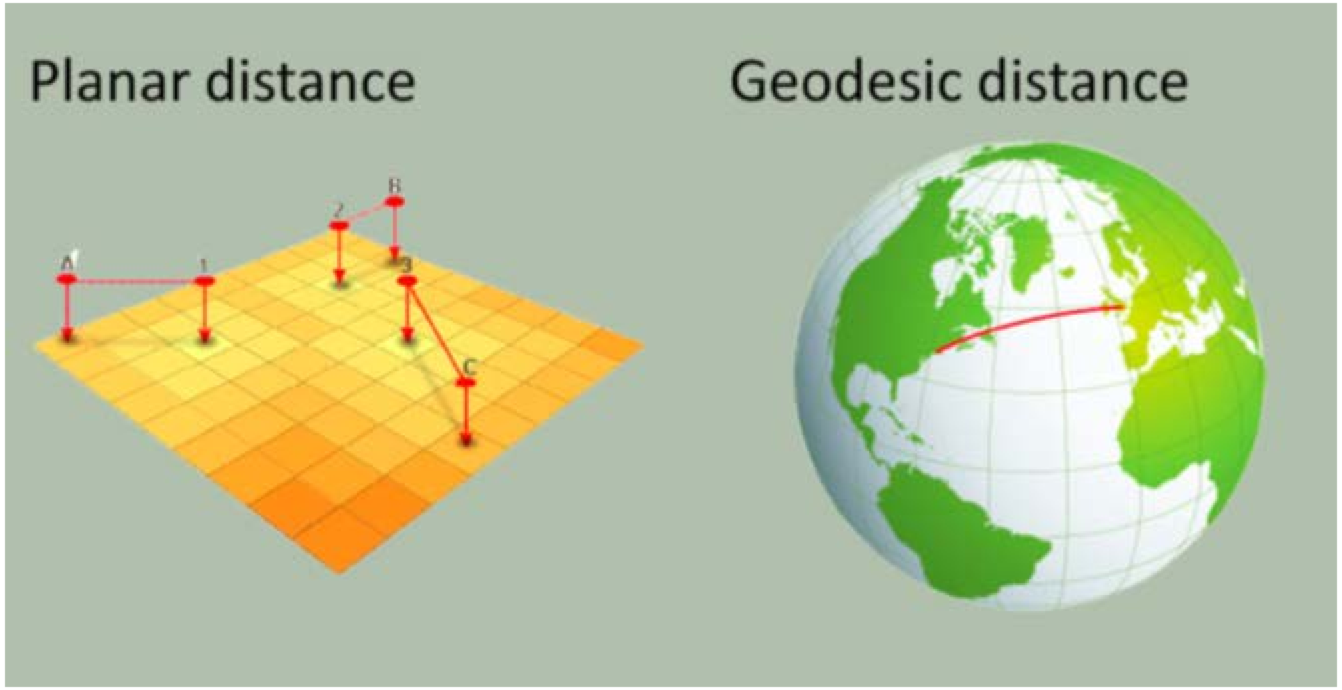
A Systematic Evaluation of Surface-Adjusted Distance Measurements using a HPC-enabled Monte Carlo Simulation

Yi Qiang[1], Barbara P. Buttenfield[2], Maxwell Joseph[2], and Carson Farmer[2]

