PLOS ONE

RESEARCH ARTICLE

Open-source data pipeline for street-view images: A case study on community mobility during COVID-19 pandemic

Matthew Martell[1]☯*, Nick Terry[1]☯, Ribhu Sengupta[1], Chris Salazar[1], Nicole A. Errett[2], Scott B. Miles[3], Joseph Wartman[4], Youngjun Choe[1]

1 Industrial & Systems Engineering, University of Washington, Seattle, WA, United States of America,
2 Environmental & Occupational Health Sciences, University of Washington, Seattle, WA, United States of America, 3 Human Centered Design & Engineering, University of Washington, Seattle, WA, United States of America, 4 Civil & Environmental Engineering, University of Washington, Seattle, WA, United States of America

☯ These authors contributed equally to this work.
* marte292@uw.edu

Abstract

Street View Images (SVI) are a common source of valuable data for researchers. Researchers have used SVI data for estimating pedestrian volumes, demographic surveillance, and to better understand built and natural environments in cityscapes. However, the most common source of publicly available SVI data is Google Street View. Google Street View images are collected infrequently, making temporal analysis challenging, especially in low population density areas. Our main contribution is the development of an open-source data pipeline for processing 360-degree video recorded from a car-mounted camera. The video data is used to generate SVIs, which then can be used as an input for longitudinal analysis. We demonstrate the use of the pipeline by collecting an SVI dataset over a 38-month longitudinal survey of Seattle, WA, USA during the COVID-19 pandemic. The output of our pipeline is validated through statistical analyses of pedestrian traffic in the images. We confirm known results in the literature and provide new insights into outdoor pedestrian traffic patterns. This study demonstrates the feasibility and value of collecting and using SVI for research purposes beyond what is possible with currently available SVI data. Our methods and dataset represent a first of its kind longitudinal collection and application of SVI data for research purposes. Limitations and future improvements to the data pipeline and case study are also discussed.

Introduction

Street-level imagery is becoming an increasingly popular form of data for research [1]. Between 2009 and 2020, more than 200 publications utilized street-level imagery from corporate sources in urban research [1]. Out of all these sources, Google Street View's Street View Images (SVI) were the most popular among academics [1–3]. Uses for SVI data include

estimating demographics [4], evaluating the built environment [5], surveying plant species [6], measuring pedestrian volume [7], among many other applications [8–10].

While SVI data can provide many useful insights for researchers, it is not without its flaws. For corporate-collected images such as Google Street View, or Tencent Street View the availability of images depends on where the companies decide to collect data, while the accessibility of these images hinges on the companies' data provision policies. For example, there is no Google Street View service in most parts of Africa. An alternative to corporate-collected images are crowdsourced SVI databases such as Mapillary [11]. These crowdsourced images sometimes may have better coverage or temporal resolution than Google Street View, at the cost of varying image quality, field of view, and positional accuracy [3, 12]. Perhaps the largest challenge with SVI data is its temporal instability. Updates to these image datasets at specific locations are infrequent, especially in rural areas [1, 13, 14]. Additionally, images frequently are not collected at a consistent time of day, or season, even within the same city. These issues make existing SVI data unreliable for temporal studies.

Typically, temporal studies involving image data use images (or video) from fixed locations. This data is used to do things such as evaluate disaster recovery [15], monitor ecological change [16], or measure urban flooding [17]. Data from fixed cameras is also used to count people [18]. The challenge with these methods is that they are fixed-location. In order to collect spatial image data for these methods, frequently a large team is required to traverse areas on foot. This challenge, along with existing SVI data's temporal issues, demonstrate the potential value of collecting longitudinal SVI data.

Our main contribution is demonstrating the feasibility of collecting longitudinal SVI data. We demonstrate this through the creation of a complete data pipeline for conducting pedestrian counts using car-based street-level imagery. The pipeline accepts raw video collected by the camera as an input and outputs a record of each pedestrian detection and their locations (latitude and longitude). This approach allows for analysis of mobility patterns with high spatial resolution and a short lag time. It alleviates the quality and field of view inconsistencies that come with crowdsourcing SVI data [3, 12], generates data that is not corporately owned, eliminates the temporal instability challenge of both kinds of data [1, 13, 14], while still maintaining the advantages of SVI data over fixed-location methods [15].

