Multi-Scale Analysis of Surface-Adjusted Geographic Distance

Andrew Eaman, Margaret Brantley, Mehran Ghandehari

Department of Geography, University of Colorado at Boulder, Colorado, USA

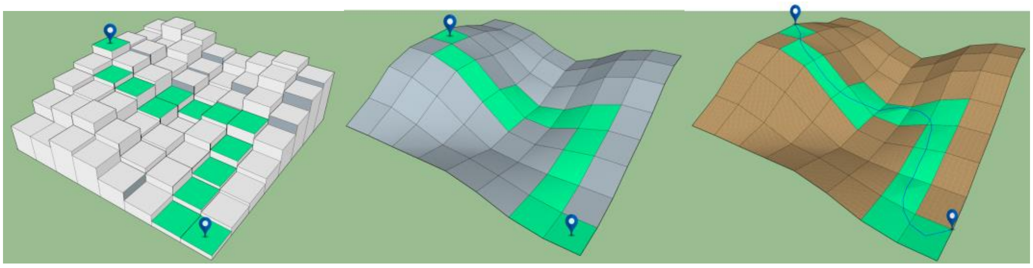
1 Introduction

Distance is one of the most important measures in spatial modeling. Distance between features and/or pixels has a significant role in proximity and neighborhood analysis (e.g., buffer, and Thiessen polygons), spatial interpolation (e.g., inverse distance weighting, and Kriging), measuring spatial relationships, dependency and autocorrelation, hot spot, pattern and cluster analysis, terrain analysis, allocation analysis, least cost path analysis, routing, path planning and network analysis, etc.

Several approaches have been employed to measure distance. Planar metrics (i.e., measuring distance along a straight line between two points, as the crow flies) are most commonly used as they are simple and intuitive. But planar metrics decrease precision and introduce uncertainty in spatial modeling because of assumptions that terrain is uniform, ignoring slope and curvature. This project incorporates surface geometries of terrain to develop a foundation of a *complete-surface-adjusted distance*. Surface-adjustment can proceed by several methods which be compared statistically; the method or methods that are selected by this procedure will be tested and validated in different data resolutions.

In order to calculate surface adjusted distance, Digital Elevation Model (DEM) is needed. DEM is currently modeled in a grid of pixels, assuming that elevation values are constant within any single pixel (rigid pixel paradigm) (Figure 1-a). This paradigm generates imprecise distance measurements to a degree dependent upon terrain roughness and landscape conditions. For individual pixels the imprecision may seem small, but additive effects across a study area can propagate dramatically. In some of the available methods, the slope of the pixels is taken into account. These methods can be called partially-surface-adjusted (Fig 1-b). But in truth, terrain can bend, twist and undulate within each pixel. In this project, polynomial functions (bi-linear, bi-quadratic, and bi-cubic) are used to simulate curvilinear pixels.

Furthermore, the sensitivity of surface adjusted distance to resolution is investigated. Distance in different resolutions (here, 10, 30, 100, 1000, and 5000 meter) are calculated and compared with the planar distance, geodesic distance, rigid pixel paradigm, and all of them are validated against finer resolution terrain data (here, 3 meter).



(a) (b) (c)

Fig. 1: Different approaches for calculating surface distance: (a) pixel to pixel method (rigid pixel paradigm), (b) partially-surface-adjusted (pixels are tilted), and (c) complete-surface-adjusted (pixels are curved).

2 Methods

There are several methods for measuring distance; some take the shape of terrain into account and others do not. Here we will compare and contrast those different methods by examining how distance is calculated and in what ways they adjust for terrain.

2.1 Planar Distance

Planar distance, often termed Euclidean distance, is the shortest length measurement between two points on a flat surface. In general, this method is most common for short distances. It assumes that the points lie on a flat surface, and uses the Pythagorean Theorem to find the shortest distance between two points. Of course the earth is not flat and points on the earth are projected in many ways causing this method to be inaccurate, especially for long distances.

2.2 Geodesic Distance

One method that incorporates the curvature of the earth is Geodesic distance, or ‘great circle distance.’ Unlike the planar method, the shortest line must follow the arc of the earth or the arc of a great circle. For any two points on a sphere, a great circle can be derived so that both points create a minor arc within that great circle. This method can be used for very long distances on earth. The curvature of the line intensifies as the points move farther away.

2.3 Pixel to Pixel Distance

Pixel to Pixel distance is a method that a line moves in a zig-zag way through the pixel centroids. Using the elevation from each pixel that intersects with a given line or transect, a Euclidean based calculation moves across the data, one pixel to the adjacent pixel, in order to calculate the shortest 3D distance between two points. This movement across adjacent pixels creates a directional bias because the pixels that intersect with the line move in only one of the eight possible directions.