1. University of Hawaii – Manoa

2. University of Colorado – Boulder

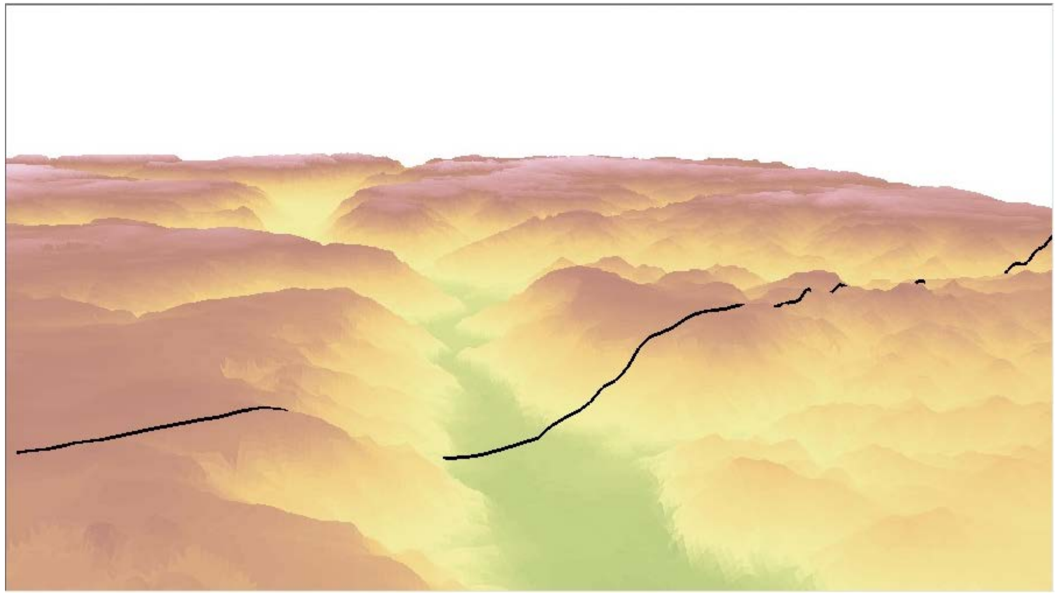
Geographical distances

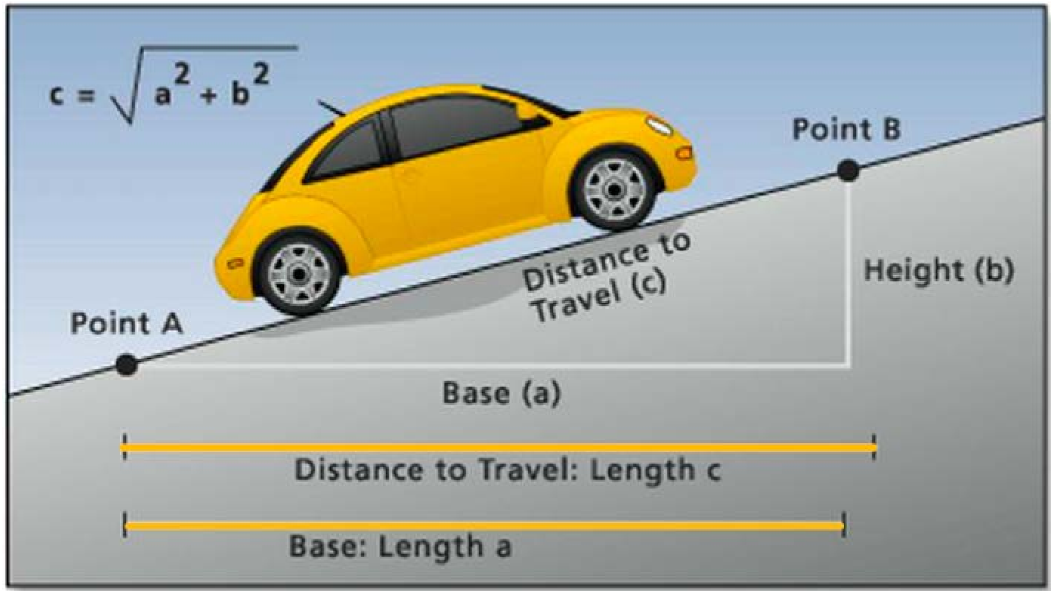


Usery *et al*. (2003) demonstrate that significant errors can occur with planar and geodesic distance and area that depend upon projection, raster resampling, and spatial resolution.

Distance on a 3D terrain surface

- Surface adjusted distance





➡a <> c !!

Why Distance Matters

- The most fundamental spatial metric
  - Travel time, travel cost, length of feature
  - Higher dimensionality metrics -> area and volume
  - Terrain derivatives -> slope, aspect, curvature
- Geospatial Analysis
  - Buffer, proximity analysis, shortest path, network
  - Spatial queries (select by location, near…)
  - Flow direction & accumulation
- Geostatistcs (spatial dependency)
  - Semi-variogram analysis
  - Kernel estimation
  - Kriging

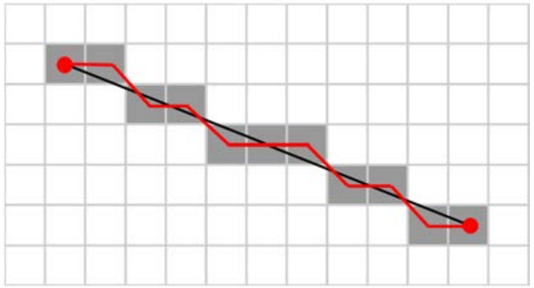
Problem Statement

- Planar and geodesic distance metrics introduce uncertainty to other spatial analyses
- Findings shown today indicate that these uncertainties increase with spatial resolution and vary with terrain type
- Current processing advances make it feasible to consider *surface-adjusted* distance metrics that account for terrain variations (instead of planar or geodesic)
- Surface adjusted metrics may reduce uncertainties and obviate the need to work at finest resolutions, given the scarcity of LiDAR data in rural and developing regions.
- Balance the reduction in uncertainty against the added computational intensity. This is a pilot demonstration of surface adjusted distance metrics.

