A simple graph theoretic method provides accurate range area estimates

John Alroy[1]

[1]Macquarie University

October 2, 2020

Abstract

Calculating spatial ranges of species and individuals is a crucial problem throughout ecology. However, sample size biases can be strong, and defining range boundaries can be difficult. These hurdles can be overcome by calculating areas without calculating boundaries. The first step is to algorithmically define a graph that connects the spatial points where observations have been made. The routine generates a small number of short edges that form a pattern resembling a mosaic. The edge lengths are summed, squared, divided by the edge count, and multiplied by a known constant to obtain a total area estimate for the shape. This non-parametric mosaic area method can work with irregular outlines and clumped point distributions. It is more accurate than convex hull, kernel density, and hypervolume estimation according to simulation analyses. Mosaic area calculations can be used in areas ranging all the way from conservation biology to morphometrics.

Abstract

Calculating spatial ranges of species and individuals is a crucial problem throughout ecology. However, sample size biases can be strong, and defining range boundaries can be difficult. These hurdles can be overcome by calculating areas without calculating boundaries. The first step is to algorithmically define a graph that connects the spatial points where observations have been made. The routine generates a small number of short edges that form a pattern resembling a mosaic. The edge lengths are summed, squared, divided by the edge count, and multiplied by a known constant to obtain a total area estimate for the shape. This non-parametric mosaic area method can work with irregular outlines and clumped point distributions. It is more accurate than convex hull, kernel density, and hypervolume estimation according to simulation analyses. Mosaic area calculations can be used in areas ranging all the way from conservation biology to morphometrics.

INTRODUCTION

One of the most fundamental problems in theoretical ecology is estimating the extent of a shape in two-dimensional space from point data. Two categories of data are relevant: occurrences of species and of individuals.

Species ranges are important at large scales because geographic range patterns are a bedrock of biogeography and macroecology, telling us about such things as provincialism (Kreft & Jetz 2010) and latitudinal diversity gradients (Lawrence & Fraser 2020). Estimating ranges based on expert opinion, species distribution modelling, or otherwise is of great importance in conservation biology (Maréchaux *et al.* 2017).

At the scale of individuals, home ranges have been studied intensively by wildlife biologists for decades (Burt 1943). The availability of large data sets derived from GPS technology calls the value of the concept into question (Kie *et al.* 2010), but interspecific comparisons of home range data are of such broad interest that this information remains relevant. For example, the allometry of home range size is a classical topic in macroecology (Kelt 2001).

Shape areas also come up in the field of niche modelling, which addresses high-dimensional spaces in addition to two-dimensional spaces (Blonder *et al.* 2014; Junker *et al.* 2016; Qiao *et al.* 2016). Additionally, the field of multivariate morphometrics is relevant: estimating the area of occupancy of a morphospace by points representing species or individuals is fundamentally the same problem. It has often been tackled in the past by computing statistics that are not explicitly spatial, such as mean pairwise distances (Foote 1991), because high-dimensional spaces are often considered. However, the connection is clear.

The full list of subjects that rely on area estimation is presumably much larger. Given the breadth and depth of interest in the topic, it comes as no surprise that a plethora of methods has been proposed. The most simple is to grid observations and count occupied squares. Gridded data have been used extensively and for many years in macroecology (Simpson 1964). Under the name ”area of occupancy”, they are still used for threat status evaluation (IUCN Standards and Petitions Committee 2019) by the International Union for the Conservation of Nature (IUCN). This approach is not without merits, because occupancy can be used to estimate population size (He 2012). However, the values are scale-dependent, and gridding will underestimate if sampling is sparse relative to the scale of interest (Hartley & Kunin 2003).

Another simple alternative is to compute a convex hull around the observations, i.e., to create a minimum convex polygon, which was a popular approach in wildlife biology for many years (Hayne 1949). Convex hulls also tend to underestimate, although they will overestimate if there are holes in distributions or if there are large outliers. But likewise, the IUCN continues to use this method for determining the ”extent of occupation” of a species, a second major criterion for threat status evaluation (IUCN Standards and Petitions Committee 2019). Indeed, both approaches are still considered to be central by conservation biology researchers, not just the IUCN (Smith *et al.* 2020).

Nonetheless, field-based ecologists are strongly cognizant of bias in convex hull areas, so alternatives such as kernel density estimation have long been commonplace in that area (Worton 1989). The IUCN guidelines mention this approach only in passing (IUCN Standards and Petitions Committee 2019). A hybrid method called local convex hull nonparametric kernel estimation also is used by wildlife biologists, but its performance has been questioned (Lichti & Swihart 2011).

There are many methods other than kernel density estimation, some quite sophisticated. Recently, for example, computation of hypervolumes (Blonder *et al.* 2014) has become popular with niche modellers. This method assumes the data are bivariate normal or elliptical in their distribution, which is problematic and has been critiqued (Qiao *et al.* 2016), and which some researchers have tried to address (Jarvis*et al.* 2019). However, the method's popularity earns it serious attention. Meanwhile, palaeobiologists have used other methods such as computing maximum great circle distances. This approach makes sense when the data follow a linear trend (Foote *et al.* 2008), but it has the drawback of putting aside most of the data points.