Specifically, we use this pipeline to generate and analyze video from 37 video-collection runs in the city of Seattle, Washington, USA from May 2020 through July 2023. The video data was converted into over 4 million high-resolution images, with each data-collection run representing about 1.5 TB of image data. We used the images to create a record containing the location of each detected pedestrian, cross-referenced to the relevant GEOID [19]. To detect pedestrians in the still images, our pipeline leverages the state-of-the-art convolutional neural network, Pedestron [20]. We used the cascade_hrnet architecture benchmarked on the CrowdHuman data set [21]. Our methods and dataset represent a first of its kind longitudinal collection and application of SVI data for research purposes.

As a secondary contribution, we provide a case study based on the video data collected throughout the COVID-19 pandemic. We examine the effect of vaccine availability and local demographics on pedestrian detections, while accounting for weekly and yearly seasonality. Community mobility became a key metric during the height of the COVID-19 pandemic as government officials worked to halt the spread of the virus [22, 23]. Two of the largest and most widely used data sets for community mobility during this time were the Google Community Mobility Reports [24] and Apple Mobility Trends Reports [25]. Researchers used this data to study the incidence of COVID-19 in the US [26] and the effectiveness of government lockdown policies [27, 28], among other topics. Issues with these two data sets include mandatory opt-in, use of specific map applications, a lack of independent verification, and no long-term

data availability guarantees [26, 28–30]. Our findings demonstrate the utility of our data processing pipeline as an alternative for tracking community mobility over time and show the potential for its use in a variety of research domains.

Methods

Data collection

We collected our data as a part of the Seattle street-level imagery campaign, an ongoing series of video surveys for the purposes of documenting mobility throughout the COVID-19 pandemic. During each survey, a vehicle equipped with a 360˚ video camera is driven along a predefined route through Seattle while collecting video data and GPS metadata. The route incorporates broad neighborhood/area canvassing designed to collect data useful to multidisciplinary researchers as well as capital transects. Full details on the route design are available in Errett et al. [31]. The capital transects specifically target capitals (social, cultural, built, economic, and public health) which are theorized to be closely tied to community resilience [32]. Specific canvassing areas and capitals within Seattle were chosen to ensure a representative sample of the overall population of Seattle [31]. While the drivers try to make the surveys as consistent as possible, occasionally exogenous factors caused deviations from standard protocols. For example, during three of the surveys (05-29-2020, 06-18-2020, and 06-26-2020), protests over the murder of George Floyd caused parts of the survey route to be unnavigable.

After consulting with the University of Washington Human Subjects Division, it was determined that this study was not considered human subjects research and would not require IRB approval. The data we captured was people in public places, where they cannot expect personal privacy. As an added precaution, all data for this study was published through Mapillary [11], which automatically obscures faces.

Data processing pipeline

After video collection, the raw data is segmented into image data. The images are subsampled from video frames so that they are collected about every 4 meters. The images are then uploaded into the DesignSafe-CI Data Depot [33]. From DesignSafe, the images are transferred to the TACC Frontera high-performance computing cluster [34]. We completed all file transfers between the two services using Globus [35]. Without access to these services, or similar ones, the storage and computing requirements for this project would be intractable.

On Frontera, orthorectification is performed to the images, then pedestrian detection is performed on the orthorectified images. The orthorectification transforms the images from a single image in the equirectangular projection to two images in the rectilinear (gnomonic) projection [36]. Pedestrians are detected on each of the new images using a convolutional neural network (CNN) based on a pre-trained model from the Pedestron repository [20]. Our data represents a highly challenging detection task, as there is great variation in lighting, backgrounds, human poses, levels of occlusion and crowd density from image to image and run to run. The Cascade Mask R-CNN architecture in the Pedestron repository performed well on the CrowdHuman data set, representing a similar challenge to our data [21]. All testing and use of the CNN was performed using GPUs on the Frontera cluster. An example image after undergoing orthorectification and pedestrian detection is shown in Fig 1.