Surface-Adjusted Distance Measurements

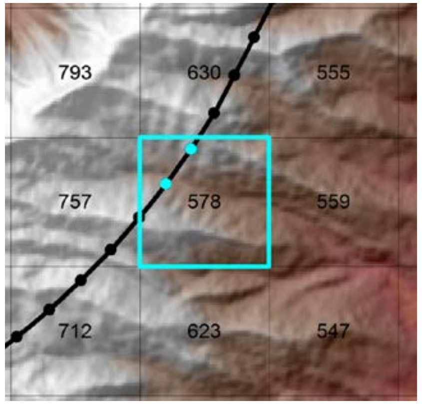
Measurements in DEM

- Pixel to pixel approach

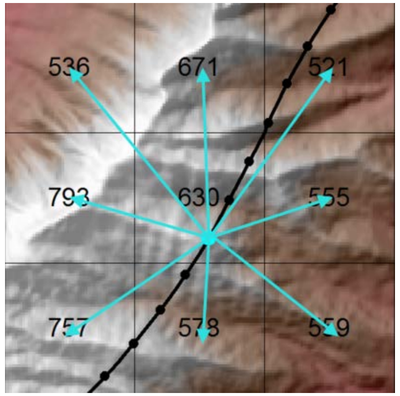


Bresenham's line (Bresenham 1965)