In any event, many existing approaches have three major flaws addressed with the new method proposed here. First, they can systematically underestimate or overestimate, depending on their properties. Consistent accuracy is a rare property. Second, they may not be particularly accurate when the data points form clumped or irregular patterns. Finally, methods that depend on a series of flexible options and parameters yield results that are indecisive and therefore not very interpretable.

As I will explain, all three problems can be solved by creating a network of points that resembles a mosaic and using the edge lengths to obtain an area estimate. This method, which has been implemented in an R package called *mosaic* , has a variety of additional applications. For example, areas of overlap between ranges are directly computable, areas of multi-dimensional shapes can be approximated, and the method allows for identifying outliers by breaking long edges.

Ecologists have used graph theory in the past, but only when working on selected topics such as landscape analysis (Foltête *et al.* 2020). The method outlined here is unrelated to any of this work. For example, existing methods that concern area estimation are founded on entirely different theory (Keith, Spring & Kompas 2019).

Before detailing the new approach, it is important to mention what this paper is and is not about. The goal is to estimate range area, not range shape. However, mosaic patterns are more intuitive approximations of range shapes than are convex hulls because they need not be convex. More importantly, range area *per se* is of central concern to biogeographers, macroecologists, allometricians, niche modellers, and even the IUCN. Second, this not a comparative benchmarking analysis. The only goal is to show that the method performs well, not to definitively prove that it outperforms every proposed alternative. Thus, comparisons will be limited to three things that are of general interest: convex hulls, kernel density estimators, and hypervolumes. Finally, many readers will have come to expect that every paper on range estimation methods will be graced with many equations and framed in terms of complex and most often parametric process models. This is not one of them. Instead, I will argue that a simple method should be taken seriously because it makes sense and it works.

MATERIAL AND METHODS

Requisite graph theory

The lengths of the edges in a mosaic graph M are used to obtain a single area estimate. Specifically, the procedure is to sum the lengths, square the sum, divide by the number of edges e to obtain a value notated $L_M\ ^2/e$, and then multiply the value by a constant derived below. The equation is implemented by the *mosaic* function *mgraph* .

A minimum spanning tree (MST) could be used for the same purpose. I discuss this idea in more detail later. MSTs are well-known and can be computed by using any of several algorithms, including a classic one proposed by Kruskal (1956). Shape area does scale tightly to the proposed function of the MST's edge lengths. For example, suppose$p$ points (= vertices) are perfectly spaced on a square grid with the grid squares having edges of length a , yielding $pa\ ^2$ as the shape's area. Keeping in mind that there are always $p - 1$ edges in an MST, the sum is $(p- 1)\ a$; the square is $(p - 1)^2 a\ ^2$; and that divided by p is close to$p\ a\ ^2$ no matter what the MST's shape.

However, real-world patterns are not grids. So in order to generalise, we need to switch from regular patterns to arbitrary distributions and discuss the theory of mosaic graphs. In a mosaic, (1) each point is connected to at least two others, and (2) no two points remain directly connected if some other point connects to both of them. Isolated triangles at the edges of the mosaic are allowed by this rule. However, on average, each empty loop (mosaic piece) is an octagon, there are four edges per piece, and there are five edges for every four points.

To understand why, note that adding any point inside an n -gon making up a mosaic piece either splits it or increases its edge count by one (Fig. 1). Suppose A is new, with nearest neighbours B and C. If B and C are adjacent, the A–B and A–C edges will be retained but the B–C edge will be discarded, increasing the mosaic piece's size. If not, then the B–A–C edges will form a wall between two new pieces. Any larger pattern will have a greater risk of splitting. Thus, the growth and splitting processes will always push mosaic pieces towards a point count that happens to be balanced around eight (caption of Fig. 1).

Whenever a piece grows, a double junction forms at point A and an edge is removed. A double junction is still created in the equally likely splitting case, but the junctions at B and C each add one edge. On average, then, a new point creates one double junction and increases the edge count (degree) of one other point. The degree of a point rarely goes past three because any new point close to a triple junction will likely pair with two of the three edges leading to it, resulting in the loss of one edge.

Thus, on average, half of the points in a mosaic are of order two and half of order three. If the order is two there is one edge per point overall, as in a simple loop, and if three, there are 1.5 per point. Therefore, the ratio of edges to points is 1.25:1, or five to four, and of points to edges, 0.8. Because every edge inside a mosaic is shared by two pieces by definition, there are four edges per piece on average. Thus, there are 3.2

points per piece. These predictions are easily confirmed by simulation using the mosaic assembly algorithm described below.

Area estimation

So, how does all of this relate to the square of the sum of lengths divided by the edge count,L_M $^2/e$, and the estimate of a shape's overall area, A ? Suppose that the average mosaic piece resembled a square, not an octagon, but also with a perimeter eight times the average edge length a . Each side would be $2a$ long and the area would be 4 a ^{2}. The overall area across the graph would therefore be the piece count times $4a$ ^{2}. In this limited case, A is justL_M $^2/e$ because there are four edges per 2 x 2 rectangle on average: given 25 rectangles, the area is 100 a ^{2}; $e = 4$ x $25 = 100$;$L_M = 4$ x 25 $a = e$ $a = 100a$; and L_M $^2/e = 100$ a $^2 = A$.