Using one GPU node on Frontera, with four NVIDIA Quadro RTX 5000 GPUs, the entire process takes about 3 seconds per original 360˚ image. Given the 4 million images we collected, this takes about 3,300 hours of computing time. While this is not a small number, when running in parallel, the whole process can be completed in a manner of days. In comparison, a human taking 10s per orthorectified image to count all the pedestrians would take over 22,000

Fig 1. Sample images from the pedestrian detection data pipeline. The left image is an original 360˚ image from a data collection run. The image on the right is the right-hand side of the original image after orthorectification and pedestrian detection (both sides of the image are processed separately). There are two pedestrians that were detected by the algorithm (in red bounding boxes).

https://doi.org/10.1371/journal.pone.0303180.g001

hours to complete the same task. File compression/decompression for file transfer also takes a substantial amount of time. Since we used DesignSafe as our main data storage platform, we had to transfer files to/from the Frontera supercomputer to perform our pedestrian detection. To avoid overloading the file transfer system, we compressed the images from each run into a tar file prior to transferring the files to Frontera. This file compression/decompression can take several hours per run, but can be performed in parallel with the detection algorithm since they are on different systems. After compression, file transfer using Globus [35] takes minutes.

In post-processing, the pipeline filters out low-confidence detections (defined as any detection with less than 80% confidence) and associates the remaining high-confidence detections to U.S. Cenus Bureau GEOIDs [19]. We arrived at this confidence level after tuning for the precision and recall of the CNN classifier. Specifically, the pipeline filters based on the output of the second to last layer of the CNN, known as a *softmax layer*. For a k–class classification problem, the softmax layer will output a k–dimensional probability vector, where each ith entry of the vector gives the probability that the original input to the CNN belongs to class i.

The final stage of post-processing is GEOID matching, where latitude and longitude metadata are cross-referenced to disjoint geographic regions (e.g. U.S. census tracts or block groups) and their respective GEOID codes. The cross-referencing code assumes the availability of shapefiles describing the geometry of the geographic regions. Aggregating the pedestrian detections according to U.S. Census Bureau GEOIDs [19] is necessary for analyses using socio-demographic data collected by the census. Additionally, the pedestrian detections can easily be cross-referenced with custom geometry defined using popular geographic information system software, such as the capitals data used in route construction and our analysis.

Following the GEOID matching step, the pedestrian detections data is written to a tabular format file (e.g. comma separated values). This file is an "analysis-ready" data product, in the sense that it is readable by most popular statistical analysis software (R, SPSS, Stata, etc.) and can be easily merged with other datasets using the GEOID column(s). A visual depiction of the entire pipeline is seen in Fig 2. Full code and a manual for following our process is available at https://github.com/marte292/rapid-data-pipeline.