Here, the implemented method for calculating pixel to pixel surface distance is explained; first, extract by mask tool is used to find where the transect and the DEM overlaps or intersects. The importance of extract by mask instead of the polyline to raster conversion tool, is the extract by mask obtains the elevation values of each pixel whereas polyline to raster tool leaves the elevation values out of the attribute table. Then the rasterized transect is converted to point data – one point per pixel. Then, it is necessary to obtain the X, Y and Z coordinates of each point. Because the points are not in order, the slope of the line must be calculated using the start and end points of the transect in order to determine a specific method for extracting the elevation of points from the attribute table. If the slope is negative, the attribute table will correctly move through the point values. But, if the slope of line is positive, the order of points that have the same y coordinate is reversed. In fact, when the slope is positive the order of the points within the attribute table are essentially reversed because of the left -to-right direction ArcGIS reads objects into the table. Once the points have been put in the correct order.

2.4 Partially Surface Adjusted Distance

The next four methods all partially adjust for terrain, meaning known elevation values are used to interpolate a given sample point on a line. That is, the elevation of each sample point is interpolated based on the underlying and/or surrounding pixels. Then the 3d sample points along the transect are used to reconstruct the 3d shape of the transect and calculate distance.

Closest (Based on the Closest Samples within a Pixel)

In this method, it is assumed that pixels are flat and pixel values are assigned to all of the sample points that lie within a pixel. That is, all of the sample points that are located on one pixel would get the same elevation.

Triangular Irregular Network (TIN)

The TIN method, uses a series of triangles that are irregularly spread over a plane. Then the elevation of sample points are linearly interpolated based on the 3d triangles (facets).

Natural Neighbor

The Natural Neighbor method works based on the Voronoi diagram. This diagram takes sample points and finds the corresponding voronoi cell. Then by adding the sample points a new diagram is created. The two diagrams are then used to find the difference in cell area before and after adding the sample point in order to calculate the weight of each neighbor to interpolate the elevation of sample point.

Weighted Average

Weighted average is a deterministic interpolation technique that is based on a weighted average of the elevation of neighbor pixels. Neighbor pixels that are closer to the sample point being interpolated would have a higher weight on average than points that are farther away. Different functions can be used con-ceptualized this concept. The simplest and most common way is the so-called "Inverse Distance Weighting" (IDW) method that the weight of each neighbor is the inverse of distance-squared to the unknown point. This method so-called Inverse Distance Weighting (IDW) that is commonly used for continues data such as terrain.

2.5 Completely Surface Adjusted Distance

The next three methods are based on local polynomial fitting with varying amounts of terms and degrees of the variable X and Y. The increasing amount of terms are meant to create a surface with more freedom to undulate. The number of terms is based on the number of neighboring cells around the central cell where the transect intersects the raster data.

Bi-linear

The first method, bi-linear, assigns an elevation to the sample point, by taking into account four terms determined by the four immediate surrounding cells. It fits a first order polynomial to the four surrounding cells. For example the center cell will have four directly adjacent neighbors, one for each cardinal direction.

Bi-quadratic

Similarly, the bi-quadratic method incorporates the center cell's values and the adjacent eight cells, or both the cardinal and intercardinal directions, in order to calculate the 3D surface where the sample points lay. Using nine terms as opposed the bi-linear's four, allows for more freedom of undulation for the surface calculated. Incorporating a few more terms increase the number of local maximums and minimums.

Bi-cubic

In the bi-cubic method the number of terms is increased to sixteen. The surrounding eight are incorporated, just as in the bi-quadratic method, as well as the one more cell in each of the cardinal and intercardinal directions. This method calculates a surface with the even more freedom of undulation which creates a surface with more local maximums and minimums.

3 Data Sources and Study Area

The study area was chosen based on the availability of 3 meter resolution DEMs to be used as a measure of validation. While the other resolutions are available in all the continental US, coverage at such a fine resolution is much more limited. . The study area is located in western North Carolina between the coordinates of 3990107.02 m N-3933263.173045 m N, and 388680.868520 m W - 525652.441471 m W. Its location at the southeast end of the Appalachian mountain range, provides a variety of elevations, ranging from 209 to 1602 meters. In order to test the effectiveness of each method across resolutions, the DEM was downloaded at 3, 10, 30, 100, 1000, and 5000 m resolutions. The 3, 10, and 30 m resolutions were downloaded from Data Gateway (https://gdg.sc.egov.usda.gov/). The 100 and 1000 m resolutions were downloaded from SRTM (http://dds.cr.usgs.gov/srtm/version2_1/). The 5000 m resolution was provided by Barbara Buttenfield, who was given them by USGS. The DEMs are in NAD_1983_UTM_Zone_17N coordinate system and Projection. The data was downloaded as a series of individual tiles that had to be mosaicked into a single DEM and then clipped to the same extent.

On the study area, we randomly positioned five transects in the area ranging from 39,196 m to 105,173 m (planar), to be assigned surface information in order to find a value of distance for each of the methods in each resolution.

4 Results and Discussion

The distance of transects have been calculated based on 3 meter sample size using different methods and resolutions. Here a set of graphs are discussed that represent the effectiveness of each method taken the 3 meter resolution DEM as a benchmark, and the variation of distance in different resolution.

