Studying Biases in Google Street View (GSV): Analyzing Spatio-Temporal Patterns in Data Availability


Joseph Chen
University of California, Berkeley


Jingfeng Yang
University of California, Berkeley


Wenjing Yi
University of California, Berkeley


Jon Froehlich
University of Washington, Seattle


Michael Saugstad
University of Washington, Seattle


Adam Visokay
University of Washington, Seattle

PROJECT SIDEWALK

MAKEABILITY LAB

CREATE
Center for Research and Education on Accessible Technology and Experiences

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Berkeley College of Computing, Data Science, and Society

Introduction

GSV as a Tool: Google Street View's comprehensive coverage is a vital input for downstream urban planning, accessibility studies, environmental research, and machine learning.

Current Research Gap: The patterns and frequencies of GSV updates are not well-studied, leading to a lack of understanding about potential data biases.

Potential Biases:
- Spatio-temporal variability in GSV imagery affects systematic observation.
- Geographic disparities in image availability, as indicated by prior studies.

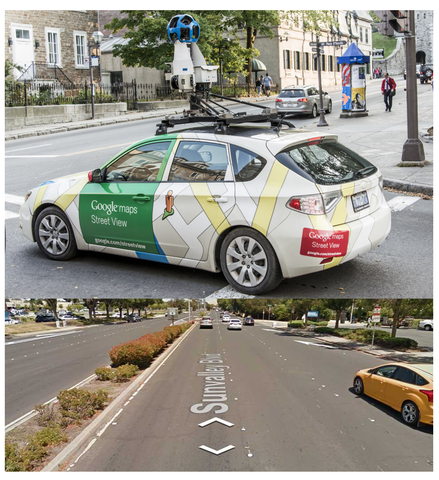
Project Aim:
- Analyze GSV's update mechanisms.
- Identify and understand potential biases in data collection.
- Equip the research community with insights on GSV data reliability for scientific analysis.

RESEARCH QUESTIONS

How might we access **GSV availability data** to examine **potential data collection biases** and find pockets of imagery or lack thereof?

How might GSV availability data relate to **neighborhood demographic characteristics?**

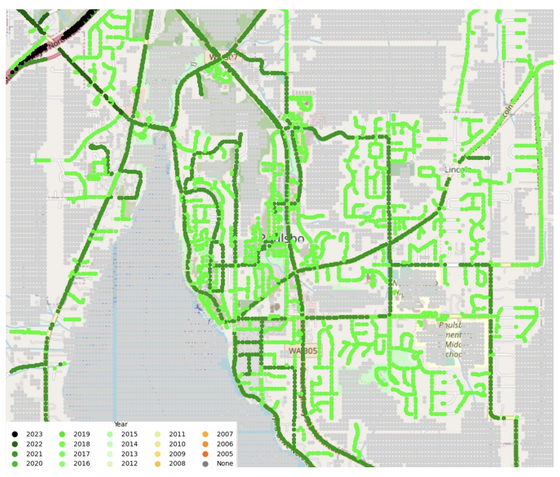
How do the **types of roads** within various neighborhoods correlate with the recency and frequency of GSV updates.



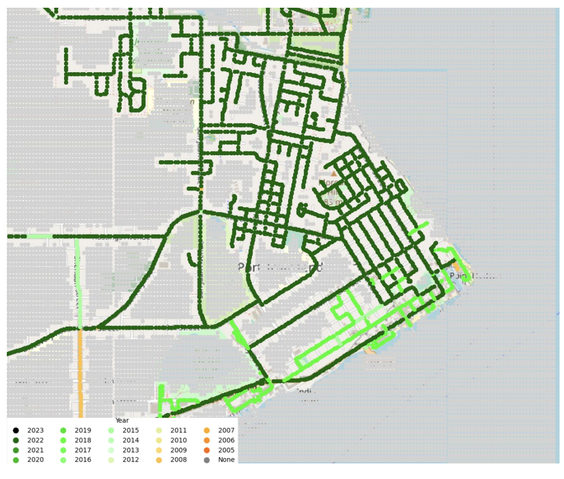
GSV DATA AVAILABILITY

Using Poulsbo, WA & Port Townsend, WA as examples

POULSBO



PORT TOWNSEND



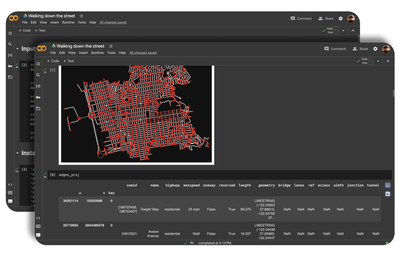
We used a scraper to download and visualize Google Street View (GSV) latest date availability data for Poulsbo and Port Townsend. The deeper the shade of green, the closer the GSV date is to the present, while gray indicates no available GSV images.

METHODOLOGY

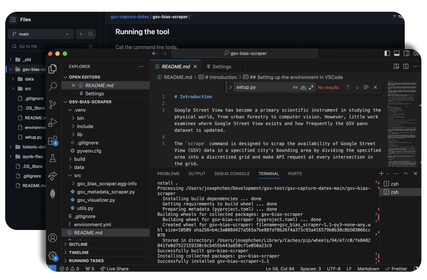
DATA COLLECTION

TOOL DEVELOPMENT

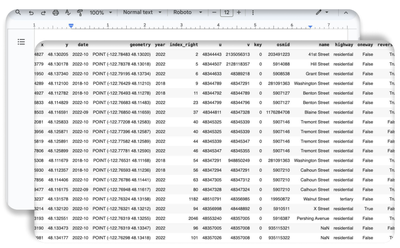
RESEARCH AND ANALYSIS



Utilized Google Colab Notebooks to streamline the systematic collection of GSV metadata, ensuring the process is both user-friendly and methodically clear



Created command-line tools to optimize the extraction of metadata, featuring customizable parameters for targeted city selection and precise data resolution preferences, as well as to view historical data availability, not just current



Selecting and analyzing numerous U.S. cities to understand how road network types, demographic variables, geographic location, and urban density affect GSV data availability

HYPOTHESIS TESTING PROCEDURE

Whether there is a significant difference in the recency of GSV data across different road types.