Case study: Community mobility in Seattle during the COVID-19 pandemic

Data processing. All analysis is performed using the Python programming language version 3.11 [37]. The initial data product as outlined in the previous section is a list of detections,
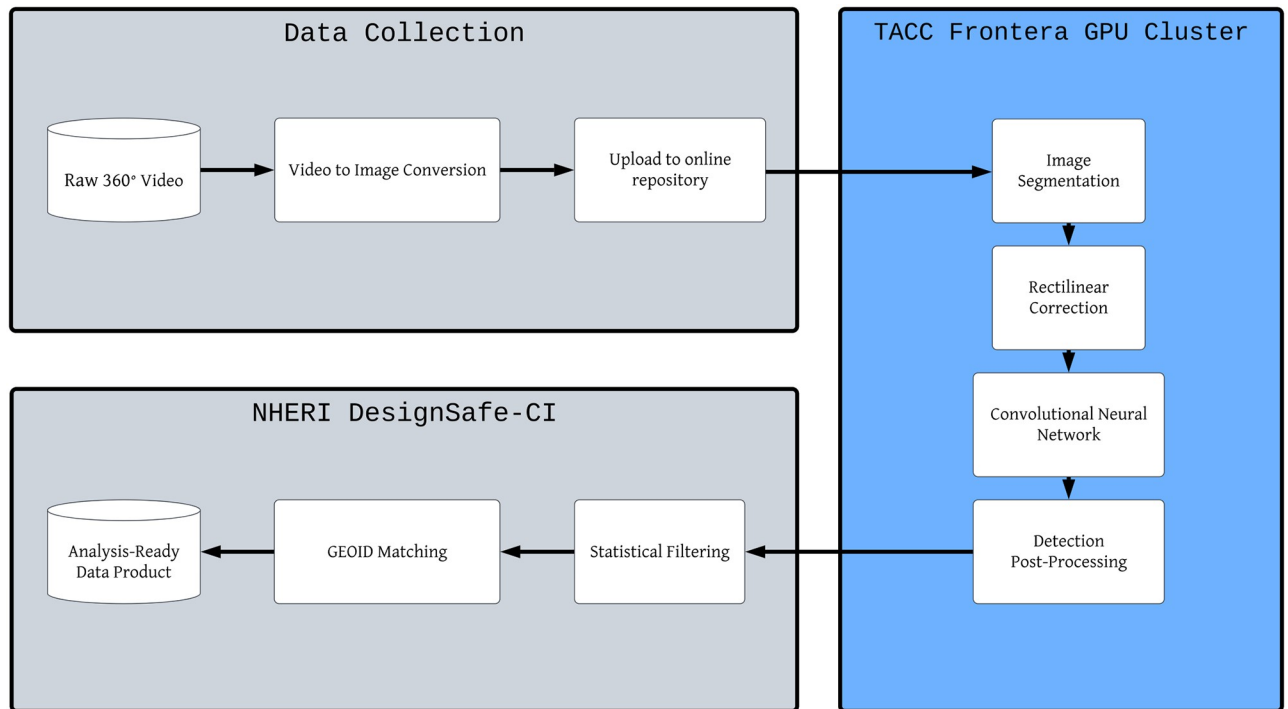
Fig 2. Flowchart of the data processing pipeline. The parts of the flowchart in gray occur on NHERI DesignSafe-CI, while the right-hand part in blue is done on the Frontera cluster.

https://doi.org/10.1371/journal.pone.0303180.g002

alongside the date of collection, geolocation, and GEOID. We also utilized a similar list of the images themselves with the same features. The last dataset we utilized is the median household income data and racial demographic data from the 2019 American Community Survey (ACS) 5-year estimates. We aggregated the detections and image data for each data collection survey at the census tract level, then matched each census tract's total number of detections and images to its respective demographic and income data.

We utilized the data from 36 of the 37 surveys, omitting data from 10-29-2020. A heavy rain event caused the survey to be stopped early due to poor video quality. For each survey, we divided the number of detections in each census tract by the number of images collected in the tract to create a normalized 'detections per image' metric. This is a necessary step as the number of images in each tract may change survey to survey due to circumstances outside our control, such as construction or community events altering the route.

The last step in data processing was to transform some of our data to be represented by categorical variables. The date of each survey was coded both as either a weekend or weekday, and by the season. The date was also coded as either being before, or after the date that vaccines became publicly available. Income data was coded to be one of 5 levels that were used during route design. These brackets were $48,274 and below, $48,275 to $80,819, $80,820, to $110,536, $110537 to $153,500, and $153,501 and above. Lastly, the proportion of the census tract's population that identifies as non-white was coded as an indicator variable, with '1' corresponding to areas that are 55.5% white or more. We determined this threshold using Jenk's natural breaks optimization. This left us with a dataset of 3171 observations to be used for analysis. Each observation represented a census tract with a detections per image value, as well as values for each of the categorical variables defined above.

Initial regression analysis. Using the processed data, we conducted a regression analysis to understand the relationships between predictors of interest and pedestrian traffic. Based on the known literature, we hypothesized that season, day of the week, COVID-19 vaccine availability, income level, and demographics all would have an impact on pedestrian traffic. We implemented a linear regression model to identify which of these factors are identified as statistically significant (α = .05). We chose this modeling approach for its simple interpretability, as our modeling goal is to describe. The regression model is detailed below:

$$Y = \beta_0 + \beta_1 \times I_{vaccine} + \beta_{2..4} \times C_{season} + \beta_5 \times I_{weekend}$$
$$+ \beta_{6..9} \times C_{incomelevel} + \beta_{10} \times I_{demographicindicator} + \epsilon, \tag{1}$$

