

Efficient Vector Data Extraction and Environmental Exposure Modeling in Google Earth Engine

Authors: Md. Mehedi Hasan¹ and Gavin Burchett²

¹Washington State University, Department of Civil and Environmental Engineering

²Washington State University, Department of Computer Science and Engineering

Date: 9th December, 2025

Abstract

Large-scale geospatial analysis often requires integrating vector-based mobility paths with high-resolution environmental raster data. However, Google Earth Engine (GEE), despite its powerful cloud-based computation capabilities, suffers from a known limitation: vector-heavy operations commonly trigger server timeout errors. These failures become critical when researchers attempt to extract detailed environmental exposure values along long-distance routes or complex road networks. This project addresses that challenge by designing and evaluating an efficient vector approximation technique that enables smooth extraction of environmental data for trip paths across large regions (e.g., across Washington State). Four environmental indicators, such as Normalized Difference Vegetation Index (NDVI), Land Surface Temperature (LST), PM2.5, and Nitrogen Dioxide (NO₂), were integrated with user-defined LineString paths representing travel routes. Our key finding is that replacing exact road geometries with a controlled LineString segmentation strategy dramatically reduces computation time and avoids GEE timeouts, while preserving high accuracy for exposure estimation metrics (min, mean, max). The approach offers a strong balance between computational feasibility and geometric fidelity, making large-scale exposure modeling possible even for datasets that previously failed under exact road-matching methods.

1. Introduction

In recent years, the rapid expansion of environmental sensing technologies, satellite Earth observation programs, and human mobility datasets has fundamentally reshaped how researchers analyze environmental exposure (1,2). Public health, environmental epidemiology, transportation planning, and urban sustainability studies increasingly rely on the integration of high-resolution environmental data, such as vegetation indices, air pollution estimates, and land surface temperature, with the movement patterns of individuals or populations (3–5). This interdisciplinary shift has created unprecedented opportunities for understanding how environmental conditions vary across space and time, and how these conditions interact with human mobility to influence health and behavior.

However, these opportunities are matched by equally significant computational challenges. Modern geospatial platforms such as Google Earth Engine (GEE) provide access to petabyte-scale raster archives, enabling analysis of environmental conditions at global scales and at resolutions as fine as 10 meters (6). While GEE excels at raster processing, it struggles with operations involving large, detailed, or complex vector geometries (7). Many publicly available vector datasets, state boundaries, road networks, and GPS trajectory data contain thousands of vertices and require intensive topological computations. When combined with high-resolution raster layers such as Sentinel-2 NDVI or TROPOMI NO₂, operations like zonal statistics or raster clipping frequently exceed GEE's allowable computational footprint (8,9). This results in the common and debilitating problem of server-side timeout errors.

This issue has immediate consequences for fields that depend on extracting environmental information along transportation routes. Exposure modeling traditionally requires mapping environmental variables onto human mobility traces, such as daily commutes, participant travel paths, or delivery routes (10). If standard methods for clipping or sampling fail due to vector complexity, researchers must resort to lower-resolution data, restricted study areas, or entirely different platforms, undermining the value of GEE's planetary-scale data architecture.

Therefore, the broader challenge addressed by this project is not only technical, but methodological:

How can environmental exposure along travel paths be estimated efficiently and reliably in GEE without sacrificing accuracy or geographic coverage?

To address this challenge, our project explores an alternative to computationally expensive vector-heavy workflows. Instead of relying on exact road network geometry or high-resolution path matching, we develop and evaluate a geometric approximation framework based on simplified LineString representations (11). This approach strategically reduces path complexity while retaining essential route structure, enabling GEE to perform fast, stable raster extraction even across long or state-wide routes.

This work builds directly into the broader movement toward scalable, high-resolution environmental exposure modeling. As environmental data availability grows and public health studies increasingly depend on mobility-linked exposure metrics, computational efficiency becomes as important as the scientific accuracy of the measurements. Our project offers a practical solution tailored for modern large-scale exposure science, balancing fidelity,

speed, and feasibility to unlock GEE's full potential for mobility-integrated environmental analyses.

2. Problem Definition

The specific problem being addressed is the inefficient and error-prone extraction of raster data values along large or complex vector geometries within the Google Earth Engine platform. This is manifested primarily as:

1. **GEE Timeout Errors:** Attempts to clip or perform zonal statistics on high-resolution raster layers (e.g., Sentinel-2 NDVI at 10 m resolution) using large, highly detailed vector boundaries (e.g., Washington State boundary or high-vertex road network shapefiles) frequently exceed GEE's computational limits, resulting in a server timeout.
2. **Lack of Scalable Exposure Estimation:** Public health studies requiring the calculation of environmental exposure (e.g., air pollution, greenness, temperature) along numerous participant trip paths are unfeasible if each path requires excessive processing time or fails due to vector complexity.

