Comparing Line Feature Morphology with Scale Specific Sinuosity Distributions: A Modified Earth Mover's Distance

Barry J. Kronenfeld[a]*, Barbara Buttenfield[b], Ethan Shavers[c] and Larry Stanislawski[c]

[a] Department of Geology and Geography, Eastern Illinois University, Charleston, Illinois, USA
[b] Department of Geography, University of Colorado, Boulder, Colorado, USA
[c] U.S. Geological Survey, Center of Excellence for Geospatial Information Science, Rolla Missouri, USA
* bjkronenfeld@eiu.edu

Keywords: scale-specific sinuosity, earth mover's distance

Scale Specific Sinuosity

Many geographic features have been observed to possess fractal characteristics. Unlike mathematical fractals that are self-similar across all scales, however, geographic features often exhibit different degrees of complexity at different scales of analysis (Buttenfield 1989). For this reason, standard metrics such as sinuosity and fractal dimension that describe shape complexity using a single numerical value cannot fully capture the fractal nature of geographic objects.

Recently, a scale-specific sinuosity (S3) metric was proposed to capture the degree of complexity inherent to a geographic feature at a particular scale (Stanislawski et al, 2019). Derived from the Richardson plot, the S3 metric is defined as the ratio of reduction in observed log feature length to increase in log stride length (unit of measurement), where feature length and stride length are taken as surrogates for complexity (i.e., sinuosity) and scale, respectively. Because the S3 metric is scale-specific, it produces different values depending on the chosen stride length interval. Typically, an S3 distribution is computed across a series of stride length intervals to describe morphometric complexity across a range of scales.

Derivation of the S3 metric is driven by a need to characterize stream sinuosity at various scales in differing terrain types. Models of sinuous flow patterns predict meanders of different sizes, preferred wavelengths and wavelength variance depending on slope and resistivity conditions (Lazarus and Constantine 2013). Terrain derivatives are also widely acknowledged to be scale-sensitive, and this calls for sinuosity metrics that are closely coupled to the scale of measurement. The conventional metric for bend geometry is sinuosity, commonly defined as the ratio of the path length of a stream channel to the straight-line distance between channel endpoints. This effectively summarizes bend geometry at a single scale. The capability to quantify sinuosity across a range of scales affords opportunities to examine sinuosity at a deeper level, offering advanced insights into relationships between stream channel geometry and changing hydrologic characteristics (discharge, erosion, aquifer replenishment, etc.).

Analytical applications of scale-specific sinuosity require a metric of difference between two S3 distributions. For example, analysis of temporal change in river bend geometry entails comparison of S3 distributions measured at two or more points in time. Similarly, classification of linear features requires a way to quantify the difference in S3 distributions between any pair of features, or between a single feature and a class prototype. Difference metrics form the basis for clustering algorithms as well as ordination methods such as principal components analysis.

Here, a modified earth mover's distance metric (d_{EMD*}) is proposed for use in analyzing the difference between two S3 distributions. The proposed metric is a variation on the standard earth mover's distance (d_{EMD}) between probability distributions that accounts for differences in both probabilities $p(x)$ and values x (Rubner et al. 1998). As illustrated in Figure 1, d_{EMD} is defined as the minimum cost of transforming one distribution into the other by moving mass around, where movement cost equals mass times distance. In the case of a probability distribution, "mass" refers to the probability density and "distance" is the difference between values in the two distributions.

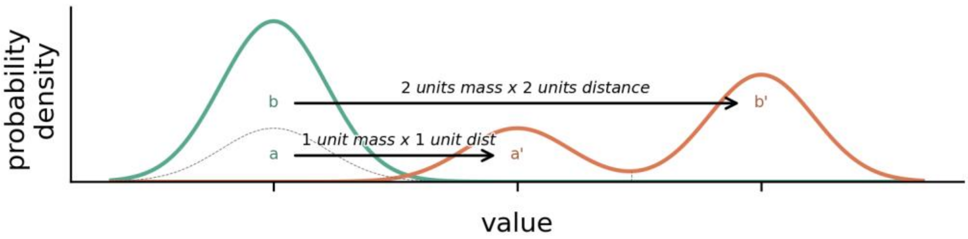

Figure 1: Illustration of earth mover's distance (d_{EMD}). To transform the left (green) distribution into the right (orange) distribution, one unit of mass at (a) is moved one unit of distance to (a´) and two units of mass at (b) are moved two units of distance to (b´) for a total cost of 1 + 4 = 5.