Because the pieces actually average out to octagons, it might seem that the area of each one would be the area of a regular octagon, which is 2 $(1 + 2^{0.5})$ a $^2 = 4.828a$ ^{2}. Thus, we might estimate A as $4.828/4L_M$ $^2/e = 1.207L_M$ $^2/e$. However, the maximal size of any polygon is reached when it expands in all directions to become regular (because it most closely approximates a circle). No matter what the construction process, polygons subject to any kind of randomness must be smaller. Thus, the 4.828 figure may be too high.

Nonetheless, simulations provide no evidence to support this hypothesis. A good explanation is that the average edge in a mosaic abuts a larger-than-average piece by definition. For example, if half of the mosaic consists of 6-edge pieces and half of 10-edge pieces, the average edge abuts a shape of $(6^2 + 10^2)/16 = 8.5$ edges, not eight. The larger a piece, the more closely it approximates a circle, the shape having the lowest perimeter-to-area ratio: a square with a perimeter of eight has an area of four, whereas a circle with the same perimeter (circumference) C has an area of$C$ $^2/(4$ pi$) = 5.093$. This effect seems to cancel out the overestimation due to irregularity in polygon shapes, and as a result, throughout the rest of this paper I employ the $4.828/4 = 1.207$ regularisation constant.

Turning briefly to MSTs, which can be computed using the *mosaic*package's *tgraph* function, each includes about 4/5 as many edges as a mosaic because the edge:point ratio is nearly one in a large MST. However, an MST's total length should be less than 4/5 of the corresponding mosaic's length because an MST should avoid many long edges. Perhaps, the MST on average simply avoids the longest out every five edges. It can do this because there are four points for every five edges in the mosaic (see above), and there is a near 1:1 ratio of edges to points in an MST. However, the choice may come down to only two edges because the others can't be avoided: if the points form a line, the MST must either cross from the left and leave out the last edge or vice versa. The longer segment when a line is subdivided at random comprises 3/4 of the length on average, so the MST's length should be $(3 + 1/4)/5 = 65\%$ that of the mosaic's. Thus, if L_T is the length of the MST, then instead of $A = 1.207L_M$ $^2/e$ we would predict$A = 1.207/0.65^2L_T$ $^2/e = 2.857L_T$ $^2/e$. However, in practice, MST-based area estimates are highly problematic because the 0.65 constant seems to vary in simulation according to the shape of the object: for example, it is higher for circles and rings, and actually close to 0.8. Therefore, an MST-based approach is not recommended.

Mosaic algorithm

The divisive algorithm implemented by the *mosaic* package's function *mgraph* produces mosaics with small sums of edge lengths, and is as follows. (1) All points (= vertices) are connected to all other points. (2) The edges are ordered from longest to shortest and inspected in turn. (3) If (a) the two connected points i and$j$ are both connected to a third point and if (b) i and$j$ are each connected to at least three points in total, then the edge is broken. If not, then it is kept.

For example, suppose there is a triangle. No edges can be broken because no point is connected to more than two others. If instead there are four points, at first each of them is connected to three others, so the longest edge (say, between points 1 and 4) is examined and discarded. Points 2 and 3 are now still triple junctions. Furthermore, each connects to the other via both 1 and 4. Therefore, the edge between 2 and 3 is also broken, resulting in a quadrilateral.

Two examples of mosaics produced by this algorithm are shown in Figs. 2A and B. As predicted, the number of edges connecting to each point is most often two, frequently three, occasionally four, and very rarely five or more. Lines very rarely cross. The algorithm occasionally creates a line at the outline of the overall shape that connects two pieces instead of belonging to a piece itself. There happens to be an example at the lower right of Fig. 2A. The method handles internal gaps in ranges well specifically because it rarely draws an edge across a gap, as long as there are enough surrounding points to complete a short circuit (Fig. 2B). Drawing squares around the edges makes it easier to visualise the contribution of each edge to the overall area estimate (Figs. 2C, D).

Details of implementation

The mosaic algorithm requires O(p ^{3}) comparisons where p is the point (= vertex) count. It can be speeded up to O(p ^{2}) by only examining edges connecting nearby pairs. Specifically, if i and j are the endpoints, then if$i$ is one of the 20-odd nearest neighbours of j , the edge should be examined; and vice versa. The reason for the cutoff of 20 is that in the final mosaic, no matter how computed, edges are usually short and points are always sparsely connected (Fig. 2). This algorithm will skip an edge if there are two or more large and tight clusters of points each having more points than the cutoff, in which case the user needs to decide whether a higher one should be imposed. The neighbours-only algorithm speeds up the calculations so much that a mosaic of 400 points can be arranged within about a second on a laptop computer. A simulation producing 1000 sets of 20-point mosaics takes about four seconds.

Outliers and long edges can of course inflate area estimates. A good, simple means of handling this problem is to exploit the preceding algorithm. Instead of only examining edges if either point is in the neighbourhood of the other point, one can require that each point is in the other's neighbourhood. This "mutual neighbour" criterion leads to removing edges that go to individual outliers or small clusters of outliers, in addition to long edges between large groups of outliers. It is used as the default in the analyses reported here.