Figure 2 shows the Root Mean Square Error (RMSE) of transects for each method of distance measurement in a resolution. Each graph in this figure represents a different resolution, with methods on the x-axis and error on the y axis. Initially the most obvious trend is the high error of planar, geodesic, and pixel-to-pixel methods. These can be attributed to the lack of surface information for the first two, and the directional bias for the last. In order to more closely examine the results of the remaining methods, planar, geodesic, and pixel-to-pixel were isolated into separate graphs. From this visualization, we can draw several conclusions about the effectiveness of methods. The first is that the methods we implemented outside of arc are consistently more accurate than the methods that arc provides. We also found that the weighted average method provides a reliable measurement across resolutions, always having one of the lowest RMSEs. In contrast, arcTIN and arcNN have a high RMSE in all resolutions. Finally, and surprisingly, in graphing our RMSEs the tendency for error to increase as the number of order of polynomial function is revealed. This finding is unexpected because the higher number of terms in a polynomial equation permits a smoother measurement of undulating surface.

Figure 3 shows the residuals (i.e., distances subtracted from the calculated distance in 3 meter resolution DEM), or the difference of each transect from its benchmark distance. The x axis represents the method used to measure distance, and the y axis represents the residuals. Each transect is represented by a different colored point. Planar, geodesic, and pixel-to-pixel methods were separated into their own graphs for the same reason mentioned above. The most noticeable pattern in this graphs is the negative values for the closest and arc closest methods, indicating a consistent overestimation of distance. This is due to the assumption in the closest method of a uniform terrain within the pixel, when in actuality sample points have different elevation. For this reason surface distance will be overestimated. However, this pattern becomes less prominent as resolution becomes coarser. In contrast, the other methods have a much higher tendency to underestimate distance measurements. Further inspection reveals the increased accuracy of distance measurements in finer resolution which systematically decreases with coarser resolution. Finally an irregular pattern in the 5000 meter resolution is noticeable, and led us to inspect the source data more closely. From this it was revealed that our 5000 meter data is resampled from 100 meter data.

The trend of distance change from one resolution to another is different for one of the transects. It shows that transect length and terrain type are important factors that need to be controlled in our experiments. The next phase of this project would be focused on the effects of terrain type on surface adjusted distance. Furthermore, the RMSE of some of the methods decrease as the resolution increase. This unexpected result is due to the sample size. We used the same sample size for all of the resolutions. Therefore, each resolution has an effective sample size and more investigation is required in this respect.

To find the best method for interpolating elevation of sample points in a regular grid another scenario has been designed. 1000 random points are generated all over the study area, and the elevation of each sample points are interpolated using different interpolation techniques (weighted average, bi-linear, bi-quadratic, bi-cubic). Also first and second order neighbors are used in different ways in these interpolation techniques. Then the interpolated elevation of samples for each resolution and interpolation method is compared against fine resolution DEM (here, 3 meter resolution). Residuals and RMSE are calculated and different types of graphs are generated to assess the accuracy of each method across resolutions. Figure A3 (in the appendices) shows the residual points of each random point. Each column shows a different resolution with the y axis showing residuals, and the x axis showing the method used. In each resolution, residuals are distributed around the zero value, and their range around zero increases as resolution becomes coarser. However, in the 5000 meter resolution an obvious irregularity in the data shows residual points that are distributed around 500, and not zero. This is a result of the resampling mentioned previously. In Figure A4, RMSE is on the y-axis, and the methods used are on the x-axis, labeled based on the number of pixels incorporated into the calculation. The graphs in this figure illustrate that the weighted average has the most accurate result and the closest method has the least accurate result and also that the optimal number of pixels to include in a weighted average calculation is 8. It also can be seen that weighted average and bi-linear methods lead to smaller RMSE in finer resolutions (10, 30, and 100 meter), but bi-quadratic and bi-cubic methods work better in coarser resolutions (1000 and 5000 meters). This means that polynomial functions of higher orders can better model the slope and curvature of pixels in courser resolutions, but can exaggerate the curvature of the terrain in finer resolutions.