Adaption to S3 distributions is motivated by the need to account for differences in both degree and scale of sinuosity (Figure 2). For example, curves (b) and (c) in figure 2 differ in degree but not scale of sinuosity, whereas curves (c) and (d) differ in scale but not degree. Here the "scale" of sinuosity is synonymous with the sizes of bends, so that the bend sizes ranging from small (Figure 2c) to medium (Figure 2d) and large (Figure 2e) correspond to scales that may be described as "fine," "medium" and "coarse". By analogy with d_{EMD}, "mass" refers to the degree of sinuosity and "distance" is the difference between scales of the two distributions.

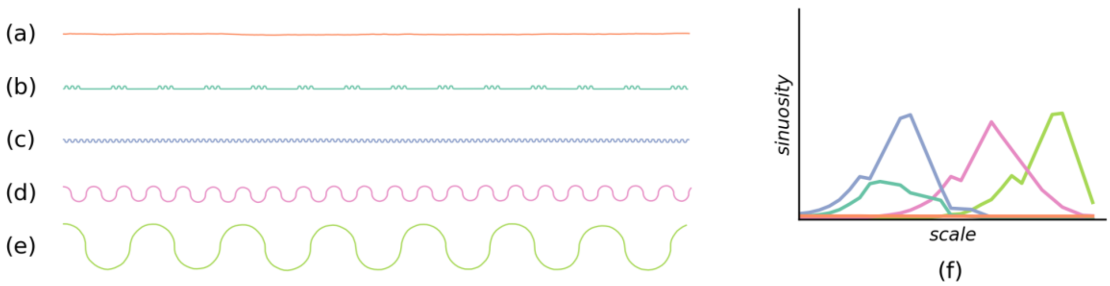

Figure 2: Curves with differing degrees and scales of sinuosity. (a) no sinuosity, (b) low sinuosity at a fine scale, (c) high sinuosity at a fine scale, (d) high sinuosity at a medium scale, (e) high sinuosity at a coarse scale, (f) corresponding S3 distributions.

Deficiency in Common Metrics

Let $\{\sigma_1, \sigma_2, \dots, \sigma_n\}$ denote the centers of the n distinct scale intervals used for measurement and let s_{ai} and s_{bi} denote the S3 values indicating the degree of sinuosity at the i^{th} scale interval for two features a and b respectively. First, let us see why a simple difference metric such as the Manhattan distance will not work. The Manhattan distance between the S3 distributions of two features a and b can be defined as follows:

$$d_{Manhattan}(\mathrm{a}, \mathrm{b}) = \sum_{i=1}^{n} |s_{ai} - s_{bi}|$$

Now consider features (c), (d) and (e) in Figure 2 and their corresponding S3 distributions. All three features have approximately the same total degree of sinuosity but at different scales. It is clear that (c) is more similar to (d) than to (e), so we should resolve that:

$$d_{Manhattan}(\mathrm{c}, \mathrm{d}) < d_{Manhattan}(\mathrm{c}, \mathrm{e}) \qquad (desired\ result)$$

Figure 3 shows the actual differences between the S3 distributions of these curves. The Manhattan distances are equal to the area of the shaded regions in the figure. Although the desired result is technically achieved due to some overlap between the S3 distributions of (c) and (d), the difference is small. It is possible to construct an example where the Manhattan distances are exactly equal, but such examples are difficult to visualize graphically because of the wide tails of most S3 distributions. In any case, for practical purposes we can say that:

$$d_{Manhattan}(\mathrm{c}, \mathrm{d}) \cong d_{Manhattan}(\mathrm{c}, \mathrm{e}) \qquad (observed\ result)$$

The problem is that the Manhattan distance compares S3 values at each scale independently and does not consider the proximity of the scales at which peak values occur in the S3 distributions being compared. The same problem exists for most distance metrics, and it is this problem that the Earth Mover's Distance iss designed to address.