Although computing a high-dimensional mosaic graph is trivial, the multiplication and summation procedure only yields a sensible estimate if there are two dimensions. A good solution is to compute mosaic areas for all pairs of dimensions; multiply them; and raise the value to the $1/P$ power where P is the number of pairs. For example, in two dimensions the power is 1, and in three it is 1/3 because there are three pairs. This is analogous to projecting a high-dimensional shape on to each side of a hypercube, averaging the projected areas, and using that as a proxy for the shape's hypervolume. Although irrelevant to most ecological problems, this potential implementation makes the mosaic approach more useful in trait space analysis and morphometrics.

Mosaic hulls

As mentioned, mosaics can be illustrated simply by drawing the graphs (Figs. 2A, B), with the addition of squares around the edges helping to understand how the area calculation works (Figs. 2C, D). However, these plots are not as intuitive as convex hulls, which are simple filled polygons – and are misleading when shapes are actually convex, which is routinely true of large, real-world data sets. Computing mosaic graphs makes it possible to replace convex hulls with hulls that allow for convexity. The procedure, which is used by the *mhull* function in the *mosaic* package, starts by choosing the most extreme point in one direction along one axis or the other, and by recording which point is to the immediate left of this one. The rest of the algorithm is as follows. Suppose that the last-visited point is A, the current point is B, and B is connected by an edge to A, C, and D. (1) Points like C and D, but not A, are examined. (2) The angles between B and neighbours like C and D are computed. (3) The points are ordered relative to the angle of a line going back from B to A, and the next one to the right (say C) is selected. (4) Step 1 is revisited (so B is replaced with C). (5) The algorithm terminates when the first point is reached again, but only on the occasion that it is reached from the point to its left.

Kernel density, convex hull, and hypervolume estimates

Two-dimensional kernel density estimates (KDEs) were generated using the*kde2d* function in R's *MASS* library (Venables & Ripley 2002). Default settings were used, so the smoothing kernel was Gaussian and the bandwidth was determined by the normal reference method. Note that a large number of kernel functions exist as alternatives, as with bandwidth methods, and that these choices have a non-trivial influence on the results. However, the point here is to see what would happen if a researcher took the path of least resistance and used the most popular approach.

By convention, KDEs are used by wildlife ecologists to estimate home ranges by taking 95% confidence intervals (Powell & Mitchell 2012). Likewise, the "core area" of a range is defined as the 50% confidence interval. The 95% criterion is arbitrary and has come under criticism (Powell & Mitchell 2012). However, 95% KDEs will be considered in this paper because they are so widely used in the literature.

Convex hull areas were computed using the *chull* function in the*grDevices* R package (R Core Team 2020) plus the *st_area*function in the *sf* package (Pebesma 2018), after preparing the hulls using *st_polygon* and *st_sfc* .

Hypervolumes were calculated using the *hypervolume_gaussian*function in the R library *hypervolume* (Blonder 2019). Again, default settings were used. This meant using Gaussian kernel density estimation, a prespecified formula for calculating the number of random samples per point, a default method for estimating the KDE bandwidth, a standard deviation count of three (which is important for fixing the actual area), and a fixed quantile value of 0.95 (likewise). As with other complex methods, the ability to vary numerous flexible parameters clouds the outcome. The mosaic area method has no flexible parameters.

Data

The American black bear *Ursus americanus* was arbitrarily chosen to illustrate the method because it was expected to have a large, well-sampled range. Data were drawn from the Global Biodiversity Information Facility (GBIF) using the *occ_data* function in R library *rgbif* (Chamberlain et al. 2020) on 2 September 2020. After processing the data with the *st_polygon* and *st_sfc*functions in the R library *sf* (Pebesma 2018), convex hull areas were computed using *st_area* .

Monte Carlo simulations were used to compare the four methods. Data were placed randomly within ranges having simple geometrical shapes. Four sets of simulations were implemented. (1) Comparisons of all four methods using five points randomly drawn from within a circle in each trial. (2) Comparisons with 20 points drawn. (3) Comparisons with 20 points drawn from a circle and 80% of the data points in the right-hand side of the shape randomly removed. (4) Trials using six different shapes, with 10 points being drawn during each trial and only the mosaic area method being applied in most cases.

RESULTS

Empirical data

There are large holes in the geographic range of the American black bear, such as the Great Basin of the western United States and a large part of eastern Canada (Fig. 3A). Although the mosaic hull fills in some gaps (Fig. 3A), it emphasises these two. Squares drawn around the mosaic lines also illustrate them well (Fig. 3B). By contrast, the IUCN website depicts a solid historical range extending throughout almost the entirety of North America (https://www.iucnredlist.org/species/41687/114251609).

Mosaic area estimates for individual five-year historical intervals show some random scatter, but no definite trend, and good consistency with the overall estimate of 602.8 equal-area degree cells (Fig. 4). By contrast, the overall estimate based on a convex hull is 1853.9 degree cells, not only much higher than most of the individual mosaic areas but much higher than convex hull areas for the same intervals. These patterns reinforce the point that the hull areas are doubly biased: they are too low when a data set is small (curve vs. upper line), but they are too high when a distribution has gaps (upper line).