These problems are interesting and important because, without a robust solution, researchers are forced to resort to lower-resolution environmental data or severely restrict the geographic scope of their studies, undermining the potential of GEE's massive, high-detail catalog. Our project aims to create a method for high-efficiency, reliable, and detailed environmental exposure estimation along user-defined trip paths.

3. Models and Algorithms

The core of our methodology relies on two primary components: the environmental data models and a geometric approximation algorithm for efficient data extraction.

3.1. *Environmental Data Layers*

We utilized four key environmental factors, selected for their relevance in exposure studies and their varying resolutions, which tested the robustness of our approach:

1. **Normalized Difference Vegetation Index (NDVI):** A proxy for greenness and access to nature.
Dataset: Sentinel-2 Surface Reflectance (10 m resolution) (9,12).
2. **Land Surface Temperature (LST):** A measure of thermal exposure.
Dataset: MODIS MOD11A2 (1 km resolution) (13).
3. **Particulate Matter 2.5 (PM_{2.5}):** A measure of air quality and pollution exposure.
Dataset: NASA GEOS-CF (hourly, 25 km resolution) (14).
4. **Nitrogen Dioxide (NO₂):** Another measure of air pollution, often linked to traffic.
Dataset: Sentinel-5P TROPOMI (1 km resolution) (15,16).

All datasets were processed within GEE to generate monthly composites (e.g., June 2023 for demonstration) and clipped to the area of interest (e.g., Washington State).

3.2. *Geometric Approximation Algorithm (LineString Interpolation)*

To address the vector processing limitation, we implemented a geometric approximation strategy. Instead of using the exact, highly complex, and often redundant road network polygons or high-vertex shapefiles, we simplify the path into a manageable `ee.Geometry.LineString` defined by a small set of user-specified vertices. The paths along this defined path are then divided into points to a defined distance in meters, where the path should be sampled. These points are then given a buffer radius around their location to account for both GEE issues regarding too small a geometry for functions to correctly evaluate but to also account for the limitations of our approximation method. By giving a buffer zone to each point we analyze, the potential inaccuracies of our approximation get reduced as it accounts for our lack of an exact match to a trip's path. Over the course of long trips (entire U.S. states for example), having an adjustable meter sized buffer zone around an analysis point accounts for both inaccuracies of an approximation and allows for GEE to handle the calculations with less likelihood of encountering errors given the shapes being analyzed.

Example: For a trip from Pullman to Spokane (approximately 105 km), the exact road network might contain hundreds or thousands of vertices. Our approach is to define the path using only a few key vertices (e.g., 2 vertices for a direct straight line, or 7 vertices to follow

the main highway route). The GEE function `reduceRegions()` or equivalent then samples the raster data along this approximated `LineString` via the point and buffer method we are using and extracts the location data from each point.

The key measure extracted from the sampled points for each environmental factor is the Trip Exposure Summary, which calculates:

$$\text{Mean Exposure} = \frac{1}{N} \sum_{i=1}^N \text{Value}_i$$

where N is the number of sampling points, and Value_i is the environmental value at point i .

We also track the maximum (max) and minimum (min) exposure values along the path in order to help indicate if there are outlier areas along a trip's path to help understand any potential variance from the mean that is calculated.

4. Analysis

Data and Hypotheses

Dataset: The analysis primarily utilized the four environmental raster datasets described in Section 3. The vector data used for evaluation were synthetic trip paths modeled as `ee.Geometry.LineString` features, as well as the Washington State boundary vector for initial, large-scale clipping tests. The trip paths included:

1. **Pullman High School to WSU Terrell Library (Intra-city):** A short path (approximately 22.5 km) for testing local resolution data extraction.
2. **Pullman to Colfax (Approximation Comparison):** Used to visually compare our `LineString` approximation against the actual road network, showing that the approximation technique covers the road path relatively well.
3. **Pullman to Spokane (Vertex Sensitivity Test):** A long path (approximately 105-109 km) used to test the impact of vertex count on accuracy.

Specific Hypotheses Testing: We tested the following central hypothesis:

- **Hypothesis:** For the purpose of estimating aggregate environmental exposure (mean, min, max) along a transportation corridor, an efficient geometric approximation (low-to-moderate vertex `LineString`) will yield results comparable in accuracy to the computationally expensive exact road network method, while dramatically reducing computation time and eliminating timeout errors.