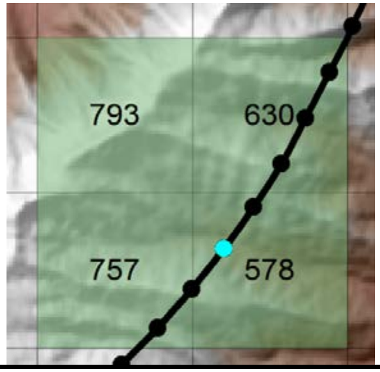
- Closest centroid



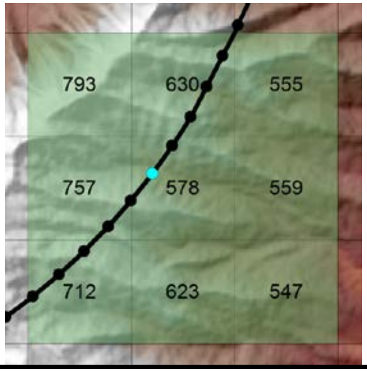
- Weighted Average



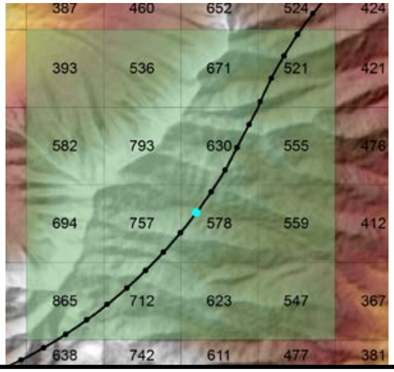
- Bi-Linear



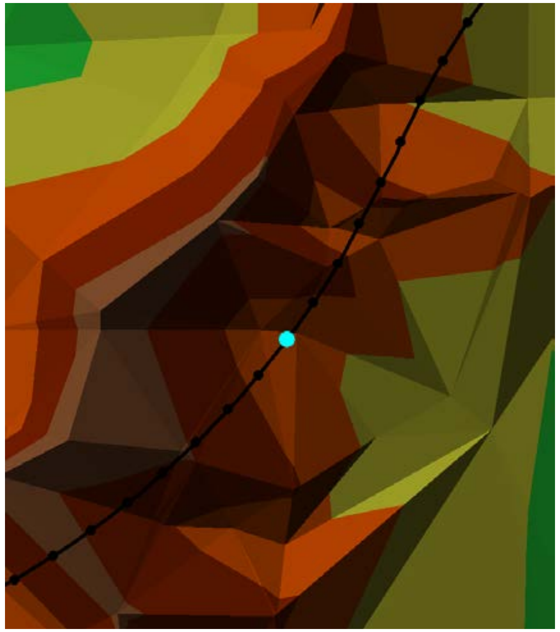
- Bi-Cubic



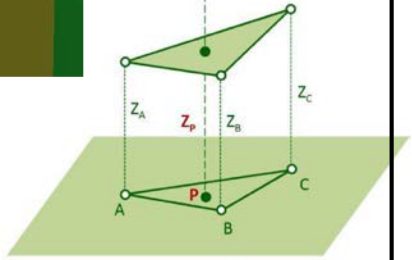
- Bi-Quadratic



Measurements in TIN



TIN

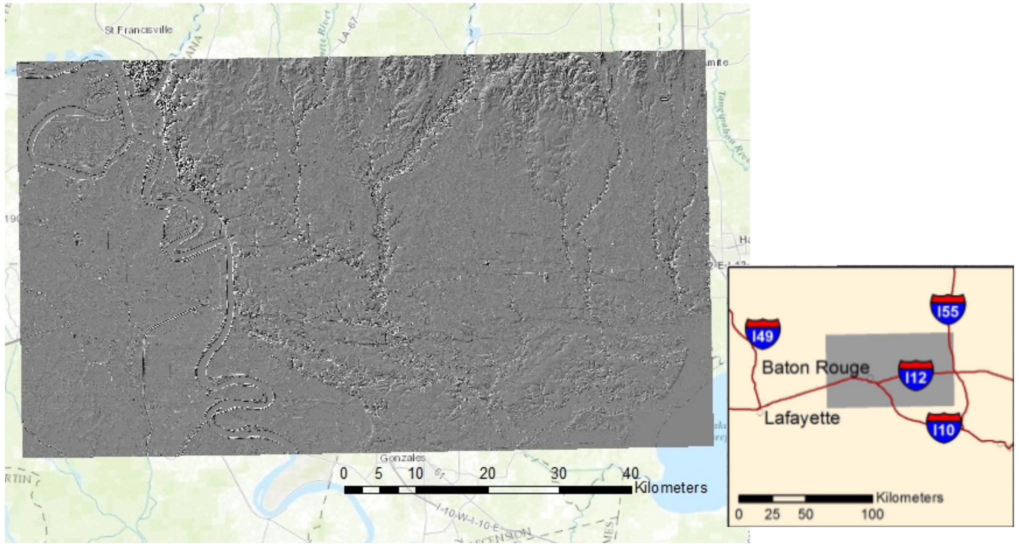


Methods

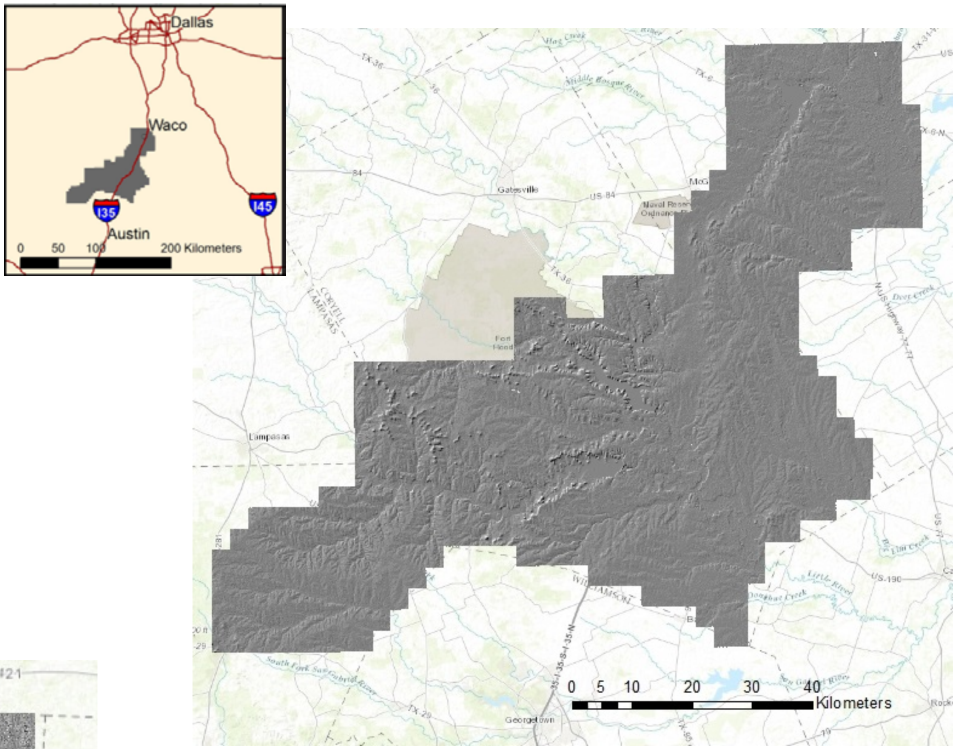
- Compare surface adjusted distance and benchmark:
  - **Benchmark:** distance measured at 3m LiDAR DEM (closest centroid)
  - **Surface-Adjusted Distance:**
    - Using different surface-adjusted methods
    - In DEMs with different resolutions (10m, 30m, 100m, 1000m)
    - In study areas of different landscapes
  - **To solve the function:**