Simulated data

Mosaic areas are already centred on actual range sizes when sample sizes are very small (five points per circle: Fig. 5A). Convex hulls consistently underestimate by a large margin, as expected. Less intuitively, the remaining two methods consistently overestimate. Based on the r^2 values (caption of Fig. 5), hypervolumes and mosaic areas are similarly precise. Thus, the issue is accuracy instead of precision.

Twenty data points per trial (Fig. 5B) is still a very low figure because it has long been recommended that at least 50 data points should be used to fix home ranges (Seaman *et al.* 1999). Here, the mosaic area values are still the only ones centred on the line of unity. Specifically, the median of ratios taken against known values is 0.95. The other three methods all fail. The 95% KDE and hypervolume estimates are still too high, with median ratios of 1.97 and 1.54. As expected, convex hull areas are biased in the opposite direction, with a median ratio of 0.60. The best one could say for these three methods is that their biases do not reverse as sample size increases.

Note that 95% KDEs are no more accurate than anything else when the sample size is five (caption of Fig. 5A) and are not very close to mosaic areas ($r^2 = 0.8568$ for KDEs vs. mosaic areas). These facts call 95% KDEs into question: they have no particular justification (Powell & Mitchell 2012), they are too high (Fig. 5), and they are not highly replicable using the best method discussed in this paper.

Spatial clustering of the data (Fig. 5C) biases the mosaic area values only weakly (median estimate:known area ratio 0.80), causes convex hull areas to fall short almost by the entire 50% that is possible (ratio 0.52), and also lowers the values for 95% KDEs and hypervolumes. However, they are still overestimates (1.41 and 1.27).

Mosaic areas also can handle a variety of range shapes even when only 10 points are sampled (Fig. 6). Median ratios of estimated to known areas are not far from one for most shapes: circles (1.00), squares (1.03), rectangles (1.17), and three-quarter rectangles (1.19). Results are worse for pairs of squares (2.06) and particularly rings (2.14). The first figure is philosophically problematic because it is hard to say whether two nearby clumps really should be considered separate shapes. If not, then 2.06 may be a reasonable compromise. With respect to rings, each one excludes half the area of the enclosing circle, so the approximate 2:1 ratio means that the method essentially treats rings as circles at this very low sampling level (if not at high levels: Figs. 1B, D). By contrast, ring areas are dramatically overestimated by 95% KDEs (6.12) and hypervolumes (4.75). These patterns are not illustrated because the ratios speak for themselves (and to save space). Again, shape solidity is a widespread assumption that is important for some methods, but not so much for the new one.

In general, the high performance of mosaic area estimation given this broad array of shapes is perhaps not too surprising because the underlying logic assumes that any shape can be covered adequately and accurately by a series of circuits connecting points, which stands to reason. The surprise is that reasonable results can be obtained with very small data sets.

DISCUSSION

Quantifying range areas is not only important across ecology and evolutionary biology, but societally important. Specifically, because the IUCN uses this information to determine the status of threatened species (IUCN Standards and Petitions Committee 2019), the issue of accuracy is no small matter. In light of this fact, one would hope that the IUCN would reconsider its focus on simple extents and areas of occupation and also embrace more current methods of area inference.

On this point, mosaic area estimation has several crucial advantages. Unlike convex hulls, which are the basis of the extent of occupation criterion, mosaics are independent of sample size and have a built-in routine for handing outliers. Unlike counts of occupied cells, which are interpreted as areas of occupation, they are independent of scale in addition to sample size. Unlike KDE and hypervolume calculations, mosaic areas are not upwards biased when shapes are solid, even when sample sizes are small (Fig. 5A). Unlike those methods, mosaic calculation has no flexible parameters and assumes nothing about the underlying shape of the distribution. And unlike all the other methods discussed here, mosaic area computation is explicitly formulated to handle the problem of irregular and non-random point distributions, with even strong patterning having little effect (Figs. 5C, 6). Autocorrelation is a major concern in this field (Noonan *et al.* 2019).

Much more needs to be done with range area estimation. For one thing, more in-depth testing of a broader range of methods would be desirable. Papers proposing and testing methods, especially those related to KDEs, are numerous, and there is no space even to summarise them: see Walter*et al.* (2015), Junker *et al.* (2016), Qiao *et al.*(2016), Jarvis *et al.* (2019), and Noonan *et al.* (2019) for recent examples. I put forth, however, that based on the current results, even the more complex methods are unlikely to outperform mosaic area estimation by a large margin. For this hypothesis to be disproven, conventional 95% KDEs and hypervolumes would have to be shown to be quite poor estimators because they are already substantially worse than mosaic areas. If there really is a much better parametric method, then the most likely candidate might be another kernel density estimator of some kind (Noonan *et al.* 2019).

Another possibility is that a better mosaic-related method might be found. For example, perhaps one could allow for denser connectivity of points or more complex weighting of edges in area calculations. Also, the algorithm for selecting edges might perhaps be further optimised without imposing a heavy computational burden. Advantages of altering the graph theory are unclear, and even if possible, further optimisation may not be particularly helpful.

