. Clearly, the point (x_0, y_0) is in C since $d((x_0, y_0), (x_0, y_0)) = 0$ since d is a metric. Now let $(x, y) \in C$. Then $d((x, y), (x_0, y_0)) = 0$ and again since d is a metric, we must have that $(x, y) = (x_0, y_0)$. Thus, $C = \{(x_0, y_0\}$.

Exercise 10:
By the definition of a circle, we know that a circle in the taxicab metric of radius r and centered at the point (x_0, y_0) will be given by the set $C = \{(x, y) \in \mathbf{R}^2 : |x - x_0| + |y - y_0| = r\}$. This is a square rotated forty-five degrees from the coordinate axes. The disc contains

the interior of this square.

Exercise 12: Proof:

(1) Identity of indiscernibles: It is clear by definition that $d_D(x, y) = 0$ iff $x = y$.
(2) Symmetric: There are two cases. If $x = y$ then $d_D(x, y) = 0 = d_D(y, x)$. If $x \neq y$ then $d_D(x, y) = 1 = d_D(y, x)$.
(3) Triangle Inequality: Let x, y, z be elements in our set.
Case 1: Suppose that $x \neq y$, z is arbitrary. We have that $d_D(x, y) + d_D(y, z) = 1 + d_D(y.z) = 2$ if $y \neq z$ and $= 1$ if $y = z$. In either case, the result is greater than or equal to 1, which is the maximum magnitude of $d_D(x, z)$.
Case 2: Now suppose that $x = y$, z is arbitrary. We have that $d_D(x, y) + d_D(y, z) = 0 + d_D(y, z) = d_D(y, z)$. If $x = z$ then $d_D(x, z) = 0$ and the inequality holds. If $x \neq z$ then $d_D(x, z) = 1$ but also $y \neq z$ since $x = y$, so $d_D(y, z) = 1 = d_D(x, z)$ and the inequality again holds.

We describe the circles and discs in the metric d_D. Since d_D only takes on the values 0 and 1, our circles will either be of radius 0 (a point) or of radius 1 (circles of all other radii will be the empty set). The circle of radius 1 centered at the point x_0 will include all points that are distance 1 from x_0, or all points not equal to x_0. Thus, it will be the entire set (call it S) minus the center x_0. There are two possible (closed) discs centered at a given point x_0, given by $D_0 = \{x \in S : d_D(x, x_0) \leq 0\}$ and $D_1 = \{x \in S : d_D(x, x_0) \leq 1\}$. D_0 is again simply the point x_0 (observe that this is also the open disc of radius 1), while D_1 includes all points which are either 0 or 1 unit away from x_0. This includes both x_0 as well as all points not equal to x_0, so this is thus the whole set S.

Exercise 13: This is not a metric, as it violates both properties (1) and (3) (identity of indiscernibles and the triangle inequality). Property (1) clearly does not hold since $d(x, x) = 1$ for any x in our set S. For a specific counterexample to the triangle inequality, let $S = \{1, 2\}$. Then $d_D(1, 2) + d_D(2, 1) = 0 + 0 = 0 < 1 = d_D(1, 1)$.

Exercise 15: Proof:
(1) Identity of indiscernibles: Let $v = (x, y)$. Then $d_{ew}(v, v) = \sqrt{(x - x)^2 + 9(y - y)^2} = 0$. Now suppose that $v = (x_1, y_1), w = (x_2, y_2)$ and $d_{ew}(v, w) = 0$. We have that $0 = \sqrt{(x_1 - x_2)^2 + 9(y_1 - y_2)^2}$, and squaring both sides yields $0 = (x_1 - x_2)^2 + 9(y_1 - y_2)^2$. Since the two

terms sum to 0, we have that $(x_1 - x_2)^2 = 0$ and $9(y_1 - y_2)^2 = 0$, or $x_1 = x_2$ and $y_1 = y_2$ so $v = w$.

(2) Symmetric: Let v, w be as above. Then $d_{ew}(v, w) = \sqrt{(x_1 - x_2)^2 + 9(y_1 - y_2)^2}$ $= \sqrt{(x_2 - x_1)^2 + 9(y_2 - y_1)^2} = d_{ew}(w, v)$.

(3) Triangle Inequality: Again let v, w be as above and let $u = (x_3, y_3)$. Then $d_{ew}(u, v) + d_{ew}(v, w) =$
$= \sqrt{(x_3 - x_1)^2 + 9(y_3 - y_1)^2} + \sqrt{(x_1 - x_2)^2 + 9(y_1 - y_2)^2}$
$=$

To consider the circle of radius r centered at (x_0, y_0) in this metric, we let $r = d_{ew}((x, y), (x_0, y_0))$ which yields $r = \sqrt{(x - x_0)^2 + 9(y - y_0)^2}$. Putting this into the standard equation for an ellipse we obtain $r^2 = (x - x_0)^2 + 9(y - y_0)^2$ or $\frac{(x - x_0)^2}{r^2} + \frac{(y - y_0)^2}{\frac{r^2}{9}}$. This is exactly an ellipse centered at (x_0, y_0) whose vertical span is one-third that of its horizontal span.