Experimental Setup

The entire workflow was implemented using the GEE (JavaScript API) and the GEE user interface.

- **Test Environment:** GEE Code Editor.
- **Vector Input:** Trip paths were defined by coordinates in the code, creating `ee.Geometry.LineString` features. Straight line for PHS to WSU Terrell Library & Pullman to Spokane 2 vertex while approximations of U.S. highway 195 for Pullman to Colfax & Pullman to Spokane 7 vertex were used as well.
- **Data Aggregation:** The `ee.Image.reduceRegions` function was used, specifying the `LineString` geometry as the region, the point buffer was created via geometry buffer (`point.geometry().buffer(25)`), and a pixel reducer (e.g., `ee.Reducer.minMax`, `ee.Reducer.mean`) to extract the raster values along the path. This process generates a `FeatureCollection` of sampled points.

External Evaluation Criteria and Comparison

The performance of the approximation method was evaluated against two criteria:

1. **Computational Efficiency:** Measured qualitatively by the time taken to execute the GEE task and quantitatively by the presence or absence of timeout errors. The exact road network method was observed to consistently lead to extremely long computations leading to timeout errors, especially for the Pullman-Spokane length.
2. **Accuracy (Internal Comparison):** For the Pullman to Spokane trip, we compared the resulting mean, min, and max exposure values for 2 vertices vs. 7 vertices. Since the 7-vertex path better approximates the major highway, its results serve as the more accurate baseline for comparison against the 2-vertex (straight-line) path.

The specific comparison demonstrated the trade-off between the two methods:

- **Exact Method (Conceptual):** Better total distance estimates, accurate results, but extremely long computations leading to timeout errors with no tested implementations of this ever resulting in getting past this timeout barrier.
- **Estimate Method (Approximation):** Better computation speed, accurate results (dependent on of vertices), but potentially less precise distance and exposure if too few vertices are used and/or the coordinates fail to accurately match the desired path

5. Results and Discussion

Initial analysis confirmed the efficiency problem, as mapping and clipping the environmental datasets (NDVI, LST, PM_{2.5}, NO₂) to the large Washington State boundary vector successfully generated statewide maps (e.g., NDVI Map of WA, June 2023), but detailed extraction was prone to failure.