Finally, mosaic area estimation is fundamentally non-parametric and depends on deduction from fundamental graph theory and geometry. Some will see this as a disadvantage. This matter touches on a deep paradigm conflict in statistics that concerns a simple question: should every method be model-based and fall within the domain of maximum likelihood or Bayesianism? Strong assumptions and flexible options come with model-based methods, and full objectivity comes with this one – in addition to high performance. Thus, in a field bursting with methods of many kinds, there may be room for a different approach to the deep problem of determining the areas of unknown shapes.

ACKNOWLEDGMENTS

I thank Roger Close, Daniel Falster, Michael Foote, Michael Gillings, Matthew Kosnik, and Anikó Tóth for helpful suggestions.

REFERENCES

Blonder, B. (2019). hypervolume: high dimensional geometry and set operations using kernel density estimation, support vector machines, and convex hulls. R package version 2.0.12.

Blonder, B., Lamanna, C., Violle, C., & Enquist, B. J. (2014). The n-dimensional hypervolume. *Glob. Ecol. Biogeogr.,* 23, 595-609.

Burt, W. H. (1943). Territoriality and home range concepts as applied to mammals. *J. Mammal.,* 24, 346-352.

Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L., & Ram, K. (2020). *rgbif: interface to the Global Biodiversity Information Facility API* . R package version 3.2.0.

Foltête, J.-C., Savary, P., Clauzel, C., Bourgeois, M., Girardet, X., Sahraoui, Y. *et al.* (2020). Coupling landscape graph modeling and biological data: a review. *Landscape Ecol.,* 35, 1035-1052.

Foote, M. (1991). Morphologic patterns of diversification: examples from trilobites. *Palaeontology,* 34, 461-485.

Foote, M., Crampton, J. S., Beu, A. G., & Cooper, R. A. (2008). On the bidirectional relationship between geographic range and taxonomic duration. *Paleobiology,* 34, 421-433.

Hartley, S., & Kunin, W. E. (2003). Scale dependency of rarity, extinction risk, and conservation priority. *Conserv. Biol.,* 17, 1559-1570.

He, F. (2012). Area-based assessment of extinction risk. *Ecology,*93, 974-980.

IUCN Standards and Petitions Committee (2019). Guidelines for using the IUCN Red List categories and criteria. Version 14. Prepared by the Standards and Petitions Committee. Downloadable from http://www.iucnredlist.org/documents/RedListGuidelines.pdf.

Jarvis, S. G., Henrys, P. A., Keith, A. M., Mackay, E., Ward, S. E., & Smart, S. M. (2019). Model-based hypervolumes for complex ecological data. *Ecology,* 100, e02676.

Junker, R. B., Kuppler, J., Bathke, A. C., Schreyer, M. L., & Trutschnig, W. (2016). Dynamic range boxes – a robust nonparametric approach to quantify size and overlap of n -dimensional hypervolumes. *Meth. Ecol. Evol.,* 7, 1503-1513.

Keith, J. M., Spring, D., & Kompas, T. (2019). Delimiting a species' geographic range using posterior sampling and computational geometry.*Sci. Rep.,* 9, 8938.

Kelt, D. A., & Van Vuren, D. H. (2001). The ecology and macroecology of mammalian home range area. *Am. Nat.,* 157, 637-645.

Kie, J. G., Matthiopoulos, J., Fieberg, J., Powell, R. A., Cagnacci, F., Mitchell, M. S. *et al.* (2010). The home-range concept: are traditional estimators still relevant with modern telemetry technology?*Phil. Trans. R. Soc. Lond. B Biol. Sci.,* 365, 2221-2231.

Kreft, H., & Jetz., W. (2010). A framework for delineating biogeographical regions based on species distributions. *J. Biogeogr.,* 37, 2029-2053.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the travelling salesman problem. *Proc. Am. Math. Soc.,* 7, 48-50

Lawrence, E. R., & Fraser, D. J. (2020). Latitudinal biodiversity gradients at three levels: linking species richness, population richness and genetic diversity. *Glob. Ecol. Biogeogr.,* 29, 770-788.

Lichti, N. I., & Swihart, R. K. (2011). Estimating utilization distributions with kernel versus local convex hull methods. *J. Wildl. Manage.,* 75, 413-422.

Marechaux, I., Rodrigues, A. S. L., & Charpentier, A. (2016). The value of coarse species range maps to inform local biodiversity conservation in a global context. *Ecography,* 40, 1166-1176.

Noonan, M. J., Tucker, M. A., Fleming, C. H., Akre, T. S., Alberts, S. C., Ali, A. H. *et al.* (2019). Analysis of autocorrelation and bias in home range estimation. *Ecol. Monogr.,* 89, e01344.

Pebesma, E. (2018). Simple features for R: standardized support for spatial vector data. *R Journal* 10, 439-446.

Qiao, H., Escobar, L. E., Saupe, E. E., Ji, L., & Soberon, J. (2016). A cautionary note on the use of hypervolume kernel density estimators in ecological niche modelling. *Glob. Ecol. Biogeogr.,* 26, 1066-1070.