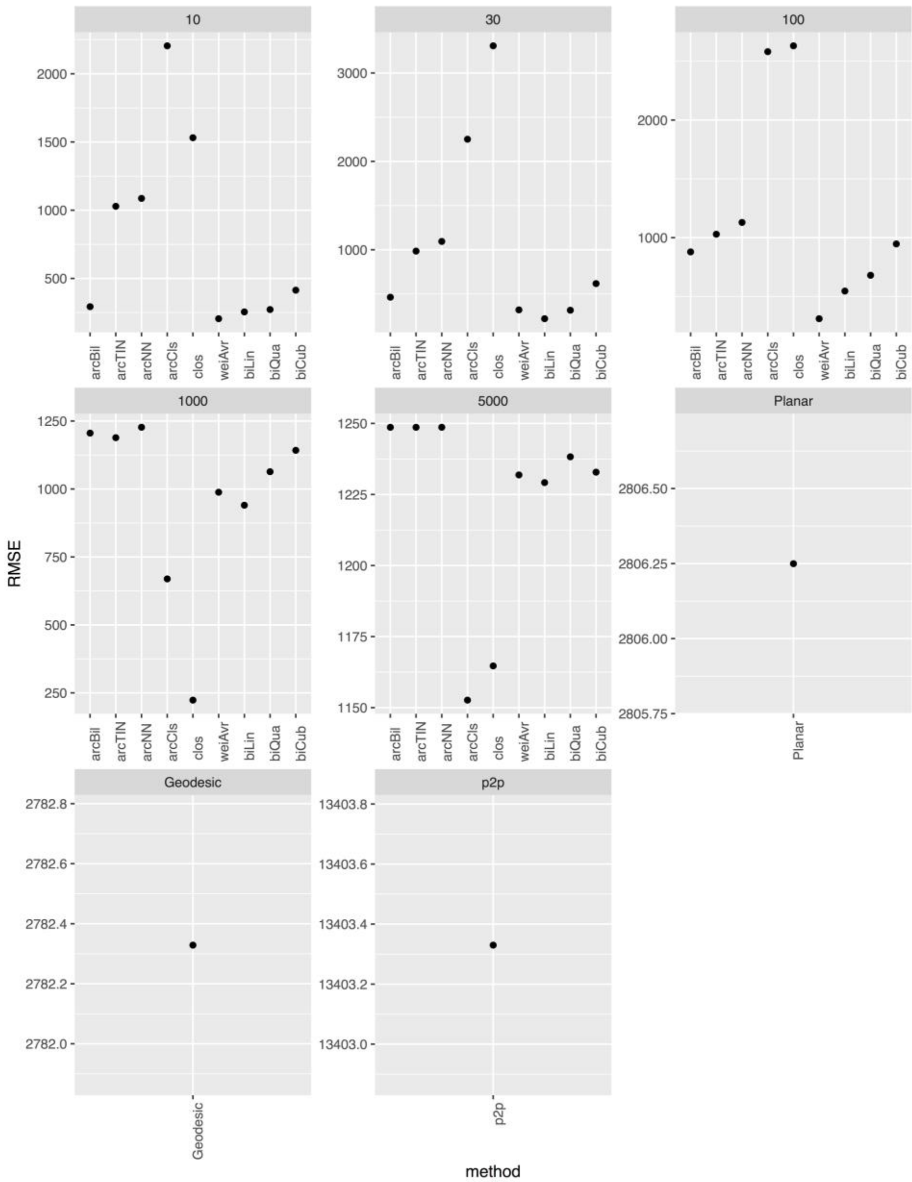


Fig. 2: Graphs of the root-mean square error. Each graph represents a different resolution, with the exception of the last three. We separated planar, geodesic, and pixel to pixel methods because the error was so high it skewed the graphs so that the patterns for the other methods were not visible.

