GEOtiled: A Scalable Workflow for Generating Large Datasets of High-Resolution Terrain Parameters

Camila Roa
Paula Olaya
University of Tennessee
Knoxville, TN, USA

Ricardo Llamas
Rodrigo Vargas
University of Delaware
Newark, DE, USA

Michela Taufer
University of Tennessee
Knoxville, TN, USA

ABSTRACT

Terrain parameters such as slope, aspect, and hillshading are essential in various applications, including agriculture, forestry, and hydrology. However, generating high-resolution terrain parameters is computationally intensive, making it challenging to provide these value-added products to communities in need. We present a scalable workflow called GEOtiled that leverages data partitioning to accelerate the computation of terrain parameters from digital elevation models, while preserving accuracy. We assess our workflow in terms of its accuracy and wall time by comparing it to SAGA, which is highly accurate but slow to generate results, and to GDAL, which supports memory optimizations but not data parallelism. We obtain a coefficient of determination (R^2) between GEOtiled and SAGA of 0.794, ensuring accuracy in our terrain parameters. We achieve an X6 speedup compared to GDAL when generating the terrain parameters at a high-resolution (10 m) for the Contiguous United States (CONUS).

KEYWORDS

High Throughput Computing, Data Partitioning, Cloud Computing, Soil moisture

ACM Reference Format:
Camila Roa, Paula Olaya, Ricardo Llamas, Rodrigo Vargas, and Michela Taufer. 2023. GEOtiled: A Scalable Workflow for Generating Large Datasets of High-Resolution Terrain Parameters. In *Proceedings of the 32nd International Symposium on High-Performance Parallel and Distributed Computing (HPDC '23), June 16–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3588195.3595941

1 INTRODUCTION

Terrain parameters such as slope, aspect, and hillshading can be derived from a Digital Elevation Model (DEM) [1]. These parameters can be generated at different spatial resolutions and are fundamental in applications such as forestry and agriculture, hydrology, landscape ecology, land-atmosphere interactions, and soil moisture prediction [2, 3]. However, generating high-resolution terrain parameters from DEMs is computationally expensive, hindering their accessibility for multiple applications.

Two of the most commonly used Geographic Information Systems (GIS) to generate terrain parameters from DEMs are SAGA GIS [4] and GDAL [5]. Both are Free Open Source Software (FOSS). SAGA is widely used in the scientific community because it focuses on accuracy. However, it uses computational resources inefficiently since it loads all elevation data onto RAM memory to compute the terrain parameters. Given the potential high resolution of DEMs (e.g., 30m, 10m, 3m), the computation becomes infeasible, considering that the size of the elevation model can surpass memory capacity for large geographical regions. On the other hand, GDAL is memory efficient because it loads the elevation data onto RAM memory in chunks while running the computation. However, generating terrain parameters from DEMs can be further optimized for high throughput computing (HTC) systems such as the cloud.

To address these challenges, we propose GEOtiled. This scalable workflow leverages data partitioning to distribute the computation of terrain parameters at a high resolution across the nodes of HTC systems while preserving accuracy and efficient memory usage. To this end, we partition the elevation data into tiles with neighborhood buffers that can be distributed into independent virtual machines (VM) instances of an HTC system. Each instance computes the terrain parameters for the assigned tile, and our workflow recomposes the tile patches into a spatial continuous. We assess the effectiveness of our workflow, GEOtiled, in terms of performance and accuracy by comparing the metrics with derived terrain parameters from SAGA and GDAL.

The contributions of this work are as follows: (a) We design a workflow that performs elevation data partitioning to exploit data parallelism in generating large datasets of high-resolution (i.e., 10 m) terrain parameters from a DEM. (b) We validate our workflow in terms of accuracy by comparing the parameters generated by our workflow to those generated by SAGA for the Contiguous United States (CONUS) region at 1 km resolution. (c) We demonstrate our workflow's effectiveness in terms of scalability by generating three critical terrain parameters (i.e., slope, aspect, hillshading) at 1 km and 10 m resolution for CONUS and comparing the performance of our method with SAGA and GDAL.

2 WORKFLOW COMPOSITION

GEOtiled comprises three stages: (i) the partition of the DEM of the region of interest (i.e., CONUS) into tiles, each with a buffer region; (ii) the computation of the terrain parameters for each tile; and finally, (iii) the generation of a mosaic for each parameter from the tiles by averaging the values of the pixels that overlap between the tiles (i.e., pixels within the buffer regions).

We use a DEM from the USGS 3D Elevation Program [1] as the input to our workflow. We pre-process the elevation data by re-projecting it to a coordinate system in metric units because it is required for SAGA and GDAL to compute terrain parameters. We

HPDC '23, June 16–23, 2023, Orlando, FL, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0155-9/23/06.
https://doi.org/10.1145/3588195.3595941

crop the DEM into a number of tiles that fit the number of processes and whose size fits the memory available. We add a buffer region to each to prevent boundary artifacts. The boundary artifacts happen because computation at a single pixel uses values from adjacent pixels; therefore, when there are no buffers, the accuracy of the computation process is impacted. We use GDAL to compute the terrain parameters on each independent tile. Last, we clean the repetitive information from the buffer regions by building a mosaic with the average values of the overlapping regions within the tiles.