Powell, R. A., & Mitchell, M. S. (2012). What is a home range? *J. Mammal.* , 93, 948-958.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Seaman, D. E., Millspaugh, J. J., Kernohan, B. J., Brundige, G. C., Raedeke, K. J., & Gitzen, R. A. (1999). Effects of sample size on KERNEL home range estimates. *J. Wildl. Manage.,* 63, 739-747.

Simpson, G. G. (1964). Species density of North American Recent mammals.*Syst. Zool.,* 13, 57-73.

Smith, J. A., Benson, A. L., Chen, Y., Yamada, S. A., & Mims, M. C. (2020). The power, potential, and pitfalls of open access biodiversity data in range size assessments: lessons from the fishes. *Ecol. Ind.* , 110, 105896.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S. Fourth Edition.* Springer, New York.

Walter, W. D., Onorato, D. P., & Fischer, J. W. (2015). Is there a single best estimator? Selection of home range estimators using area-under-the-curve. *Movement Ecol.,* 3, 10.

Worton, B. J. (1989). Kernel methods for estimating the utilization distribution in home-range studies. *Ecology,* 70, 164-168.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D. et al. (2019). CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. *Meth. Ecol. Evol.,* 10, 1-7.

Hosted file

`image1.emf` available at https://authorea.com/users/363764/articles/484477-a-simple-graph-theoretic-method-provides-accurate-range-area-estimates

Figure 1 Potential changes to mosaics resulting from the addition of a point. The point must link to two others, and it must be closer to half the points than the other because it cannot be at the exact centre of the shape. If the other two points are neighbours, their edge is dropped and the mosaic remains intact, growing by one point. Otherwise, it splits. In (A) and (B), the mosaic must grow because the new point must join neighbours. In (C), the chance is 2/3 if the point lands on one side and certain if it lands on the other. In (D), there is a 2/3 chance. In (E), there is either a 1/2 or 2/3 chance. In (F), there is a 1/2 chance. With more points, this chance falls below 1/2. Thus, growth and splitting pushes the point count towards eight.

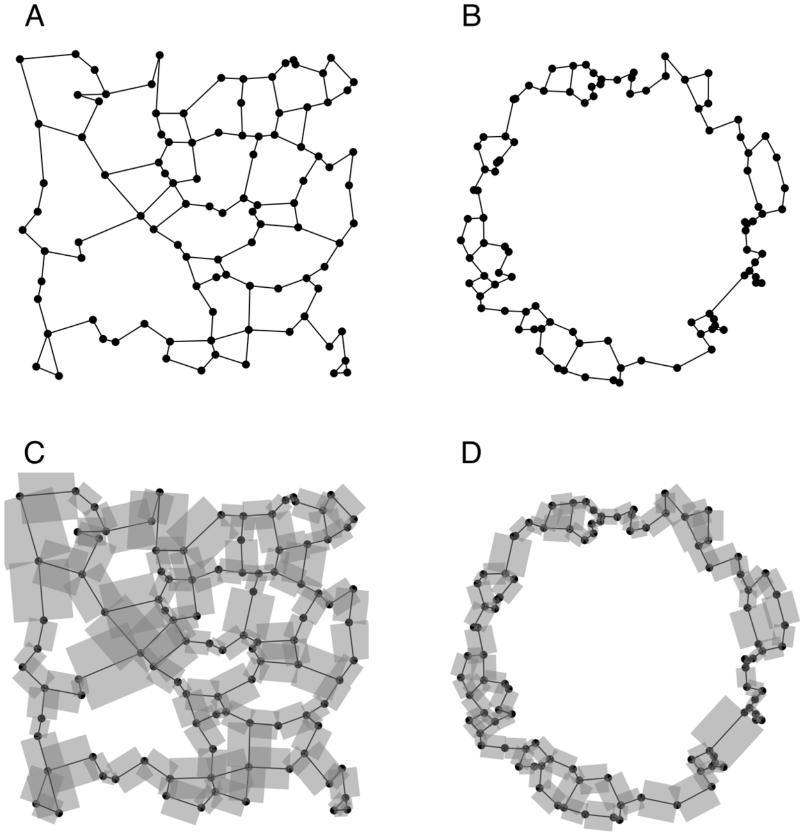

Figure 2 Mosaics of randomly located points within shapes having diameters of 1. There are 100 in each panel. (A) and (B) show the raw mosaic graphs; (C) and (D) include rectangles that indicate the contribution

of each edge to the area estimate. (A) and (C): The shape is a square. There are 56 points with two edges, 38 with three, and six with four. There are 26 mosaic pieces, so the piece/point ratio is 26/100, close to the 0.25 prediction for a compact shape. The true area is 1 and the estimate is 1.098. The median over 1000 trials is 0.975. (B) and (D): The shape is a ring. There are 74 points with two edges and 26 with three. The piece/point ratio is 14/100, which is low because the pattern is elongate. The true area is pi/8 = 0.393 and the estimate is 0.373. The median over 1000 trials is 0.370.