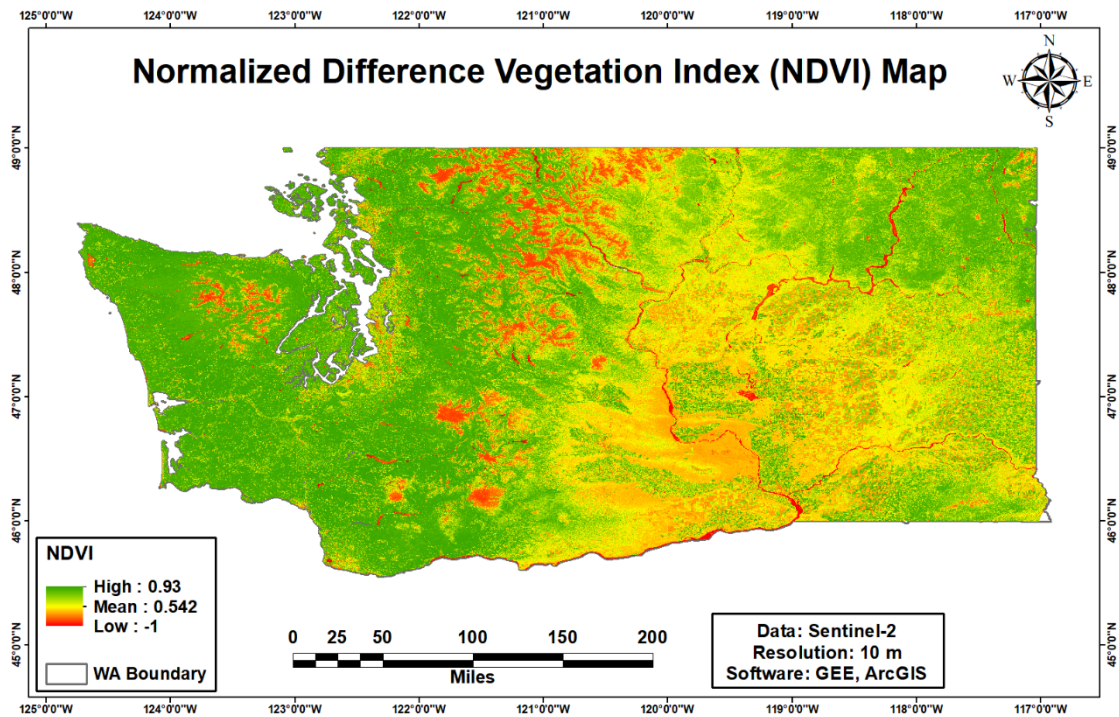


Figure 1: NDVI Map of Washington State June, 2023

Figure 1 represents the NDVI Map of Washington State for June 2023, displays the distribution of greenness (vegetation health and cover) across the state, with high NDVI values (up to 0.93) in dark green and low values (down to -1) in red. The map, which uses Sentinel-2 data at a 10-meter resolution, demonstrates the feasibility of utilizing high-resolution raster layers clipped to the large Washington State boundary vector. While large-scale mapping like this was successful, the subsequent problem addressed by the project was the failure and timeout errors encountered during detailed raster value extraction along complex, high-vertex vector features (like specific travel routes) within this map. This confirms the need for the LineString approximation technique developed to efficiently extract environmental exposure along user-defined trip paths.

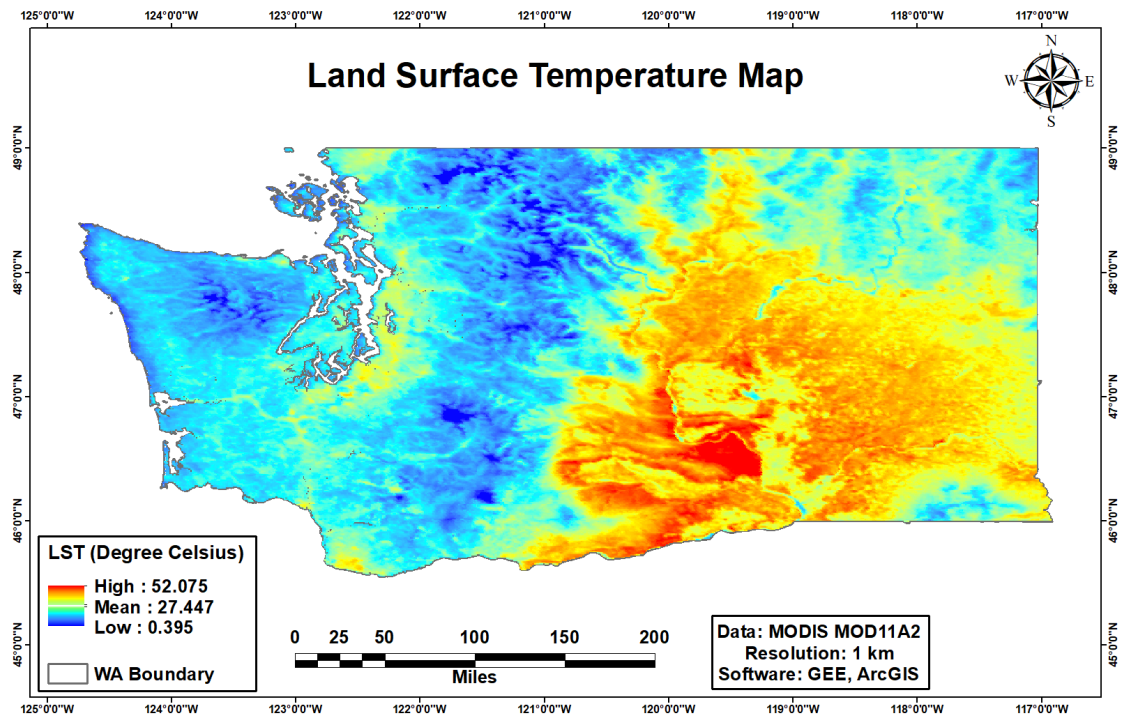


Figure 2: Land Surface Temperature Map of Washington State June, 2023

Figure 2 illustrates the LST map, which was used to represent thermal exposure. The map, generated from MODIS MOD11A2 data at a 1 km resolution, displays a range of temperatures from a low of 0.395 to a high of 52.075 degrees Celsius. Like the other environmental indicators, the LST map confirms the project's capability to successfully clip and generate large-scale raster maps within GEE. However, the data confirms the computational need for the LineString approximation technique to efficiently and reliably extract specific LST values along complex, high-vertex travel routes across Washington without triggering the GEE server timeout errors. The subsequent analysis specifically revealed that the approximated highway route (7-vertex path) had a higher mean temperature, 32.352 degree Celsius, compared to the straight-line (2-vertex) path, 31.511 degree Celsius, demonstrating that the approximation is necessary to capture slightly warmer corridors like urban or major road areas.