Our workflow is composable: it deploys existing tools such as GDAL for the computation. At the same time, it orchestrates the three stages and optimizes the computation process by introducing tiles to perform parallel computation in an environment with multiple VMs or threads. In this way, we decrease the computation time and reduce the memory usage per VM by exploiting data-level parallelism while preventing the formation of boundary artifacts.

3 ACCURACY AND SCALABILITY STUDIES

To demonstrate the effectiveness of our workflow, we present an accuracy and performance scalability study for CONUS at different spatial resolutions.

Accuracy Study. We compare terrain parameters SAGA (reference) generated versus GEOtiled for CONUS at 1 km. We observe a coefficient of determination (R^2) of 0.794, and the distribution of values generated by GEOtiled follows SAGA's as shown in Figure 1, validating the accuracy of our workflow.

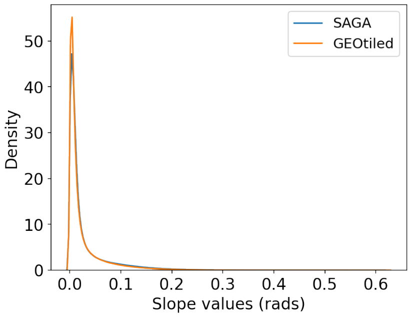

Figure 1: Distribution of slope values generated by SAGA and GEOtiled for CONUS at 1 km resolution.

Scalability Study. We compare the wall time it takes SAGA, GDAL, and our GEOtiled workflow to generate the three terrain parameters for two data scenarios in Table 1. For this study, we partition both input datasets into 400 tiles when we increase the resolution from 1 km to 10 m, the input and output data scale from hundreds of MB to TB. We test our workflow with 10 VMs on Jetstream 2, each with 8 CPU cores, 30GB RAM, and 60 GB disk.

Resolution	Number of tiles	Points per tile	Input data size	Output data size
1 km	400	37.9k	30.9 MB	69.7 MB
10 m	400	437M	341.1 GB	911.1 GB

Table 1: Data scenarios for CONUS at two resolutions.

Figure 2 shows the wall time for the slope computation; the most compute- and memory-demanding of the three parameters. We observe that the SAGA wall time is several orders larger than for GDAL and GEOtiled. Figure 3 compares the wall time for the slope computation using GEOtiled and GDAL. We do not consider SAGA because the required resources (RAM memory and time) are unfeasible. For 400 tiles, we achieve a speedup of nearly x6 with GEOtiled.

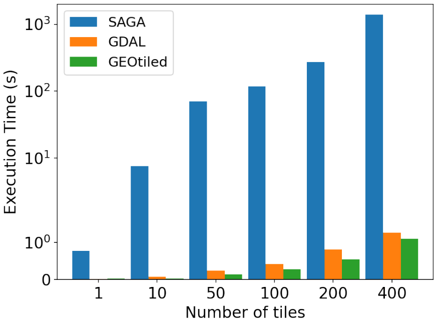

Figure 2: Wall time for SAGA, GDAL, and GEOtiled for calculating the slope for CONUS at 1 km resolution.

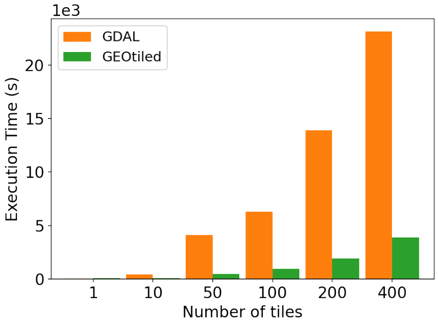

Figure 3: Wall time for GDAL and GEOtiled for the computation of slope for CONUS at 10 m resolution.

4 CONCLUSION

We present GEOtiled, a scalable workflow that generates large datasets of high-resolution terrain parameters. We demonstrate the effectiveness of our workflow by studying its accuracy and scalability performance compared to other GIS tools. We provide a Jupyter Notebook with our workflow, GEOtiled.ipynb, in https://github.com/TauferLab/SOMOSPIE.

ACKNOWLEDGMENTS

The authors acknowledge the support of the National Science Foundation through the awards #2103845 and #2103836; and the Access program through the NSF grant #2138296.

REFERENCES

[1] USGS, "3DEP: 3D Elevation Program." [Online; accessed 04-21-2023].
[2] D. Rorabaugh, M. Guevara, R. Llamas, J. Kitson, R. Vargas, and M. Taufer, "SOMOSPIE: A Modular SOil MOisture SPatial Inference Engine Based on Data-Driven Decisions," in *Proc. of the 2019 15th International Conference on eScience*, pp. 1–10, 2019.
[3] R. M. Llamas, L. Valera, P. Olaya, M. Taufer, and R. Vargas, "Downscaling Satellite Soil Moisture Using a Modular Spatial Inference Framework," *Remote Sensing*, vol. 14, no. 13, 2022.
[4] R. J.Böhner, O.Conrad and A.Ringeler, "SAGA: System for Automated Geoscientific Analyses." [Online; accessed 04-21-2023].
[5] Open Source Geospatial Foundation, "GDAL: Geospatial Data Abstraction Library." [Online; accessed 04-21-2023].