where Y is the detections per image for a date/census tract combination; $I_{vaccine}$ is an indicator for if the vaccine was available on that date; C_{season} is a categorical variable with 3 levels for summer, winter, and spring; $I_{weekend}$ is an indicator for if it is the weekend or not; $C_{incomelevel}$ is a categorical variable with 4 levels for the 4 income brackets above the lowest bracket; $I_{demographicindicator}$ is an indicator variable for if the population is 55.5% white or more. β_0 is the baseline detections per image on a weekday, not in the summer, with the vaccine unavailable, in a census tract at the lowest income level and a population less than 55.5% white. β_1 represents the change in detections per image from the vaccine becoming available, and $\beta_{2..4}$ represent the change for different seasons. β_5 represents the change from a weekday to the weekend, and $\beta_{6..9}$ represent the change to other income brackets. Lastly, β_{10} represents the change in detections per image to from an area that is less than 55.5% white to an area that is more.

In addition to the above analysis, we subset the data by only looking at detections that occurred in an image with at least one other detection. Then we calculated detections per image again, and fit the above model again with the new response variable. This same process was followed for detections with at least two, three, and four other detections in the same image. The goal of these analyses was to see if there were different trends for larger groups of people when compared with the entire data set.

Results

Data pipeline

Our main contribution, the open-source data pipeline, is publicly available on https://github.com/marte292/rapid-data-pipeline. The repository contains a process manual with step-by-step instructions on how to implement the data pipeline in Python [37]. The required Python libraries and system requirements are provided. Additionally, we provide enough code for future researchers to implement the pipeline on their own systems, with their own file structure. The pipeline is capable of processing terabytes of image data and outputting an analysis-ready data product in a matter of days (using high-performance computing, such as a single GPU node on Frontera, an academic supercomputer) with minimal human input.

Case study

Using data from the Seattle street-level imagery campaign, we calculated the number of detections per image across all data collection surveys. Fig 3 shows the detections per image for each survey, as well as the detections per image for the subset of detections sharing an image with at least 4 others. Fig 3 also displays the timestamp of COVID-19 vaccines becoming publicly available in Washington state.

Fig 3 depicts the trends over time for detections per image and detections sharing an image with at least 4 others. While both graphs exhibit similar trends overall, notably after vaccine
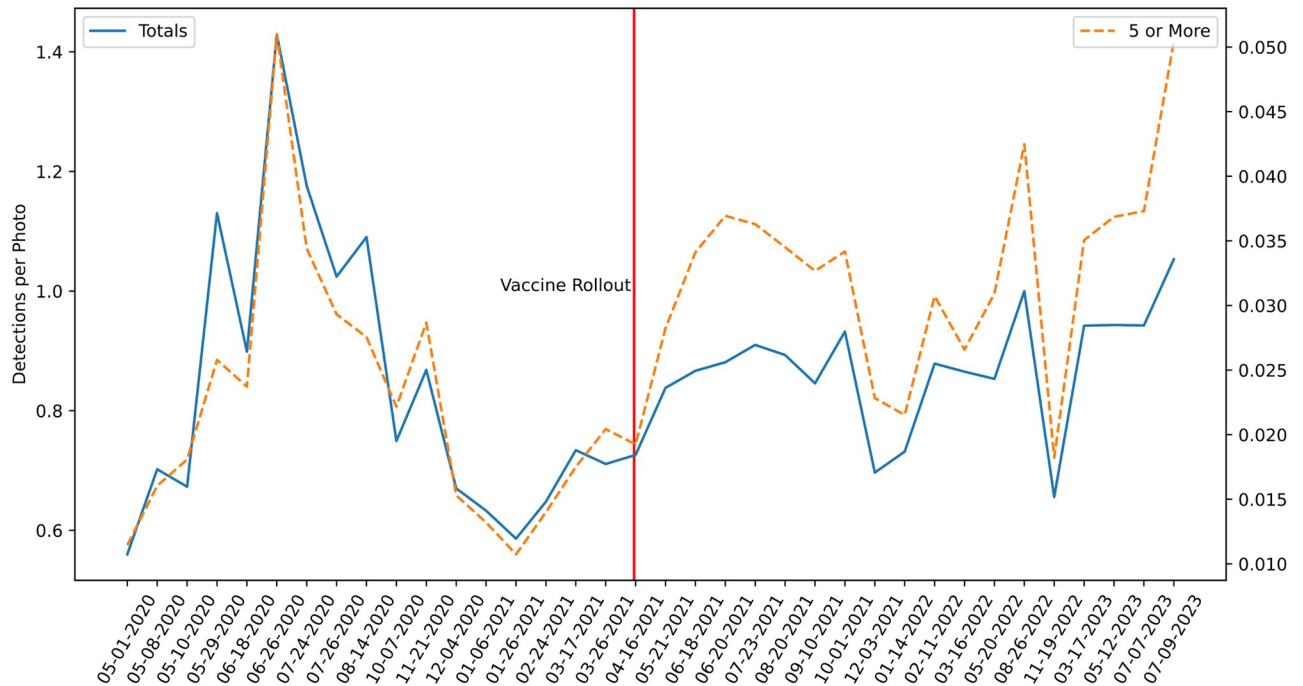
Fig 3. Time series data of the total detections per image (solid blue line, left axis), and detections per image for the subset of detections sharing an image with at least 4 others (orange dashed line, right axis). As the survey dates are irregular, all dates are included in the figure. Please note that the axis for total detections per image does not start at 0. This was done purposefully to facilitate comparison between the trends of the two graphs.

