Improving Digital Elevation Model Accuracy in Colorado through Machine Learning

Eric Gosnell

Abstract

The United States Geological Survey 3D Elevation Program (USGS 3DEP) provides high-resolution digital elevation models (DEMs) crucial for various applications in civil engineering, education, and outdoor activities. While the USGS 3DEP DEMs offer valuable insights, they suffer from errors and inaccuracies. This paper proposes a solution to improve both the accuracy and resolution of DEMs using lidar sensing drones and machine learning techniques. The proposed approach involves collecting lidar data from diverse terrain types across Colorado, processing it to match USGS 3DEP data, and training a machine learning model to generate high-resolution DEMs. The model incorporates terrain characteristic metrics such as elevation, steepness, rock type, and tree coverage to enhance the visual fidelity and accuracy of the DEMs. The accuracy of the algorithm is evaluated using drone-collected lidar data. Despite challenges such as cost, flight restrictions, and geographic limitations, the proposed approach offers a promising avenue for enhancing the quality of DEMs and providing valuable insights into terrain characteristics.

Keywords: digital elevation model, 3D Elevation Program, lidar, machine learning

Introduction

The United States Geological Survey 3D Elevation Program (USGS 3DEP) 1-meter accuracy digital elevation models (DEMs) for all of the conterminous United States, and 5-meter accuracy DEMs for Alaska. Over the past decade, the program has increased the coverage accuracy to match the demand of both the general public using digital mapping software and professionals using geographic information systems. Higher resolution data produces a more accurate model of the Earth, which has many professional applications in civil engineering, such as water flow management and housing development.

Furthermore it has practical applications for safer and more predictable exploring and adventuring in the backcountry, along with educational applications to learn about the world. However, despite the continued public interest in high resolution data, USGS is limited by resources and current technology. Flying planes across every part of the country is extremely expensive, and the lidar sensors are limited in the amount of data they can process at such speeds and distances from the ground. As such, obtaining very accurate models of the Earth, including terrain features such as trees and rocks is currently only feasible in high demand areas such as cities.

While the USGS 3DEP DEM is a highly accurate model of the real world, there are many errors and inaccuracies in the data. These include but are not limited to height inaccuracies across different flight paths, striping, quilting, and inaccuracies in error prone terrain such as cliffs, canyons, and heavily forested areas. Methods for detecting and correcting these errors in USGS 3DEP data have been developed such as multiple Fourier transform algorithms to reduce noise, the Delta Surface Fill algorithm to fill voids, mean profile filtering to average high and low deviations, and bilinear interpolation for better grid patterns.

While these methods are generally successful in creating an accurate seamless model, they all focus on correcting the existing data rather than improving the resolution. I am proposing a solution to not only correct errors in the terrain, but more importantly to increase the resolution of the data using only the existing 1-meter DEM.

Through the use of lidar sensing drones, I will be able to create a machine learning model which can process the USGS 3DEP data and output a probabilistic high resolution digital elevation model for any terrain in Colorado. This would significantly increase the visual similarity between the

real world and the 3D maps displayed on geographic information systems, which benefits all users by providing a better understanding of the terrain.

Methods

To develop the machine learning model, I must collect lidar data to train the model on, process that data to match USGS 3DEP data, train the model, then collect and process more lidar data to evaluate the accuracy of the algorithm.

Data Collection

In order to create a model that is accurate across all terrain types found throughout Colorado, data will be collected from 10 sites across the state that each exhibit terrain characteristics unique to that region. The primary metrics used in determining these regions are: elevation, steepness, rock type, and tree coverage including tree type. Elevation is important to consider because flora found at different elevations have different capacities to hold and produce soil. Higher elevation terrain will have a lower soil depth, along with higher wind speed, which leads to more exposed rock, while lower elevations have deeper soil and much smoother ground coverage. Similarly, steepness has an effect on the ruggedness of terrain, as steeper inclines will have more exposed rock and ledges. On near vertical inclines, such cliff faces or canyon walls, the USGS 3DEP data gathered by plane is prone to inaccuracies which must be considered when improving the surface quality. Rock type is another important metric as different types of rock erode in different ways. Gneiss and schist, commonly found in the Front Range are more prone to cracking and shattering, resulting in jagged terrain and scree fields. Sandstone and limestone, found more commonly from Gunnison to Grand Junction, erodes easily and smoothly

resulting in even terrain and a higher density of canyons.

Forests are very biodiverse and provide protection from wind, resulting in a much higher soil depth and smoother ground. The increased flora and strong root structure furthermore aid in reducing erosion. While all digital elevation models used in this search are bare-earth models, tree coverage and tree type are both important to consider for the effect they have when gathering data. The higher the density of tree coverage, the more lidar beams that are absorbed which leads to more noise and a higher error rate after filtering.