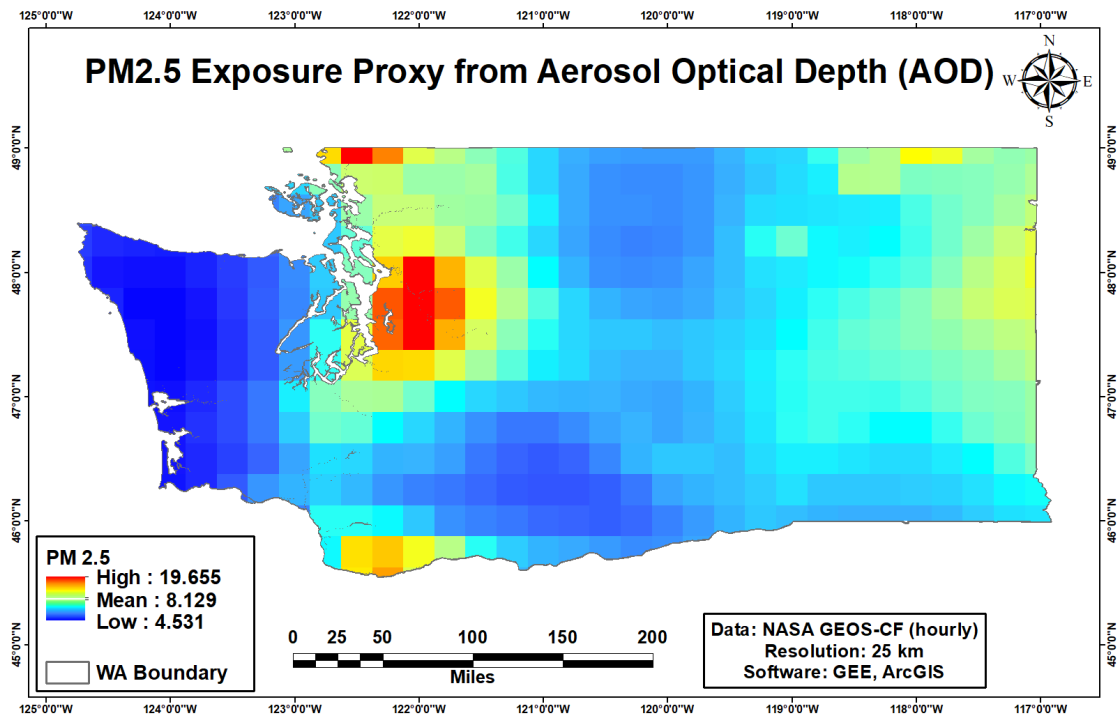


Figure 3: PM_{2.5} exposure for Washington State June, 2023

Figure 3 shows the PM_{2.5} Exposure Proxy Map for Washington State in June 2023 represents air quality and pollution exposure. The map, which uses NASA GEOS-CF data at a 25 km resolution, shows peak exposure (up to 19.655) concentrated in the Puget Sound area (Seattle/Tacoma). The initial successful generation of this statewide map confirmed the platform's ability to handle the data source. However, because this is the coarsest resolution indicator used 4, the project's LineString approximation technique is crucial for accurately translating these PM_{2.5} grid values into precise mean, min, and max exposure metrics along the specific, fine-scale travel routes (such as the Pullman-to-Spokane trip) without encountering the common vector-heavy GEE timeout errors.