rollout the graph of detections sharing an image with at least 4 others exceeds the graph of detections per image in all cases. The spike in detections seen in June 2020 is due to the large scale protests of police brutality that took place in Seattle in the aftermath of George Floyd's murder.

The full results of the linear regression model for total detections per image are displayed in Table 1. They show that the season being summer is the only significant seasonal effect. Additionally, the income bracket is a significant predictor, with wealthier areas seeing less pedestrian traffic. Finally, a census tract having a population greater than 55.5% white is a significant positive predictor. All other variables are not significant, including vaccine availability.

For the regression models using a subset of data, the results are similar to the initial model. All models have the same significant predictors as the initial model. The model using the detections sharing an image with at least one other also had the weekend as a borderline significant, negative predictor. The models using detections sharing an image with at least 3 and 4 others had vaccine availability as a significant, positive predictor. The full results of the linear regression model for detections per image with at least 4 others are displayed in Table 2, with all other regression models available in the supporting information.

Discussion

Comparison to Google Community Mobility data

Given the ability to measure community mobility through pedestrian counts, there is potential value of our pipeline for social sciences and public health research [22, 23]. At an individual

Table 1. OLS regression results for detections per image. The first three non-intercept terms represent indicator variables for the different seasons, with fall being the baseline. The Vaccine Available term represents a binary variable for whether the COVID-19 initial vaccination series was publicly available or not. Weekend is a binary variable for whether the data was collected on Saturday or Sunday. The four Income Bracket terms are indicator variables for the median income level of the census tract where the data was collected. The income brackets are defined in our methods. Lastly, the More than 55.5% White term is an indicator variable for if the census tract in question had a populace that is more than 55.5% White. Full documentation for the Python package used to make this output is available from the developers [38].

Dep. Variable:	Detections_per_Image			R-squared:		0.086
Model:	OLS			Adj. R-squared:		0.083
Method:	Least Squares			F-statistic:		29.71
No. Observations:	3171			Prob (F-statistic):		3.20e-55
Df Residuals:	3160			Log-Likelihood:		-4456.8
Df Model:	10					
Covariance Type:	nonrobust					
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1.1242	0.075	14.914	0.000	0.976	1.272
Spring	0.0783	0.053	1.471	0.141	-0.026	0.183
Summer	0.2527	0.055	4.599	0.000	0.145	0.360
Winter	-0.0046	0.058	-0.079	0.937	-0.118	0.109
Vaccine Available	0.0061	0.036	0.172	0.863	-0.064	0.076
Weekend	-0.0775	0.052	-1.483	0.138	-0.180	0.025
Income Bracket 2	-0.4689	0.081	-5.795	0.000	-0.628	-0.310
Income Bracket 3	-0.8688	0.079	-11.041	0.000	-1.023	-0.714
Income Bracket 4	-0.9938	0.086	-11.540	0.000	-1.163	-0.825
Income Bracket 5	-1.3752	0.116	-11.893	0.000	-1.602	-1.148
More than 55.5% White	0.6416	0.054	11.820	0.000	0.535	0.748

Table 2. OLS regression results for detections per image for the detections subset sharing an image with at least 4 others. Coefficients are defined the same as in Table 1.