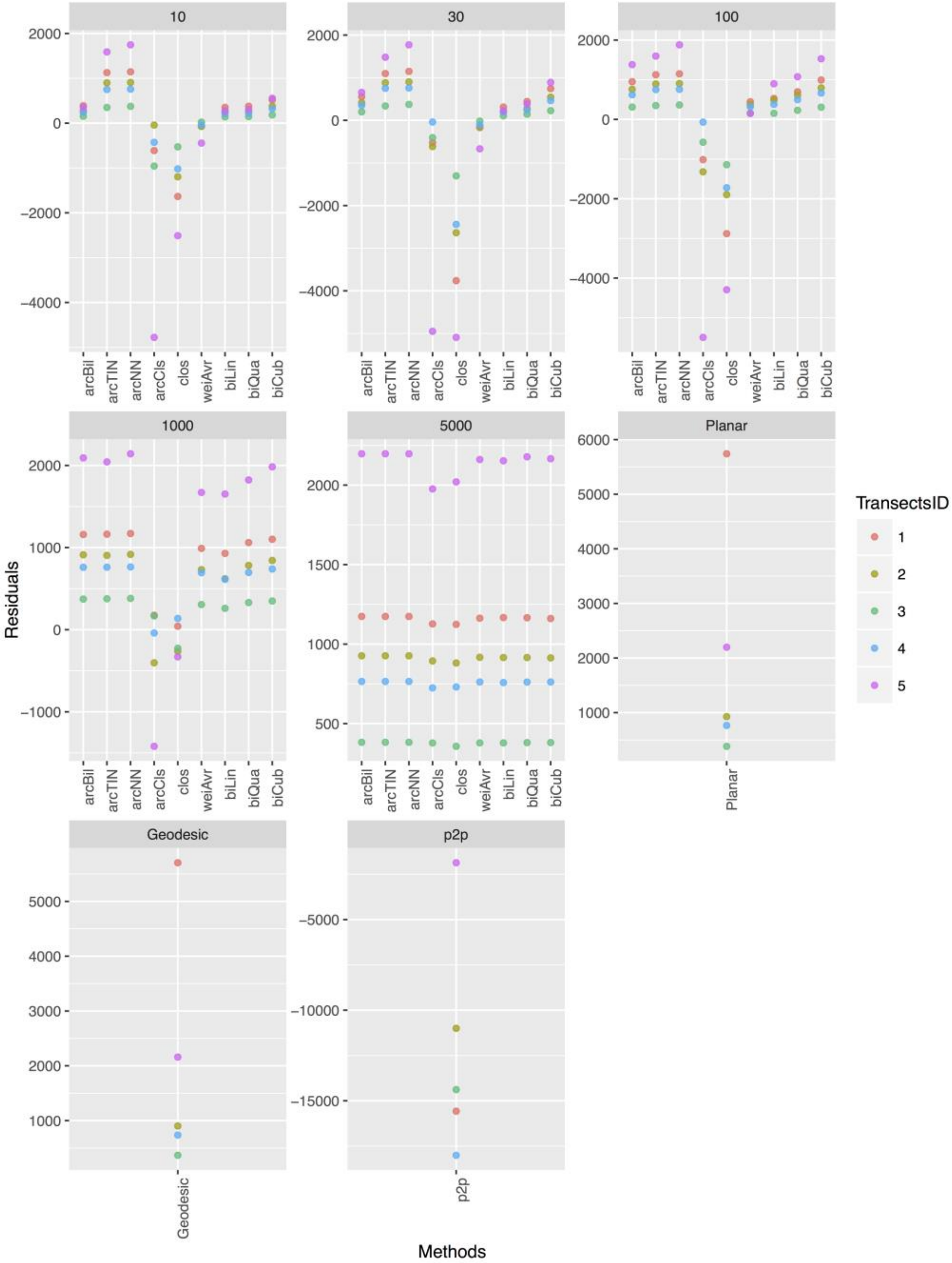


Fig. 3: These graphs shows the residuals, or the difference of each transect from its benchmark distance. The x axis represents the method used to measure distance, and the y axis difference of each transect from the benchmark distance. Each transect is represented by a different color.

5 Conclusion

In this project, calculating geographic distance using different methods in various DEM resolutions are investigated. In order to find a measure of confidence in the methods of distance measurement, we ran a code which adds surface information to different transects by looping through methods, DEM resolutions, and transects. This code enabled us to create a matrix which shows the difference between each method/ resolution pair and the benchmark distance, which was calculated on a 3M DEM. In conclusion, some interesting patterns and behaviors of surface distance have revealed, but no method has proven to be the most reliable in all of the resolutions. The results have provided a basis for further examination into the data and methods of distance measurements in order to identify possible causes for patterns and behaviors so that we can apply these methods to a diverse set of resolution and terrains in the US.

Acknowledgments

We have to express our appreciation to the Prof Babara P. Buttenfield who provided expertise that greatly assisted the project. We also have to express our thankfulness to the Stefan Leyk and Alex Stum for sharing their knowledge and pearls of wisdom with us during the course of this project.

References

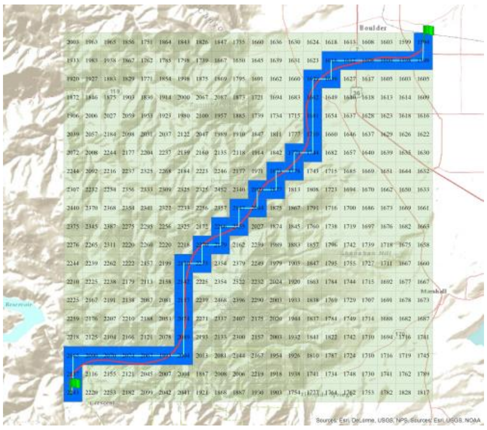
https://en.wikipedia.org/wiki/Geographical_distance

http://shreerangpatwardhan.blogspot.com/2011/01/geodesic-polyline.html

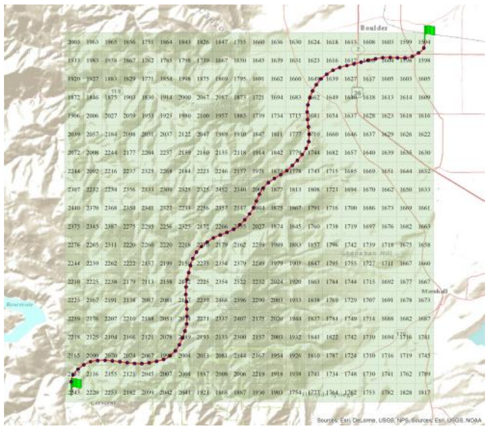
http://www.spatialanalysisonline.com/HTML/index.html?inverse_distance_weighting_idw.htm