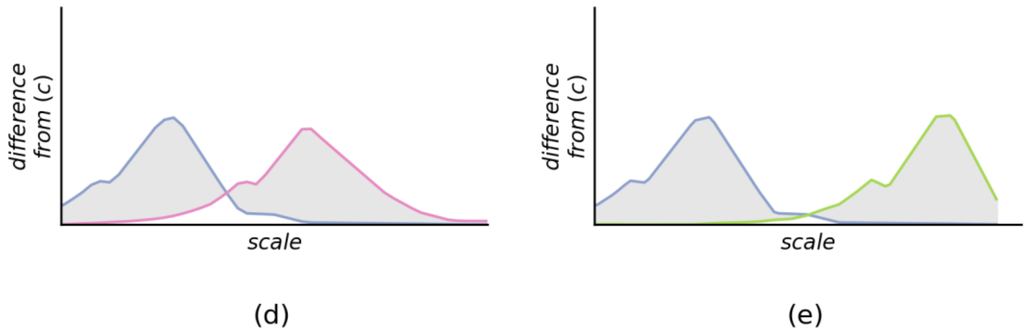

Figure 3: Differences between S3 distributions of curves shown in Figure 2. Left: difference between curves (c) and (d); right: difference between curves (c) and (e).

Triangle Inequality

Modification of d_{EMD} is required because unlike probability distributions, S3 distributions are not constrained to have equal total mass. The total mass of an S3 distribution for a given feature is proportional to the log of the total sinuosity of that feature (Stanislawski et al, 2019), where total sinuosity is the sum of S3 values across all scales. Since we wish to compare features with different total sinuosity, the constant total mass assumption cannot be met. A natural modification is to allow mass to be removed from the larger distribution to equalize their total masses, with a fixed removal cost (c_r) applied per unit mass removed. Let us refer to this metric as $d_{EMD'}$. Unfortunately, $d_{EMD'}$ will break the triangle inequality. The triangle inequality is a fundamental mathematical property of all mathematical metrics and requires that the direct distance between two data observations a and c cannot be larger than the distance through an intermediate observation b:

$$d(a,c) \leq d(a,b) + d(b,c)$$

A problem arises with the "triangle" formed between the S3 distributions of a straight line and two curves with equal sinuosity but at different scales. Consider again curves (a), (c) and (e) in Figure 2, reproduced in Figure 4 with relevant quantities illustrated. We must preserve the following inequality:

$$d_{EMD'}(c,e) \leq d_{EMD'}(c,a) + d_{EMD'}(a,e)$$

Let $m_c = m_e$ be the masses (i.e. total sinuosity) of S3 distributions (c) and (e), and let $\delta\sigma$ be the difference between the scales of sinuosity of these two distributions. Then we have the following:

$$d_{EMD'}(c,e) = m_c \times \delta\sigma$$

$$d_{EMD'}(c,a) = m_c \times c_r$$

$$d_{EMD'}(a,e) = m_c \times c_r$$

If $\delta\sigma > 2c_r$ then $d_{EMD'}(c,e) > d_{EMD'}(c,a) + d_{EMD'}(a,e)$ and the triangle inequality will be broken.

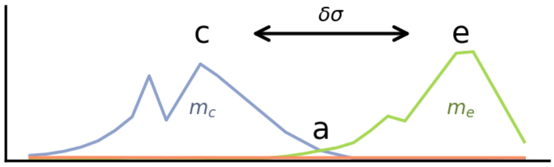

Figure 4: S3 distributions of curves (a), (c) and (e) from Figure 2 with quantities required to compute $d_{EMD'}$: m_c is the mass of the S3 distribution for feature c, m_e is the mass of the S3 distribution for feature e, and $\delta\sigma$ is the (average) difference in the scales of sinuosity for the two S3 distributions.