$$\text{Measurement error} = f\,(\text{DEM resolution, Methods, Landscape})$$

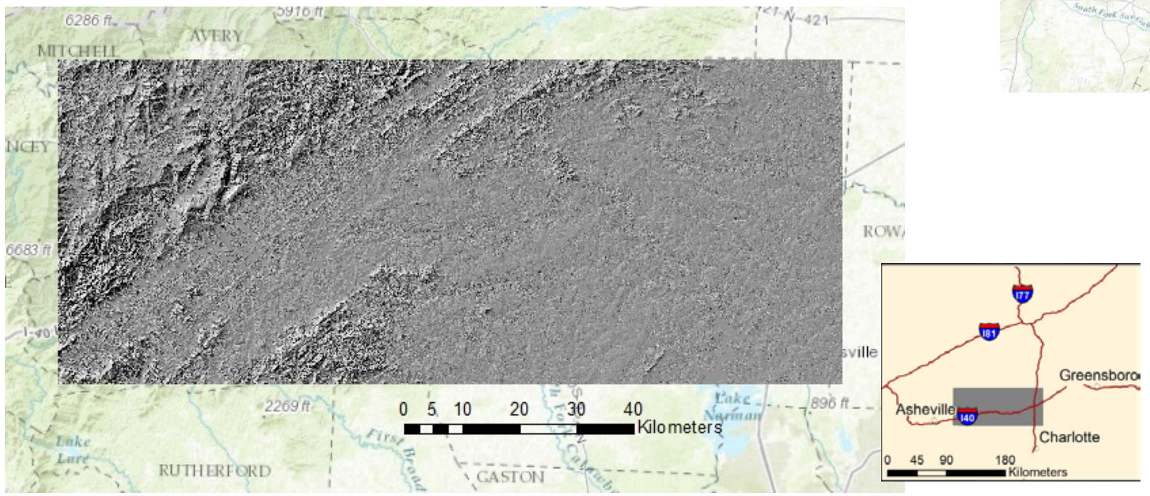
Study Areas



Louisiana

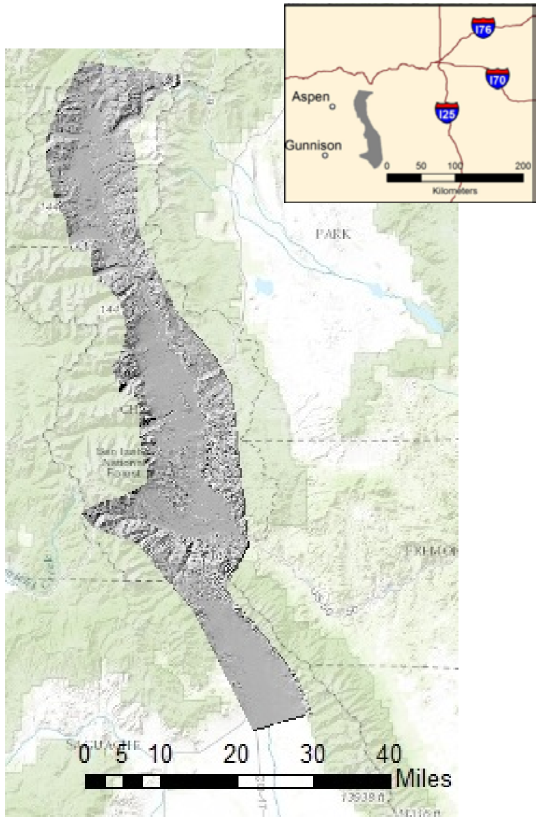


Texas

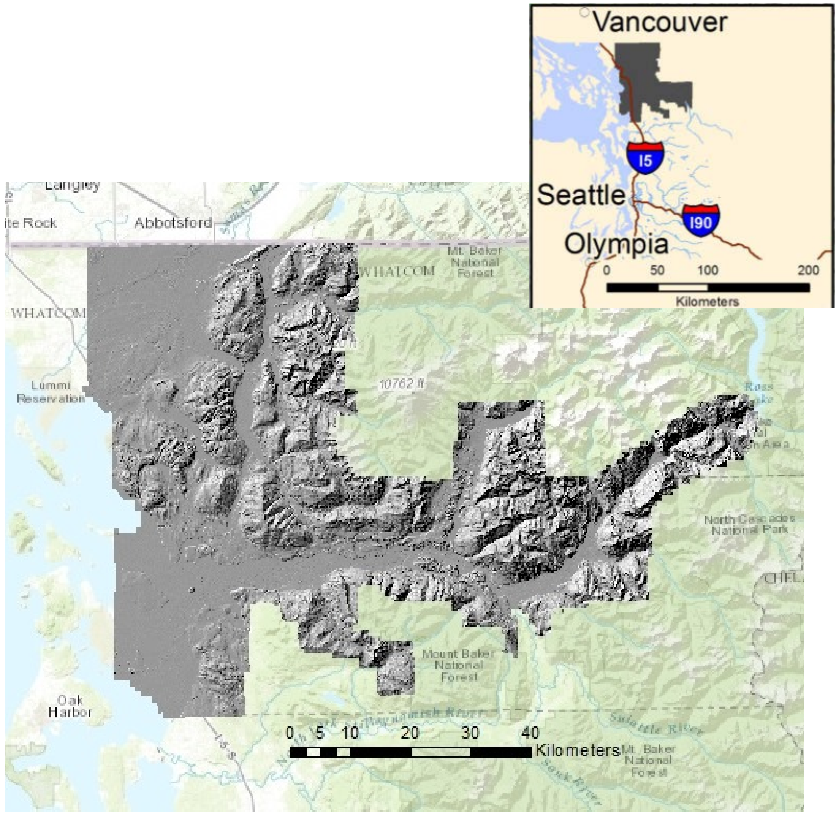


North Carolina

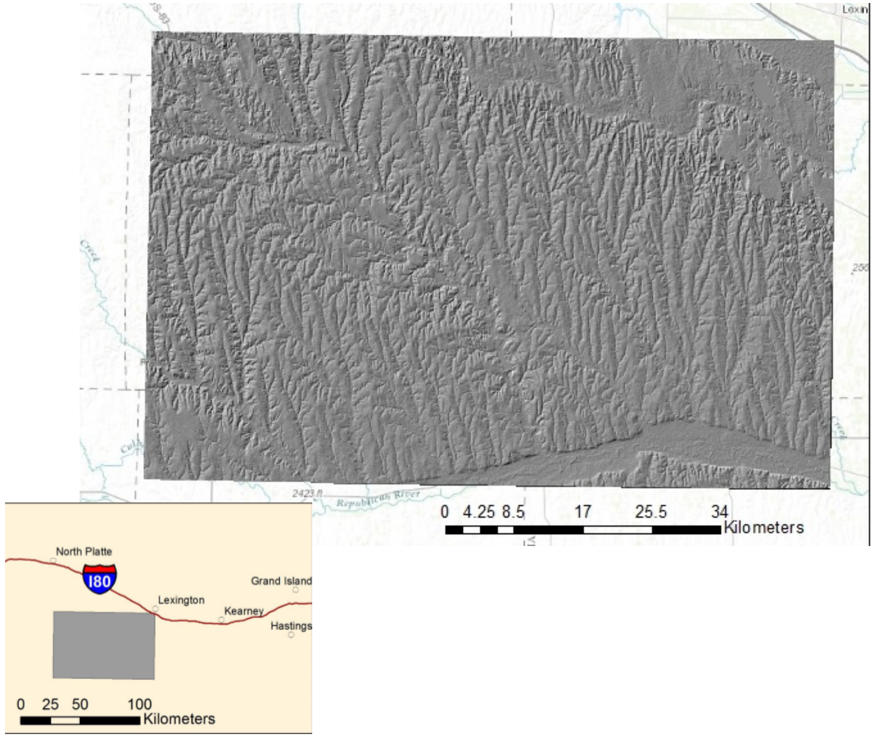
Six Study Areas



Colorado



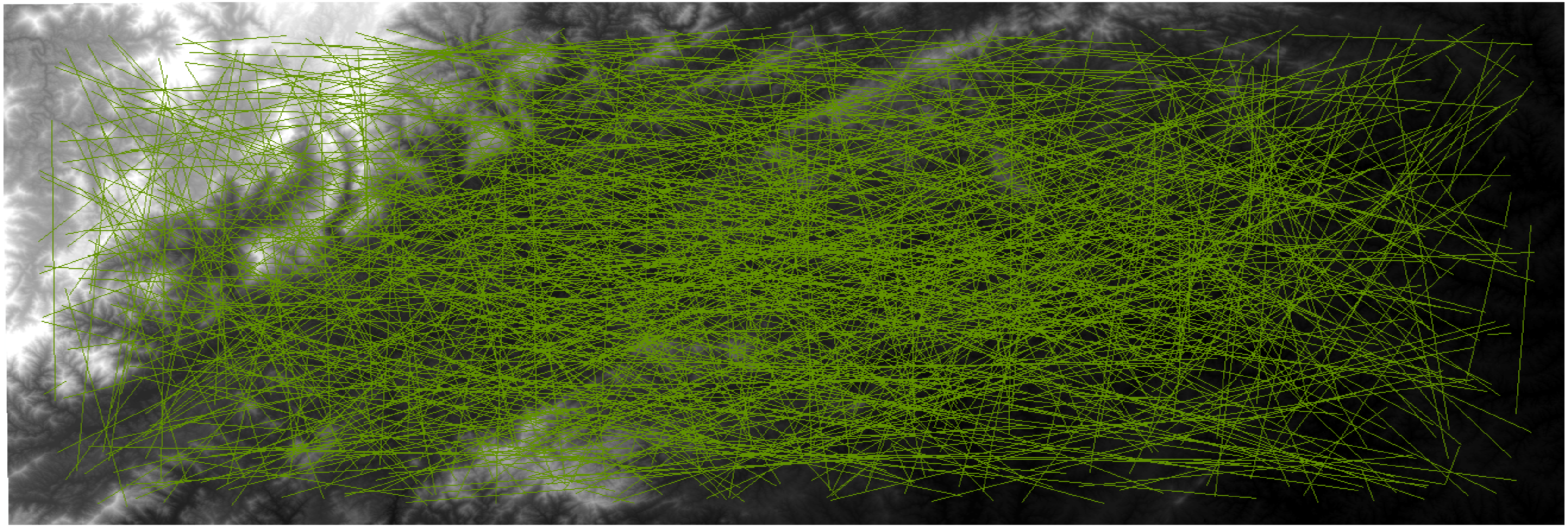
Washington



Nebraska

Simulation

1000 random transects per study area



Simulation

Nested ‘for’ loops:

```
Lv 1: for study_area in list_study_area (len=6):
Lv 2:     for DEM in list_DEM (len=4):
Lv 3:         for method in list_method (len=7):          <- 168 independent processes
Lv 4:             for transect in list_Transect (len=1000):  <- 168,000 independent processes
Lv 5:                 for point_pair in all_point_pairs (len=30 - 4000):
                          calculate distance between each pair of points:
```

Data size: 3.43GB for the largest study area

Estimated computation time for the simulation using sequential processing: 400 hours

Estimated computation time when parallelized at Lv 3: 2.38 hours, Lv 4: 8.6 seconds