http://maps.unomaha.edu/Peterson/gisII/ESRImanuals/Ch3_Principles.pdf

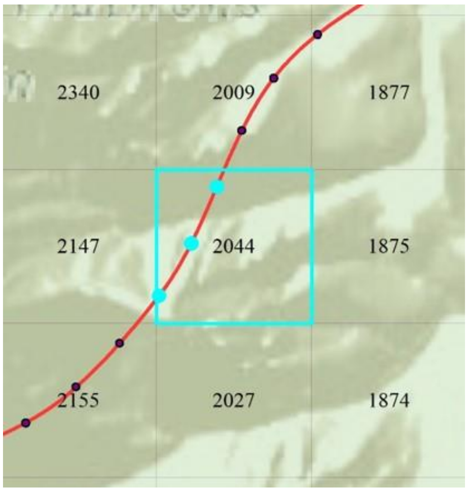
Appendices



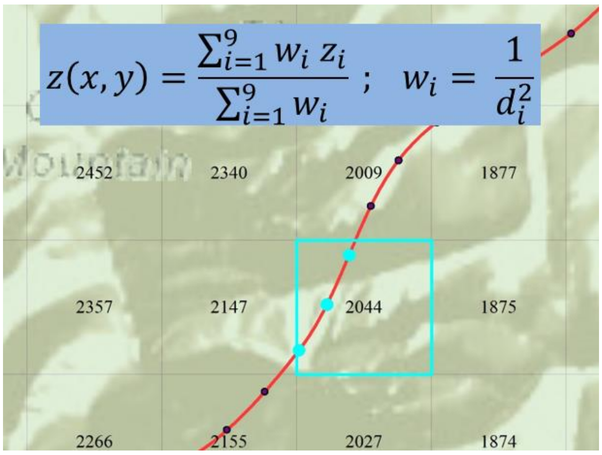
(a)



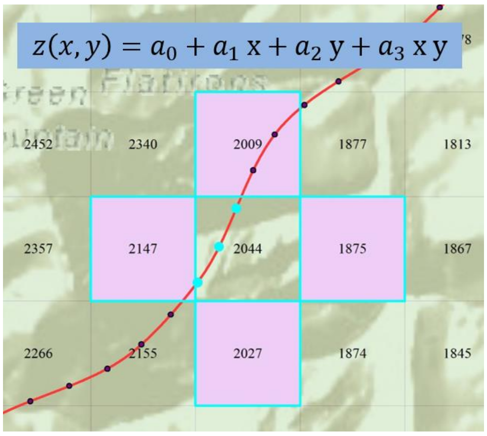
(b)



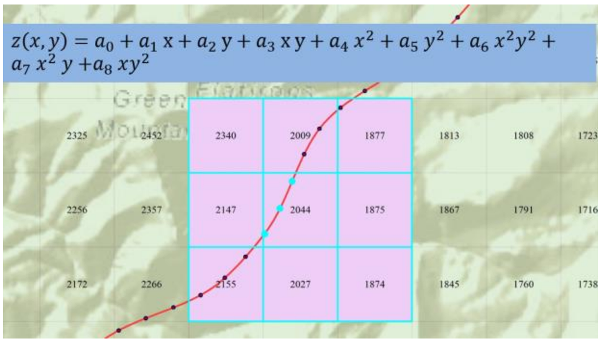
(c)



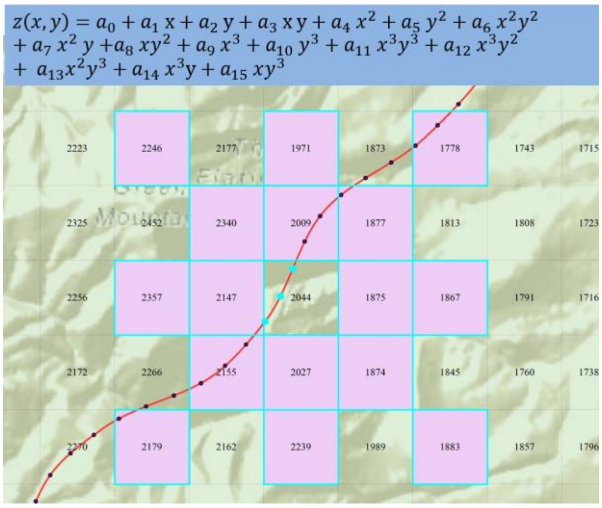
(d)



(e)



(f)



(g)

Fig. A1: (a) pixel to pixel method, (b) sample point along the transect, (c) within a pixel method, (d) weighted average method, (e) bi-linear method, (f) bi-quadratic method, and (g) bi-cubic method