Once all 10 regions have been selected based on diversity and other factors discussed in the limitations section, they will be subdivided into two non-overlapping one square kilometer quadrants. One quadrant will be used for training the model and the other will be used for evaluating the machine learning model afterwards. While a higher number of larger regions would help to improve the accuracy and capability of the training model, due to financial and time constraints detailed below, I have determined that 20 total quadrants is a realistic quantity that both provides sufficient data while minimizing cost.

Various methods exist for obtaining a high resolution DEM, and the most practical methods for my research are manually surveying land or using lidar equipped drones. Each method has its respective costs and limitations. Manual surveys are extremely time intensive and highly prone to inaccuracies, particularly in areas that are difficult to access by foot. Meanwhile, lidar drones are expensive and limited by no fly zones and other flight restrictions. After analyzing the various methods of obtaining a high resolution DEM, it is clear that lidar equipped drones offer the easiest and most accurate data.

Data Preprocessing

The data collected from the drone will form a lidar point cloud, which must be heavily processed to form a seamless bare-earth digital elevation model. For consistency, I will use the same methods used by the United States Geological Survey. This process first involves removing outliers which are more than 3 standard deviations from the mean profile filtered value for that coordinate. The removed point will be replaced with the expected mean value. The dataset must then be void filled using the Delta Surface Fill algorithm, though due to the slow and precise nature of drone based lidar data acquisition there should be little to no voids in the data.

In areas with forests or dense foliage, the ground points must be classified to create the bare-earth DEM, and any points not on the ground must be removed and void filled. Finally, I will use Fourier analysis techniques alongside density-based clustering algorithms to identify and filter out any noise, non-ground points, and remaining errors in the data.

Next a grid must be applied to the lidar point cloud data, connecting the points together in an even manner. The grid length must be a factor of the length used in the USGS 3DEP DEM grids so that specific points can be aligned to match the data in later steps. Ground points can then be interpolated onto the grid to create a 3 dimensional polygonal surface that connects each lidar point, which can then be smoothed to reduce ridges and irregularities, creating a more realistic representation of the true surface. This data can then be exported to a DEM and manually reviewed and corrected. If the total sum of lidar points across the one square kilometer quadrant is too large to be computed efficiently, the quadrant can be further split into multiple DEMs and mosaicked together to create one seamless model in the final output.

Once the digital elevation model is complete, the values can be compared to known points on the intersections of USGS 3DEP grids, and the height of my DEM can be adjusted to match these values. This is an important step in maintaining the accuracy of the final output, as all data points from USGS must remain the same so as to not reduce accuracy.

Data Processing

To train the machine learning model, I will input the 10 high resolution DEMs gathered by drone, along with the USGS 3DEP DEM for the corresponding training quadrants. Additionally, I will provide the four terrain characteristic metrics described above for the corresponding quadrants, along with the coordinates of the quadrant. This is used to form connections between the metrics and the characteristics of the DEM, which can then both be linked to a certain section of the state. This is so that after training, the machine learning model will then be able to predict the expected characteristics for any region based on a distance based weighted average of the characteristics at nearby training quadrants. In other words, an area near a rocky, jagged training quadrant will exhibit more rocky and jagged terrain, while an area near a flat smooth quadrant will be flatter and smoother.

For the machine learning model itself, I will use logistic regression along with a clustering propensity model to generate the terrain characteristics. The clustering propensity model calculates the probability that a certain data point will belong to a specific cluster, which can be used to vary the terrain generation based on the four terrain characteristic metrics. This conversion from low resolution to high resolution can be repeated 10,000 times across multiple generations until the algorithm is capable of producing realistic and consistent digital elevation models.

Accuracy Assessment

Once the training process is complete, the machine learning model will be able to take in an USGS 3DEP DEM quadrant as an input and transform it into a much higher resolution DEM similar to those gathered by drone. To test the accuracy of the algorithm, I will take the 10 testing quadrant DEMs that were gathered by drone and are known to be accurate. The 10 corresponding quadrants from the USGS 3DEP DEM can then be passed into the machine learning model, along with the location of the quadrants. In order to pass the test, all information in the input DEM must persist to the output DEM so that to ensure no information was lost in the process. The output DEM can then be compared to the true surface DEM gathered by drone and evaluated based on certain qualities that correspond to the terrain characteristic metrics. These include the frequency, size, and jaggedness of exposed rock, ruggedness of terrain, and gradient dependent deformities such as cliff banding on steep terrain. The output DEM will also be evaluated by averaging the height difference between each point in the output DEM and the true surface DEM. In the case that any of these 3 primary categories of evaluation fail significantly, the model can be retrained and improved based on the mode of failure until an acceptable machine learning algorithm is created.