Implementation

- ipyparallel

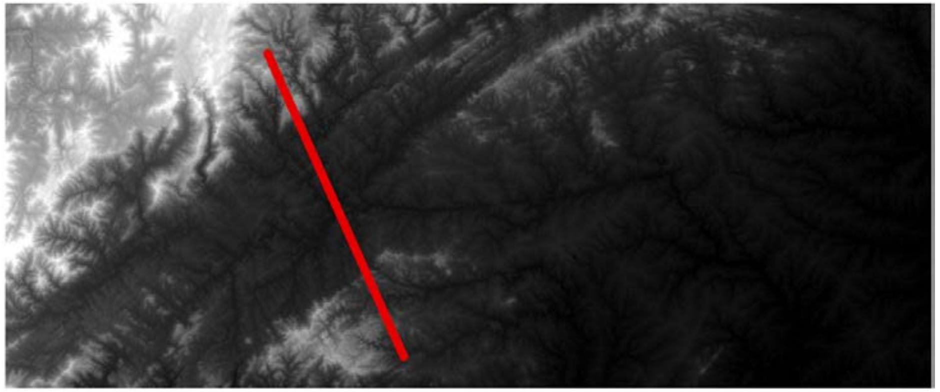
```
dist=lview.map(surfaceAdjusted.distance,
                cases_df['transect'].tolist(),
                cases_df['method'].tolist(),
                cases_df['resolution'].tolist(),
                cases_df['area'].tolist())
```

- Compute & storage in cloud
  - Data in Simple Storage Service (S3)
  - Parallel computing on Elastic Compute Cloud (EC2)
  - Containerized Jupyter notebook

amazon web services + docker

Technical Hurdles

- Using DEM buffers around transects instead of the whole DEM.



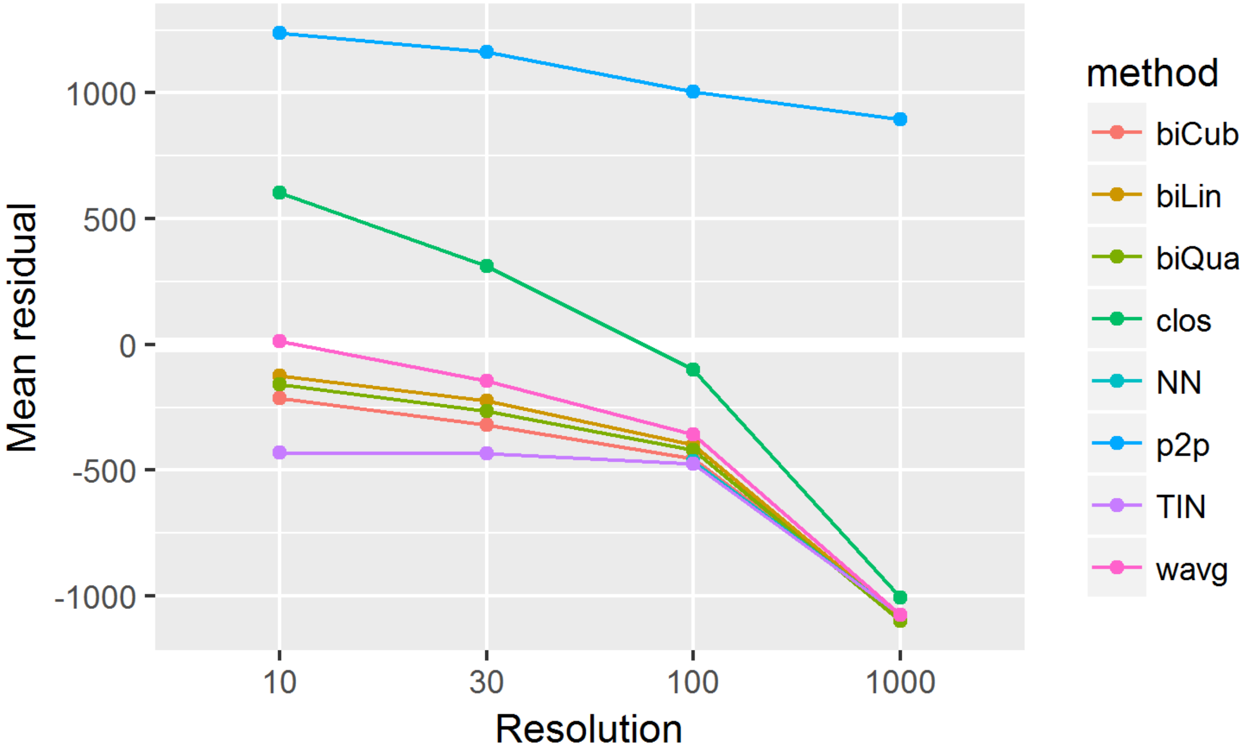
Data size of the entire DEM
- 3m resolution > 1Gb,
- 10m resolution > 100Mb

1000 transects quickly run out of memory

Buffer area around each transect
- Coded in csv files (X, Y, elevation)
- Data size of each csv: less than 1Mb

Preliminary result (Colorado)