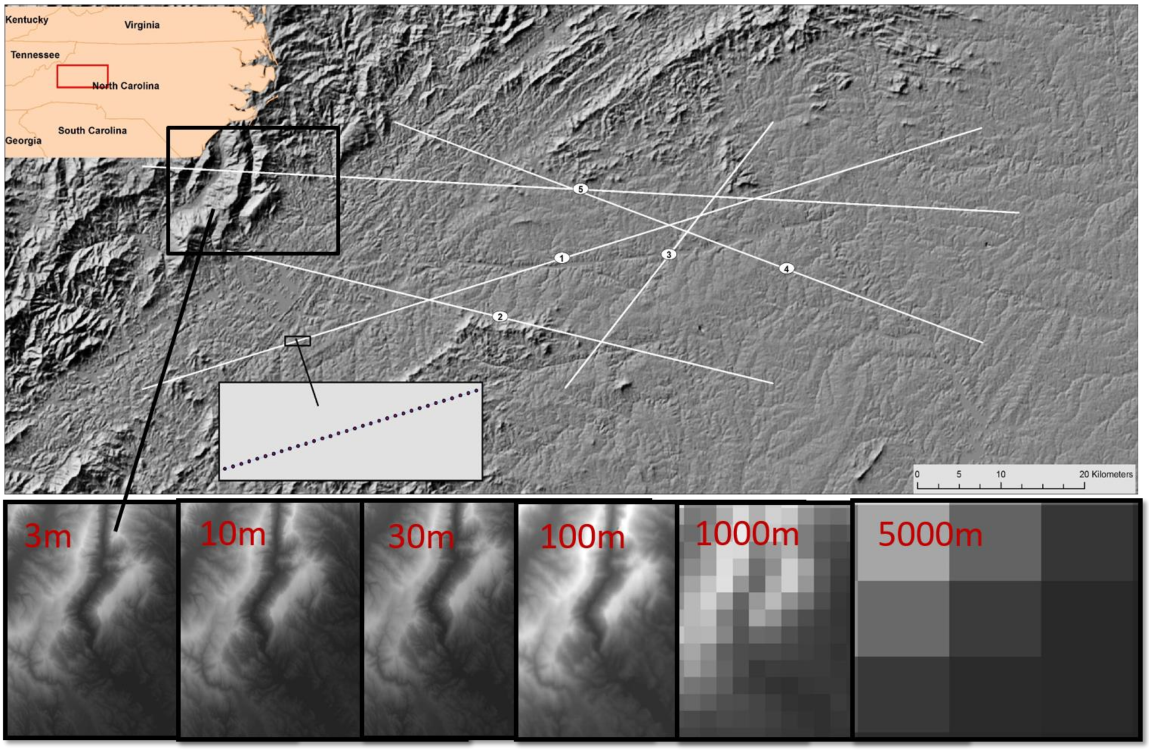


Fig. A2: Study area

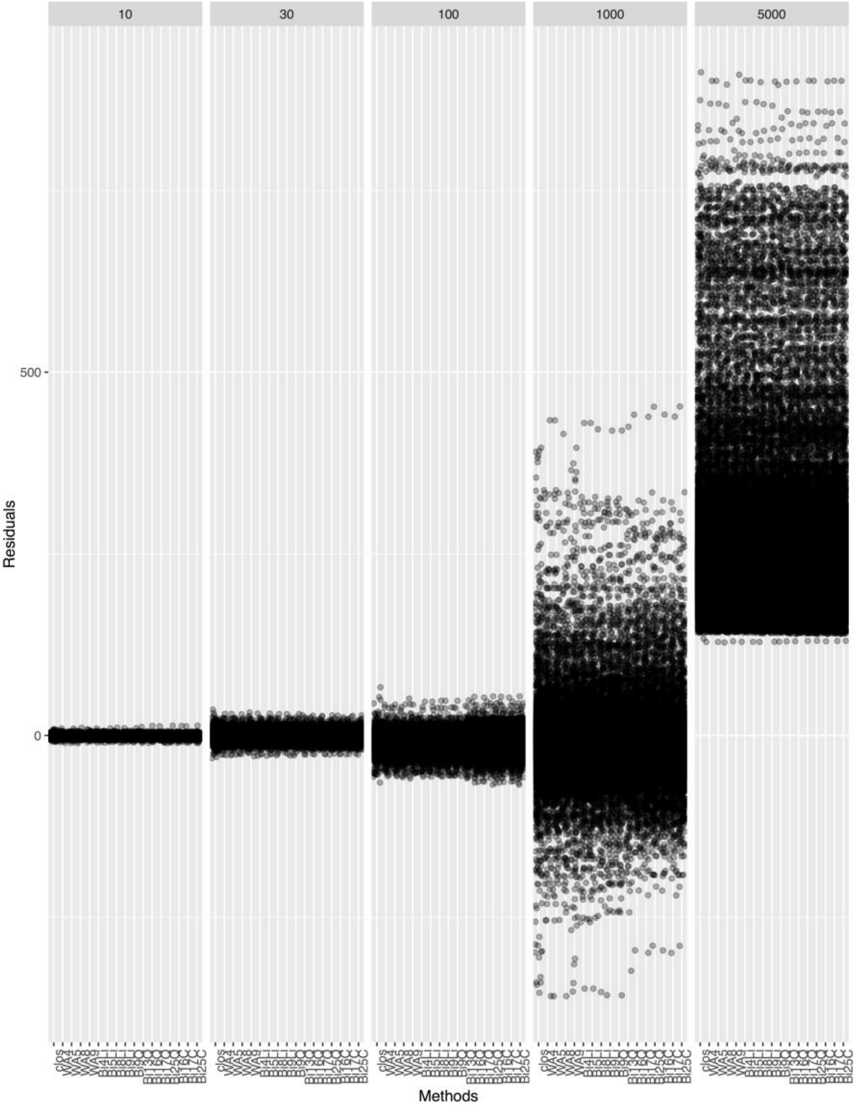


Fig. A3: Distribution of residuals for different methods in different resolutions