Modified Earth Mover's Distance for S3 Distributions

To solve this problem, it is necessary and sufficient to ensure that the cost of moving mass never exceeds twice the cost of removing mass. This leads to the modified Earth Mover's Distance (d_{EMD*}), defined as the minimum cost of transforming one distribution into the other by either moving or removing mass, with a fixed removal cost (c_r) per unit mass removed and movement cost equal to mass times distance but capped at a maximum of $2 \times c_r$.

It will be noticed that the removal cost c_r influences the resulting metric. This defines the relative importance of differences in degree vs. scale of sinuosity. The higher the value of c_r the greater the importance of degree of sinuosity relative to scale of sinuosity. Since movement cost is capped at $2 \times c_r$, the parameter c_r also defines the limit of scale difference considered in assessing similarity. If two curves have equal sinuosity at scales separated by greater than a factor c_r then the difference between them will be measured as equivalent to the difference between one of the curves and a straight line. For example, the meanders of Figure 2e are 15x as wide as Figure 2c. Since scales of the S3 distribution are defined logarithmically, setting $c_r = \log(15)$ would indicate that the curve in Figure 2e is considered as different from Figure 2c as it is from a straight line.

Implementation

The proposed metric is implemented using the transportation cost minimization algorithm *min_cost_flow* in the NetworkX Python package (Hagbert et al. 2008). The problem is structured by establishing a network with supply and demand nodes corresponding to the scale intervals in each distribution. The *min_cost_flow* algorithm requires that movement costs be integer values; this is achieved by multiplying all costs by a sufficiently large number (e.g., 10,000) to achieve the desired resolution on the scale axis and then rounding to the nearest integer. Empirical analysis shows computation time to scale up as a function of the square of the size of the input S^3 distributions.

Application

The authors stress that the present study forms an initial test and evaluation of the proposed metric. The d_{EMD*} metric is being used to classify stream features across the USA by morphometry. Data consists of 50 stream segments ranging from 37km to 206km in length derived from the National Hydrography Dataset from across the conterminous United States (Stanislawski et al. 2018). First, stream segments are divided into 10km to 20km reaches by splitting features at locations that maximize d_{EMD*} between adjacent reaches. Reaches are classified using a variant of the k-means clustering algorithm with the d_{EMD*} metric serving as the distance function. The same process is being carried out with other distance metrics, and results will be compared through visual examination and analysis of percent variance explained by each clustering solution.

Disclaimer:

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Buttenfield, B.P. (1989). Scale-dependence and self-similarity in cartographic lines. Cartographica, 26 (1), 79-100. Doi: 10.3138/4678-60MR-M036-4732

Hagberg, A.A., Schult, D.A. and Swart, P.J. (2008). Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008.

Lazarus E D, and Constantine J A. (2013). Generic theory for channel sinuosity, Proceedings National Academy of Sciences 110(21): 8447-8452

Rubner, Y., Tomasi, C. and Guibas, L.J. (1998). A metric for distributions with applications to image databases. Proceedings of the 1998 IEEE International Conference on Computer Vision, Bombay, India.

Stanislawski, L.V., Buttenfield, B.P., Kronenfeld, B.J., and Shavers, E.J. (2019). Scale-specific metrics for adaptive generalization and geomorphic classification of stream features. 2019 ICA Workshop, Abstractions, Scales, and Perception, Tokyo, Japan, July 15, 2019. Accessed May 25, 2022 at https://generalisation.icaci.org/downloads/abs2019/Abs2019_paper_9.pdf

Stainslawski, L.V., Kronenfeld, B.J., Buttenfield, B.P. and Brockmeyer, T. (2018). Generalizing linear stream features to preserve sinuosity for analysis and display: a pilot study in multi-scale data science. In Freundschuh and Sinton (eds.), Proceedings of AutoCarto/UCGIS 2018, Madison, Wisconsin, USA, May 22-24, 2018, pp. 111-119. f