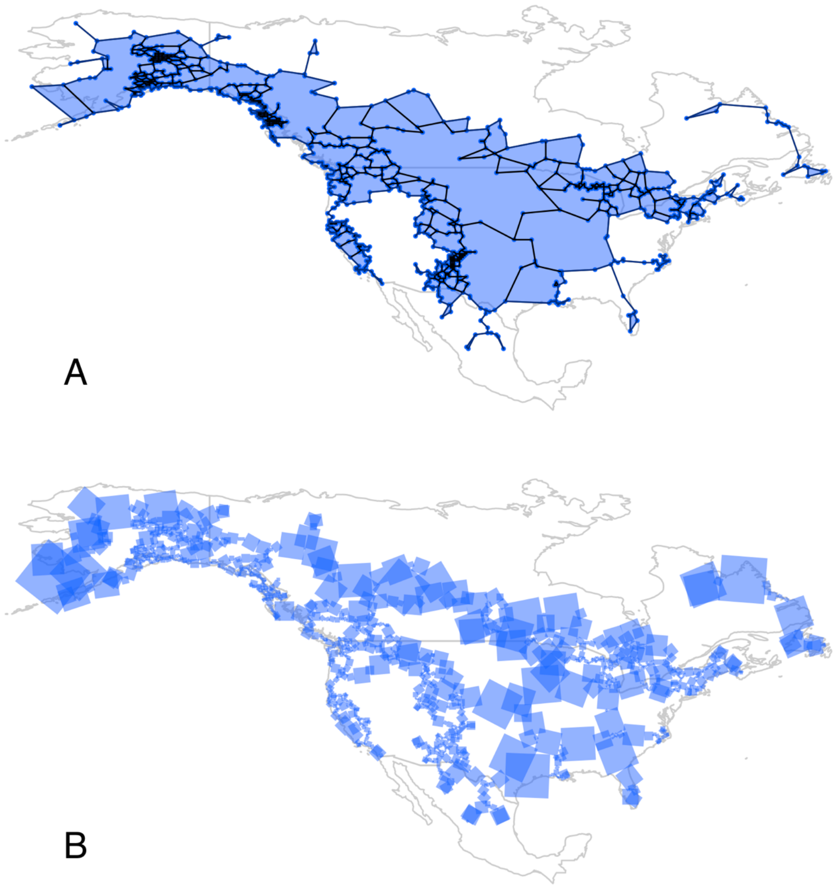

Figure 3 Data for *Ursus americanus* drawn from GBIF. There were 7068 point occurrences. Thirty-eight obvious outliers with coordinates to the east of 50º W and south of 20º N were excluded. The rest were grouped into 1162 quarter-degree cells. (A) The full set of observation points embedded in a filled mosaic. (B) The same mosaic with squares shown around each edge, as in Figs. 2C and D.

Hosted file

`image4.emf` available at https://authorea.com/users/363764/articles/484477-a-simple-graph-theoretic-method-provides-accurate-range-area-estimates

Figure 4 *Ursus americanus* mosaic areas for groups of points falling into historical intervals of five years. Here, grouping of point observations into cells was conducted separately for sets of points belonging to different intervals. Red lines = convex hull areas; blue lines = mosaic areas. Horizontal lines show values for the full data set.

Hosted file

`image5.emf` available at https://authorea.com/users/363764/articles/484477-a-simple-graph-theoretic-method-provides-accurate-range-area-estimates

Figure 5 Performance of four range estimation methods in simulation. A thousand trials are illustrated in each panel. Underlying data were drawn randomly from within circles of varying sizes, with a mean of zero and standard deviation of 0.5 on a log scale. Data for each method are offset by a factor of 10^4 for clarity. From the top, HV = hypervolumes (dark blue points); KD = 95% kernel density estimates (light blue); CH = convex hull areas (orange); MA = mosaic areas (red). (A) Sample size is five points. Adjustedr^{2} based on logged values = 0.8252 (HV), 0.7251 (KD), 0.7689 (CH), and 0.8501 (MA). (B) Sample size is 20 points.r^{2} = 0.9810 (HV), 0.8950 (KD), 0.9821 (CH), and 0.9652 (MA). (C) Sample size is 20 points, but 80% of the points in one half of the circle are removed. r^{2} = 0.9768 (HV), 0.9185 (KD), 0.9763 (CH), and 0.9699 (MA).

Hosted file

`image6.emf` available at https://authorea.com/users/363764/articles/484477-a-simple-graph-theoretic-method-provides-accurate-range-area-estimates

Figure 6 Performance of mosaic area estimation applied to ranges having six different shapes and sizes that vary from one trial to the next: 2 x 1 rectangles with one quarter removed (green); pairs of squares separated by one unit of distance horizontally and vertically (reddish violet); 4 x 1 rectangles (turquoise); squares (gold); rings with holes equal in diameter to that of the outer circle times 0.707, so half the area is empty (purple); and circles (red).r^{2} = 0.9161 (three-quarter squares), 0.9535 (pairs of squares), 0.9268 (4 x 1 rectangles), 0.9233 (squares), 0.9583 (rings), and 0.9317 (circles). Sample size is 10 data points. Data for each shape are offset by a factor of 10^4 for clarity; y-axis scale applies to the lowermost set.