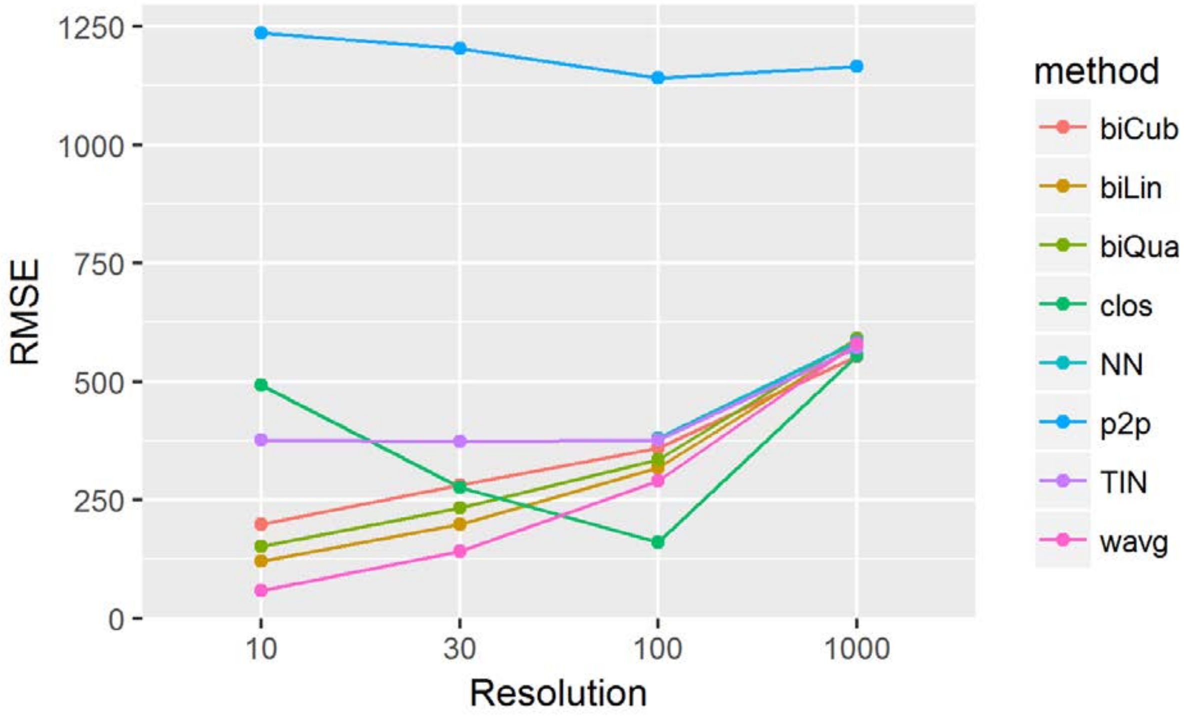
Average residual of different methods:
surface adjusted distance – benchmark distance (LiDAR)



- The residual of all distance measurements decrease as resolution becomes coarser.
- Except **p2p** and **clos**, other methods have similar performance
- **p2p** has the least accuracy (overestimating)
- **clos** crosses the zero-line (overestimating -> underestimating)
- The **TIN** method is stable under 100m resolution.

Preliminary result (Colorado)

Root-Mean-Square Error (RMSE) of different methods



		Resolution			
		10m	30m	100m	1000m
Method	biLin	120.45	197.62	319.45	604.03
	biQua	151.20	233.70	334.68	562.90
	biCub	197.75	280.08	359.52	595.50
	clos	491.39	273.77	161.78	549.99
	p2p	1235.72	1203.58	1140.67	1164.72
	TIN	375.67	371.35	375.98	580.40
	wavg	57.80	142.09	289.31	575.62

The most accurate method at different resolution

University of Colorado Boulder

University of Hawai'i Mānoa

Result Summary

- Results show distance measurement depends on DEM resolution and type of surface adjustment.
  - Surface adjusted distance computation tends to underestimate the distance as the DEM resolution coarsens (as expected).
  - **Weighted average** is the most accurate method in the finer resolutions (10m and 30m), whereas **closest centroid** is the most accurate method at the coarser resolutions (100m and 1000m).
  - The polynomial methods have a similar changing trend in different DEM resolutions. Bilinear (fewest neighborhood pixels) is the best polynomial surface adjust method. BiQuadratic is the least accurate (most neighborhood pixels).
  - At the coarsest resolution (1000m), adjustment methods give similar results.
  - The TIN method is very stable at the finer resolutions.

Conclusions and Future Work

- The premise of this work is that chronic mis-estimation resulting from planar distance metrics can impact models that rely on terrain for science, planning and decision support.
- One goal is to determine if particular surface adjusted metrics can obviate the need to work at finest resolutions, given the scarcity of LiDAR data in rural and developing regions and the computational intensity.
- This work demonstrates that the supercomputing techniques can create opportunities to answer fundamental GIS questions that cannot be answered using sequential computing.
- Implementing in Amazon cloud computing and testing additional terrain types – ongoing work.
- A guideline will be created for the choice of surface-adjusted methods at different analytical tasks.

Acknowledgements

- This research is partially supported by the Earth Lab, one of two Grand Challenge Initiatives funded by the University of Colorado (http://www.colorado.edu/grandchallenges/)
- We also acknowledge the USGS Center for Excellence in Geospatial Information Science (CEGIS) for providing data and analytic advice.

Questions are welcome!