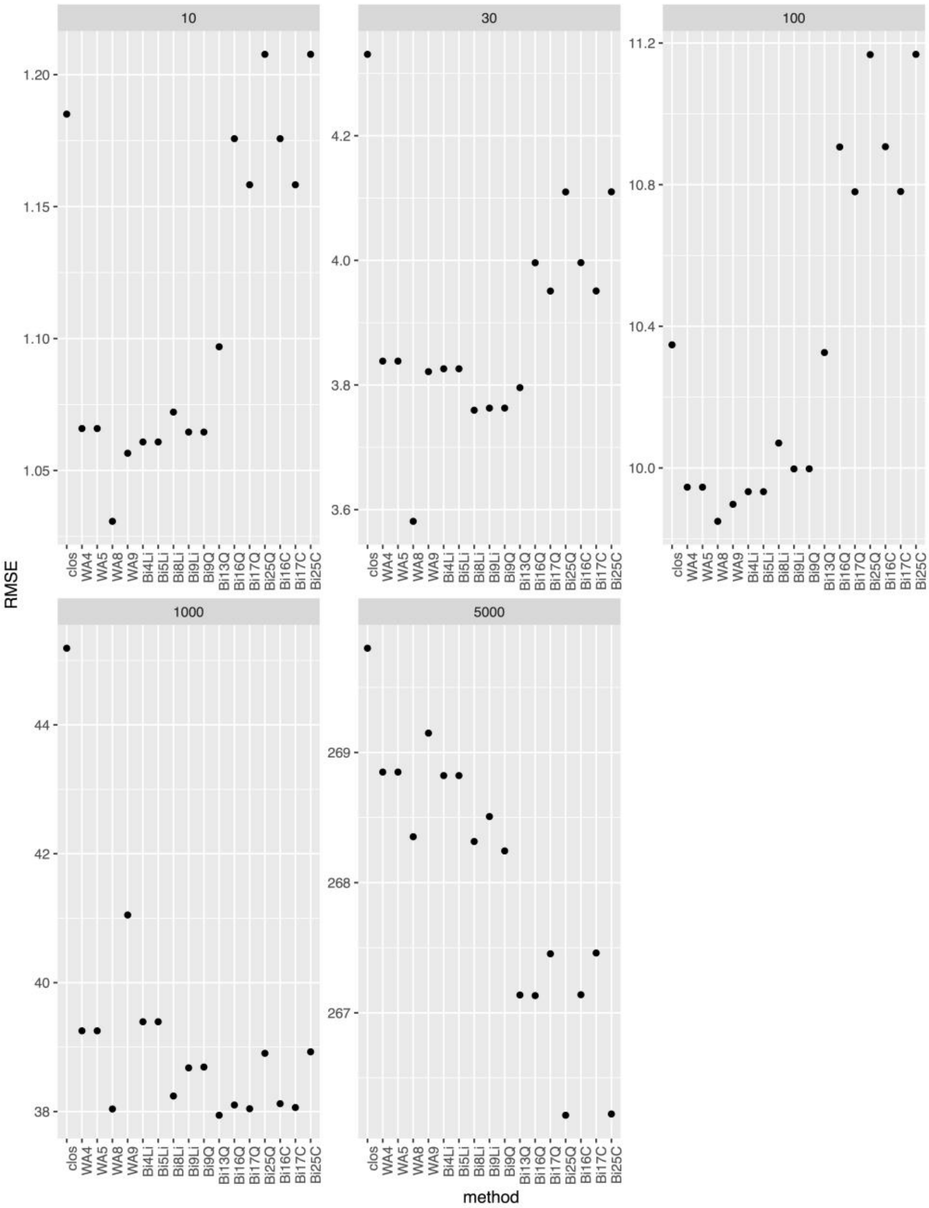


Fig. A4: RMSE of different methods in various resolutions

Pseudocode1 and Pseudocode1 are for the pixel to pixel method:

Pseudocode1:

use a searchCursor on the transect shapefile using the Id
make a temporary layer from the shapefile
Create a variable by select by attribute
extract by mask using the DEM and the selected feature
take the extracted mask and convert it into points
save output as shapefile

Note: The points of the transect are placed at the centroid of the pixel where the transect overlap the DEM.

Pseudocode2:

```
Search cursor for the masked points
    Find X,Y,Z values
    Find the start and end points
    Calculate the slope (end Y - start Y)/(end X- start X)
    If slope is greater than 0
        For each point
                If Y1 equals Y2
                Reverse the list containing Z values
Subtract the Z values of the first and next point and square it
Square the resolution and add the difference is Z values
Take the square root of the value calculated
Sum each of those sqr rooted values.
Return sum
```