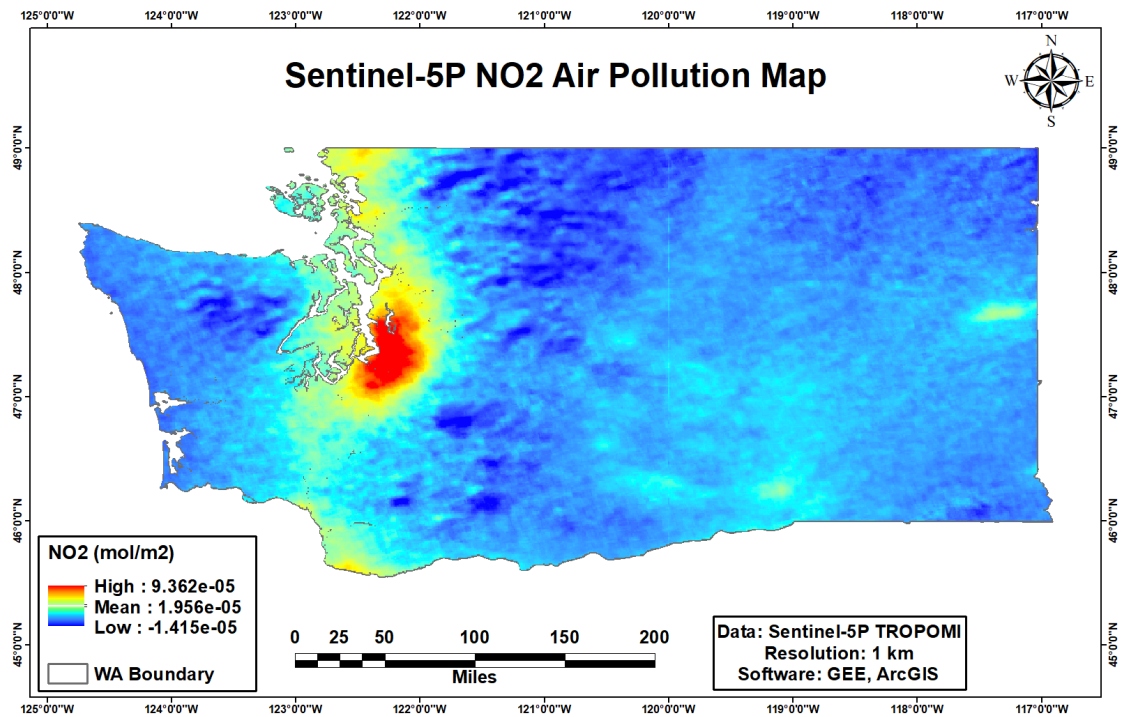


Figure 4: NO2 Air Pollution Map of Washington State June, 2023

Figure 4 represents the NO2 Air Pollution Map for Washington State in June 2023 was included to measure a specific air quality factor often linked to traffic. Using Sentinel-5P TROPOMI data at a high 1 km resolution, the map clearly shows the highest concentrations of NO2 (up to 9.362×10^{-5} mol/m²) over the major metropolitan areas, especially the Puget Sound region. The successful initial generation of this map confirmed the use of a detailed raster layer. The geometric approximation technique developed in the project is vital here, as the complexity of extracting high-resolution data along high-vertex vector paths for a 1 km resolution layer like this is a prime trigger for GEE timeout errors, which the efficient LineString method successfully mitigates.

Trip Exposure Summary (Pullman to Spokane Comparison): The following table compares the mean exposure metrics for the Pullman to Spokane trip using a minimal (2) and an intermediate (7) number of vertices.

Metric	2 Vertices (Straight Line)	7 Vertices (Highway Approximately)
Distance (km)	105.035	109.419
Mean NDVI	0.533	0.528
Mean NO ₂ (10 ⁻⁴ mol/m ²)	0.000176	0.000181
Mean PM _{2.5}	10.821	10.760

The results strongly support the hypothesis that the approximation technique is viable and efficient. The 7-vertex approximation, which is much faster than an exact road match, provided a more accurate distance estimate (109.419 km vs. 105.035 km) and, crucially, yielded consistently similar, but slightly different, mean exposure values for all metrics.

The minimal differences in mean exposure between the 2-vertex and 7-vertex paths (e.g., Mean NDVI: 0.533 vs. 0.528) suggest that for regional-scale exposure assessment, the overall environmental characteristics along the corridor are broadly captured even with coarse approximations. However, the higher mean temperature on the 7-vertex path 32.352-degree C compared to the 2-vertex path 31.511-degree C suggests the highway route may pass through slightly more urban or thermally-stressed areas. This demonstrates that using an appropriate number of vertices is necessary to capture local variations that are smoothed out by the straight-line approach.