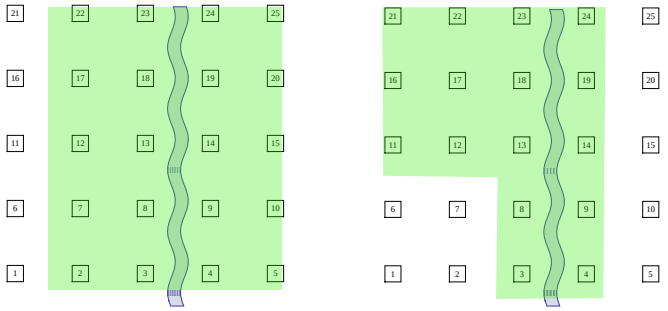
Exercise 16: We can change the vertical span by a factor of 4 rather than 3 by replacing the 9 with a 16. In general, to obtain ellipses of general but fixed radii we can simply fix the coefficients in front of the squared factors in the distance formula. To allow the algorithm to consider ellipses of any radii, we would need to consider all possible coefficients (we would of course need to limit this to an appropriate discrete set).

Exercise 17: See Figure 11

Exercise 18: This is not a metric, since for any x not located on the river, $d_{ave}(x, x) =$average$\{d(x, x), \rho(x) + \rho(x)\} =$average$\{0, 2\rho(x)\} = \rho(x) > 0$ since x is not on the river. Thus, d_{ave} is not a metric.

Exercise 19: Let $u, v \in X - S$.

(1) Identity of indiscernibles: $d_q(u, u) = 0$ is straightforward. Suppose $0 = d_q(u, v) = \min\{d_X(u, v), \min_{s,t \in S}\{d_X(u, s) + d_X(v, t)\}\}$. Then either $d_X(u, v) = 0$, in which case $u = v$, or there are $s, t \in S$ such that $d_X(u, s) + d_X(v, t) = 0$. Since both terms are non-negative, we have



(a) County 13 (b) County 17

FIGURE 11. Exercise: discs of radius 1

$d_X(u,s) = 0$ and $d_X(v,t) = 0$. Thus $u = s$ and $v = t$, contradicting $u, v \in X - S$.

(2) Symmetry is clear.

(3) Triangle inequality: let $u, v, w \in X - S$. We show $d_q(u,v) + d_q(v,w) \geq d_q(u,w)$. Without loss of generality, we have three cases. Case 1: $d_q(u,v) = d_X(u,v)$ and $d_q(v,w) = d_X(v,w)$, and the result follows from the triangle inequality on d_X. Case 2: $d_q(u,v) = d_X(u,v)$ and there exist $s,t \in S$ such that $d_q(v,w) = d_X(v,s) + d_X(w,t)$. Then $d_q(u,v) + d_q(v,w) = d_X(u,v) + d_X(v,s) + d_X(w,t) \geq d_X(u,s) + d_X(w,t) \geq \min\{d_X(u,w), \min_{s',t' \in S}\{d_X(v,s') + d_X(w,t')\}\} = d_q(u,w)$. Case 3: There exist $a, b, s, t \in S$ such that $d_q(u,v) + d_q(v,w) = d_X(u,a) + d_X(v,b) + d_X(v,s) + d_X(w,t) \geq d_X(u,a) + d_X(b,s) + d_X(w,t) \geq d_X(u,a) + d_X(w,t) \geq \min\{d_X(u,w), \min_{s',t' \in S}\{d_X(u,s') + d_X(w,t')\}\} = d_q(u,w)$.

REFERENCES

[B] Bespalchuk, P, et. al., Thyroid Cancer in Belarus after Chernobyl, International Congress Series 1299, 27-31 2007.

[C] Cramer, H. Mathematical Methods of Statistics, Princeton University Press, 1999.

[G] Goodman, J. E. and O'Rourke, Joseph, Handbook of Discrete and Computational Geometry , CRC-Press, 1997.

[Gr] Greenwald, J., Deadly Meltdown, Time Magazine, 1986.

[H] Hill, A. B. Snow-An Appreciation. Proceedings of the Royal Society of Medicine 48, 1008-1012, 1955.

[K] Kulldorff, M. A spatial scan statistic. Communications in Statistics: Theory and Methods, 26:1481-1496, 1997.

[K2] Kulldorff M. and Information Management Services, Inc. SaTScanTM v7.0: Software for the spatial and space-time scan statistics. http://www.satscan.org/, 2006.

[Ko] Koch, T. John Snow, hero of cholera. CMAJ 178(13), 1736, 2008.

[Kr] Krause, E. F., Taxicab Geometry: An Adventure in Non-Euclidean Geometry, Dover Publications, 1987.

[R] Rich, R. An ill wind from Chernobyl: The poisonous dust may have settled from the Chernobyl explosion of 1986, but whole communities, politicians and the nuclear industry accross Europe are still dealing with the fallout. New Scientist Print Edition 1765, 1991.

[S] Snow, J. On the mode of communication of cholera. 2nd Ed., London, J. Churchill, 1855.

[TT] Tango, T. and K. Takahashi, A flexibly shaped spatial scan statistic for detecting clusters, International Journal of Health Geographics 4:11, 2005.

[T] Thorton, J.L. John Snow, Pioneer Specialist-An Aesthetist. Anaesthesia. 5(3): 129–35, 1950.

[TYT] Takahashi, K., T. Yokoyama and T. Tango. FleXScan: Software for the Flexible Scan Statistic. National Institute of Public Health, Japan, 2005 - 2006.

E-mail address: frankoj2@scranton.edu

Department of Mathematics, The University of Scranton, Scranton, PA 18510, U.S.A.

E-mail address: schwellrac@ccsu.edu

Department of Mathematical Sciences, Central Connecticut State University, New Britain, CT 06050, U.S.A.

E-mail address: wood@hendrix.edu

Department of Mathematics and Computer Science, Hendrix College, Conway, AR 72032, U.S.A.