Null Hypothesis (H0): There is no significant difference in the recency of GSV data among different road types. This means that the variation observed in the dataset is due to random chance.
Alternative Hypothesis (H1): There is a significant difference in the recency of GSV data among different road types. This implies that the type of road (eg. residential, primary, secondary, etc.) has an impact on the recency of the GSV data.

	x	y	date	geometry	year	index_right	u	v	key	osmid	name	highway
0	-122.784827	48.130205	2022-10	POINT (-122.78483 48.13020)	2022	2	48344443	2135056313	0	203491223	41st Street	residential
1	-122.783779	48.130178	2022-10	POINT (-122.78378 48.13018)	2022	5	48344507	2128118357	0	5914088	Hill Street	residential
2	-122.791950	48.137340	2022-10	POINT (-122.79195 48.13734)	2022	6	48344633	48389218	0	5908538	Grant Street	residential
3	-122.764289	48.112100	2018-10	POINT (-122.76429 48.11210)	2018	9	48344789	48347291	0	281091363	Washington Street	residential
4	-122.764927	48.112782	2018-10	POINT (-122.76493 48.11278)	2018	11	48344792	48344789	0	5907127	Benton Street	residential

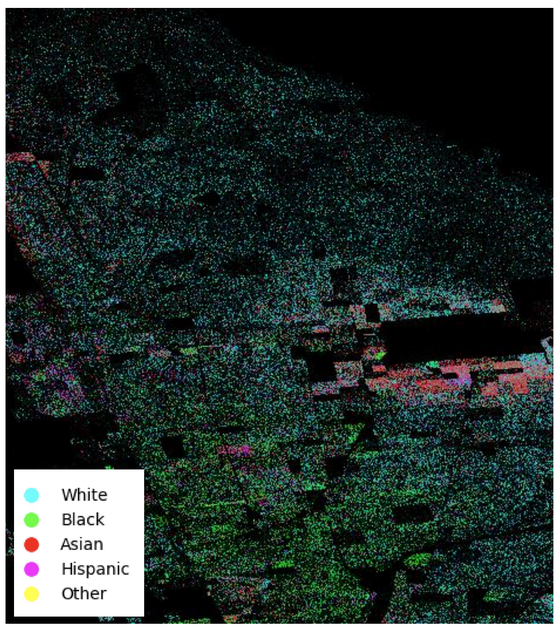
data frame used for analysis

Statistical Testing Procedure
1. Load Data
2. Data Pre-processing
3. Statistical Analysis
   - Shapiro-Wilk Test for Normality: use this test to check if the 'year' data within each road type group is normally distributed
   - Conduct a statistical test (Kruskal-Wallis) to see if there are significant differences in the recency of GSV data among different road types. Setting P-Value threshold as 0.05.
   - Non-Parametric Test
   - Comparing Medians Across Groups
   - Small & large Sample Sizes

RELATIONSHIP BETWEEN DEMOGRAPHIC AND GSV DATA

Using Berkeley as an example

DEMOGRAPHIC PLOT BASED ON US CENSUS



Using US Census data, we create a graph depicting the racial demographic composition of the entire population in various locations.

SCRAPED DATA FROM GSV BY YEAR

Employing our web scraping tool, we generate a plot that displays the years in which Google Street View (GSV) imagery was captured for various locations.

FINDINGS AND DISCUSSION

- In our study, we analyzed 20 cities in the USA, selected randomly from 2020 US Census data (10 urban and 10 rural), to examine the recency of Google Street View (GSV) data across different road types. Our findings consistently showed significant statistical differences in the recency of GSV data among road types, with no notable distinction between urban and rural areas. This conclusion is supported by p-values consistently below 0.05.
- Our analysis of demographic distributions in various cities revealed no discernible patterns in the recency of Google Street View (GSV) data when correlated with racial demographics. For visual reference, please see the accompanying graph to the left.

KEY TAKEAWAYS

- Our findings demonstrate significant differences in GSV availability and temporal recency even within a single city.
- As researchers—from sociology to urban planning to climate change—increasingly rely on GSV as a core study dataset, more work is necessary to investigate and transparently publish underlying biases in the data itself.

NEXT STEPS

- We are planning to build an interactive website in the future, where users can input the name of the city they are interested in. The website will then fetch the latest and historic dates data for the city's Google Street View (GSV) by using our well-designed scraper and provide data visualization. This will allow users to download the data locally for data processing and analysis.
- Expanding the Scope of Demographic Factors: Beyond racial demographics, analyzing other demographic variables like age, income levels, education, and occupation to see if they correlate with the recency of Google Street View (GSV) data.

REFERENCES

- Curtis, J.W., Curtis, A., Mapes, J. et al. Using google street view for systematic observation of the built environment: analysis of spatio-temporal instability of imagery dates. Int J Health Geogr 12, 53 (2013).
- Smith, Kaufman, & Mooney, Google street view image availability in the Bronx and San Diego, 2007–2020: Understanding potential biases in virtual audits of urban built environments, Health & Place, 2021
- Fry, Mooney, Roriguez, Caiaffa, Lovasi, Assessing Google Street View Image Availability in Latin American Cities, J. or Urban Health, 2020