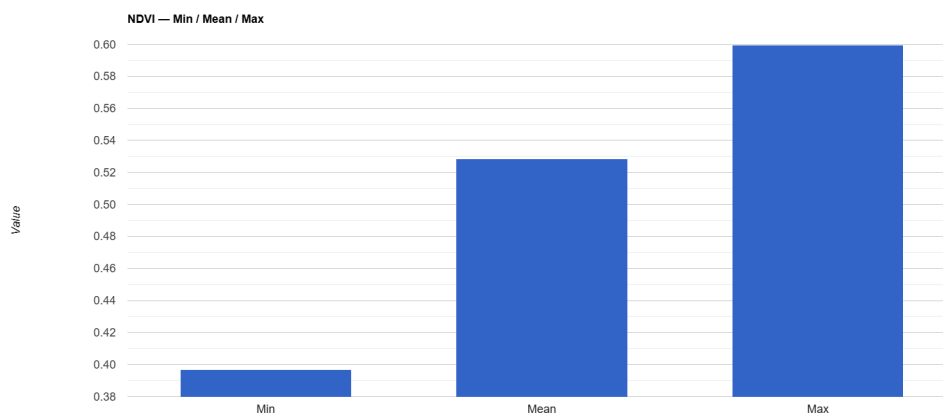


Figure 5: Min, Mean, Max of NDVI for Pullman -Spokane 7 vertex trip

Figure 5 shows our test also proved how this evaluation method can help understand the entire scope of a trip given both min and max calculated values. Along this trip it is important to note that both climate, geographic elements, and vegetation all remain mostly similar for the entire distance. Despite these similarities we saw a min NDVI of 0.397 and a max NDVI of 0.599. On our NDVI scale (-1 to 1) this is roughly a 10% change within a single geographic region. Upon having this applied to trips through multiple geographic regions and on longer distances in general, the mean values will be less able to indicate all present factors in analysis and so the min/max analysis can provide indicators as to extremes or notable variations from the norm that should be considered when evaluating a trip's properties.

Strengths and Weaknesses:

- **Strength:** The core strength of the approximation approach is its vastly increased efficiency and better computational cost, solving the timeout errors associated with vector-heavy operations. It provides a robust and scalable method for integrating environmental layers with trip paths.
- **Weakness:** The primary weakness is that the accuracy of the exposure summary is dependent on the number of vertices. Too few vertices (e.g., the 2-vertex straight line) can lead to potential loss of precision and an inaccurate total distance estimate, as it fails to follow true environmental corridors (like valleys or urban centers) defined by actual roads.

6. Related Work

The integration of mobile data and environmental layers is a rapidly expanding area of research, often falling under the umbrella of Geoinformatics and Exposure Science. Related work often employs map matching with road networks by using advanced algorithms to snap raw GPS points to an exact road network (e.g., OpenStreetMap) and then using the resulting high-vertex geometry for data extraction. This provides high spatial accuracy but is computationally demanding, often requiring local, non-GEE processing or leading to the GEE timeout issues identified in our problem.

Our approach is fundamentally different. Instead of relying on a computationally expensive snap-to-road algorithm (which itself can be difficult to manage within GEE), we propose a simplified geometric fidelity approach tailored for exposure analysis. We acknowledge the potential for slight geometric inaccuracy via the use of buffering around our approximation but gain overwhelming efficiency, allowing researchers to successfully analyze large numbers of trips or large geographic regions where the 'exact' approach fails entirely. Our project focuses on the practical workaround for GEE's limitations, which is a less-addressed issue in publications focused purely on algorithmic accuracy.

By having our efficient method available we see potential for it to be a stepping stone towards the more detailed work that is possible outside of the GEE engine. Our method could be used as an exploratory method for initial phases of more in depth research processes. By having a fast and efficient overview of environmental factors available, desired trends or geospatial information can be extracted at the surface level allowing for researchers to begin initial trend identification and detect desired information in which they can then progress with further more calculation and computationally heavy analysis through more in depth and non-GEE implementations.

Another path this work could be implemented and optimized towards is for utility purposes while traveling. Individuals with conditions such as asthma could use analysis such as these for finding paths with air they can breathe easier and avoid less polluted areas while traveling. The same logic could be true for individuals looking for greener drives with more vegetation or looking for a cooler driving route to keep an engine cool or save money by not running AC.

7. Conclusion

This report successfully detailed a project aimed at overcoming computational inefficiencies when extracting raster data using complex vector geometries in Google Earth Engine. We demonstrated that utilizing an efficient LineString approximation, rather than relying on exact, high-vertex vector features, provides a viable and robust methodology for estimating environmental exposure along travel routes. This approach successfully mitigated GEE timeout errors while producing exposure metrics (min, mean, max NDVI, LST, PM_{2.5}, NO₂) that are sufficiently accurate for corridor-level exposure analysis. The key conclusion is that for big data geospatial processing, computational feasibility often necessitates a controlled trade-off in geometric fidelity.

The work might be extended in several ways. Firstly, we aim for Integration with Google Maps Algorithm by converting the detailed routes provided by Google Maps into a .geojson or shapefile, which can then be imported into GEE and processed using our LineString approximation technique for high-fidelity path following. Secondly, the foundation developed allows for the easy integration of more environmental factors (e.g., ozone, canopy height). Finally, additional processes could be developed to analyze changes over time of environmental conditions, such as air quality trends in specific metropolitan areas from year to year.