Cost and Limitations

As mentioned above, lidar drones are expensive, each costing upwards of $10,000 for the drone, lidar sensor, GPS, and other onboard sensors. There are other options however, including lidar drone rental services which charge between $300 and $600 per day depending on the sensor and drone quality. The drone is needed to survey 20 square kilometers of land spread out throughout the state. Including both the time

to survey the land, and travel between locations, I estimate that it will take 5 days to collect all necessary data. At an average cost of $450/day, the drone rental total will be $2250. Adding in remaining costs of travel across the state, namely for gas, the total cost required to carry out this research is $2500.

The next major limitation on this research is drone flight restrictions across the state. Locating 10 unique sites and obtaining drone flight and lidar survey permissions for those regions may prove to be difficult. Wilderness areas and national parks in Colorado do not allow for take-off, operation, or landing of drones and so these areas must be ruled out. In the national forests, drones are limited to flight below 400 feet, and weighing less than 55 pounds. Finally, on any privately owned property, land owner permission must be obtained before being able to fly over the land.

Once data for all 20 quadrants have been gathered and the machine learning algorithm has been created, it will still be geographically limited in its capabilities. Since all training and testing locations are in Colorado, the model is only known to be accurate for Colorado. While the machine learning algorithm can certainly take in a USGS 3DEP DEM and terrain characteristics from somewhere outside Colorado, the accuracy will decrease the distance and variance in the terrain characteristics. To make this algorithm capable of transforming all areas in the US, many more regions would need to be sampled from across the country and the model would need additional training on them. Seeing as the demand for higher resolution is nation-wide this would be an ideal final product, however the scope is too large, time intensive, and costly for my research.

Another limitation with the produced digital elevation model is that while it contains all the data currently available, the higher resolution points that fill the gaps in between are simply a prediction of what is most likely to exist at that location, not what actually exists there. This may lead to scenarios where users are misled by the model, though since the USGS 3DEP data is already at such high resolution, the differences between my model and the real world will be on such a small scale that it should not have any serious impact. Despite this, I will still have a disclaimer on the product that it is generated using artificial intelligence and is not guaranteed to be accurate.

Conclusion

The novel method proposed will improve the quality of life of civil engineering professionals, outdoor enthusiasts, and any other user of 3D mapping technology. Despite the constraints and limitations, the machine learning model will be able to generate high quality terrain across all of Colorado, and the methods described above can be used to expand the model to other parts of the country.

Another avenue for improvement of the model is to include three dimensional non-ground based objects such as trees and houses or other manmade structures. Additionally, there can be coloring improvements to better match the satellite imagery to the newly generated high resolution terrain.

Personal Statement

As a Colorado native, I am an outdoor enthusiast and spend a significant amount of time in the backcountry, along with planning potential trips using geographic information systems. I have often found myself wishing it was higher resolution so I could get a

deeper understanding of the terrain. As a computer science student at the University of Colorado Boulder, I have the necessary skills to create the algorithm described above.

Works Cited

[1] Danielson, Gesch (2011) "Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010)" https://pubs.usgs.gov/of/2011/1073/pdf/of2011-1073.pdf

[2] Stoker, Miller (2022) "The Accuracy and Consistency of 3D Elevation Program Data: A Systematic Analysis" https://www.mdpi.com/2072-4292/14/4/940

[3] Callahan, Berber (2022) "Vertical accuracy of the USGS 3DEP program data: study cases in Fresno County and in Davis, California" https://www.proquest.com/docview/2644080479?sourcetype=Scholarly%20Journals

[4] Siddiqui (2011) "Automating the Correction of USGS Digital Elevation Models Using Fourier Analysis and the Mean Profile Filter" https://www.asprs.org/wp-content/uploads/2010/12/Siddiqui.pdf

[5] Berber, Ustun, Yetkit (2012) "Comparison of accuracy of GPS techniques" https://www.sciencedirect.com/science/article/pii/S0263224112001674

[6] Jung, Jung (2023) "A Scalable Method to Improve Large-Scale Lidar Topographic Differencing Results" https://www.mdpi.com/2072-4292/15/17/4289

[7] Gesch, Oimoen, Evans (2014) "Accuracy Assessment of the U.S. Geological Survey National Elevation Dataset, and Comparison with Other Large-Area Elevation Datasets—SRTM and ASTER" https://pubs.usgs.gov/of/2014/1008/pdf/ofr2014-1008.pdf