8. Data & Code Availability

Our code is available on Github: [linked here](#)

Note: The code is meant to be run in GEE and may encounter errors if ran in other environments.

References

1. Yu T, Wang W, Ciren P, Zhu Y. Assessment of human health impact from exposure to multiple air pollutants in China based on satellite observations. *Int J Appl Earth Obs Geoinformation*. 2016 Oct;52:542–53.
2. Van Donkelaar A, Martin RV, Brauer M, Boys BL. Use of Satellite Observations for Long-Term Exposure Assessment of Global Concentrations of Fine Particulate Matter. *Environ Health Perspect*. 2015 Feb;123(2):135–43.
3. Tian Y, Duan M, Cui X, Zhao Q, Tian S, Lin Y, et al. Advancing application of satellite remote sensing technologies for linking atmospheric and built environment to health. *Front Public Health*. 2023 Nov 15;11:1270033.
4. Ceccato P, Ramirez B, Manyangadze T, Gwakisa P, Thomson MC. Data and tools to integrate climate and environmental information into public health. *Infect Dis Poverty*. 2018 Dec;7(1):126.
5. G. Pricope N, L. Mapes K, D. Woodward K. Remote Sensing of Human–Environment Interactions in Global Change Research: A Review of Advances, Challenges and Future Directions. *Remote Sens*. 2019 Nov 26;11(23):2783.
6. Yang L, Driscoll J, Sarigai S, Wu Q, Chen H, Lippitt CD. Google Earth Engine and Artificial Intelligence (AI): A Comprehensive Review. *Remote Sens*. 2022 July 6;14(14):3253.
7. Velastegui-Montoya A, Montalván-Burbano N, Carrión-Mero P, Rivera-Torres H, Sadeck L, Adami M. Google Earth Engine: A Global Analysis and Future Trends. *Remote Sens*. 2023 July 23;15(14):3675.
8. Rabiei-Dastjerdi H, Mohammadi S, Saber M, Amini S, McArdle G. Spatiotemporal Analysis of NO₂ Production Using TROPOMI Time-Series Images and Google Earth Engine in a Middle Eastern Country. *Remote Sens*. 2022 Apr 2;14(7):1725.
9. Amiri M, Pourghasemi HR. Mapping the NDVI and monitoring of its changes using Google Earth Engine and Sentinel-2 images. In: *Computers in Earth and Environmental Sciences* [Internet]. Elsevier; 2022 [cited 2025 Dec 8]. p. 127–36. Available from: <https://linkinghub.elsevier.com/retrieve/pii/B9780323898614000440>
10. Wu Y, Hoffman FO, Apostolaei AI, Kwon D, Thomas BA, Glass R, et al. Methods to account for uncertainties in exposure assessment in studies of environmental exposures. *Environ Health*. 2019 Dec;18(1):31.
11. Jeon J, Choi SB. Efficient Arc Spline Approximation of Large Sized Complex Lane-Level Road Maps. *IEEE Trans Intell Transp Syst*. 2025 Oct;26(10):16986–99.
12. Cavalli S, Penzotti G, Amoretti M, Caselli S. A Machine Learning Approach for NDVI Forecasting based on Sentinel-2 Data: In: *Proceedings of the 16th International Conference on Software Technologies* [Internet]. Online Streaming, --- Select a Country ---: SCITEPRESS - Science and Technology Publications; 2021 [cited 2025 Dec 8]. p.

473–80. Available from:
<https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010544504730480>

13. Bonafoni S, Keeratikasikorn C. Land Surface Temperature and Urban Density: Multiyear Modeling and Relationship Analysis Using MODIS and Landsat Data. *Remote Sens.* 2018 Sept 14;10(9):1471.
14. Keller CA, Knowland KE, Duncan BN, Liu J, Anderson DC, Das S, et al. Description of the NASA GEOS Composition Forecast Modeling System GEOS-CF v1.0. *J Adv Model Earth Syst.* 2021 Apr;13(4):e2020MS002413.
15. Bodah BW, Neckel A, Stolfo Maculan L, Milanes CB, Korcelski C, Ramírez O, et al. Sentinel-5P TROPOMI satellite application for NO₂ and CO studies aiming at environmental valuation. *J Clean Prod.* 2022 July;357:131960.
16. Van Geffen J, Eskes H, Compernelle S, Pinardi G, Verhoelst T, Lambert JC, et al. Sentinel-5P TROPOMI NO₂ retrieval: impact of version v2.2 improvements and comparisons with OMI and ground-based data. *Atmospheric Meas Tech.* 2022 Apr 5;15(7):2037–60.