Dep. Variable:	Five_Or_More_Peds_per_Image			R-squared		0.059
Model:	OLS			Adj. R-squared:		0.056
Method:	Least Squares			F-statistic:		19.78
No. Observations:	3171			Prob (F-statistic):		7.09e-36
Df Residuals:	3160			Log-Likelihood:		4286.3
Df Model:	10					
Covariance Type:	nonrobust					
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.0393	0.005	8.213	0.000	0.030	0.049
Spring	0.0006	0.003	0.180	0.857	-0.006	0.007
Summer	0.0100	0.003	2.870	0.004	0.003	0.017
Winter	-0.0039	0.004	-1.057	0.291	-0.011	0.003
Vaccine Available	0.0093	0.002	4.109	0.000	0.005	0.014
Weekend	0.0009	0.003	0.270	0.787	-0.006	0.007
Income Bracket 2	-0.0289	0.005	-5.626	0.000	-0.039	-0.019
Income Bracket 3	-0.0468	0.005	-9.373	0.000	-0.057	-0.037
Income Bracket 4	-0.0541	0.005	-9.889	0.000	-0.065	-0.043
Income Bracket 5	-0.0688	0.007	-9.376	0.000	-0.083	-0.054
More than 55.5% White	0.0309	0.003	8.960	0.000	0.024	0.038

level, higher physical activity is known to predict better physical [39, 40] and mental health [41–43], and is associated with higher self-reported satisfaction and quality of life [44, 45]. In an aggregate sense, mobility is theorized to be an intermediate variable through which socio-economic deprivation affects vulnerability to infectious disease [46, 47], resilience to disasters [48], and exposure to environmental hazards [49]. In light of this body of literature, we argue that the use of pedestrian counts to assess mobility could be a differentiating factor in researching social and health inequity. One extremely common source of mobility data during the COVID-19 Pandemic has been Google Community Mobility Reports [24] and Apple Mobility Trends Reports [25]. While there have been improvements in recent years [50], there are known representation and self-selection biases with existing mobility data captured by smartphones and other internet-based data collection methods [51–55].

Given the large number of publications using smartphone data as the foundation for their work, a natural question is how our data compares to smartphone mobility data. Comparison between our data set and the still publicly available Google Community Mobility Reports data can reveal some of the similarities and differences between the two data sets [24]. Google Community Mobility data is reported at the county level in the United States. Since Seattle is in King County, Washington, the King County data is what we use to draw the comparison.

Google Community Mobility data does not provide raw mobility numbers, but rather is reported as a percentage change from the five-week period of Jan 5–Feb 6, 2020. This data is collected from smartphones running the Android operating system with location history turned on, which is off by default. The data is baselined by day of the week, so data from a given Monday is compared to the median of the five Mondays in the baseline window to calculate a percent change. Additionally, it is unclear how exactly Google quantifies mobility. It is mentioned that it combines number of visitors to a location with amount of time spent in that location, but no specifics beyond that are provided.

Google mobility data is broken down into different categories. The category that most closely aligns with one of the categories used in our analysis is parks. Although Google's data classifies parks as official national parks and not the general outdoors, it does not indicate how it accounts for city or state parks. Our own data for park locations is based on the City of Seattle's official classifications.

Fig 4 shows a comparison of our detections per image data against Google Community Mobility data. Note that not all surveys are included because Google Community Mobility data stopped being provided on October 15, 2022. Overall, the trends between the two data sets are remarkably similar, lending further credibility to our data collection procedure. The more notable differences in the graph are from the months of November 2020 through August 2021, where the Google mobility data shows a larger drop followed by an increase in community mobility than was visible through our own data.

One plausible explanation for this is the upwards sampling bias that occurs when using smartphone data [56, 57]. Our data set captures anyone on the street, including individuals experiencing homelessness, who are less likely to have smartphones. This population was on the streets throughout the entirety of the pandemic, so they were consistently captured by our data collection efforts. This consistent baseline pedestrian count could lead to a lesser response to vaccine rollout and winter weather in our own data in comparison with Google's. Additionally, there is a known income gap in both vaccination rates and smartphone ownership [58, 59]. This gap could drive the increase in the Google Mobility data during vaccine rollout.
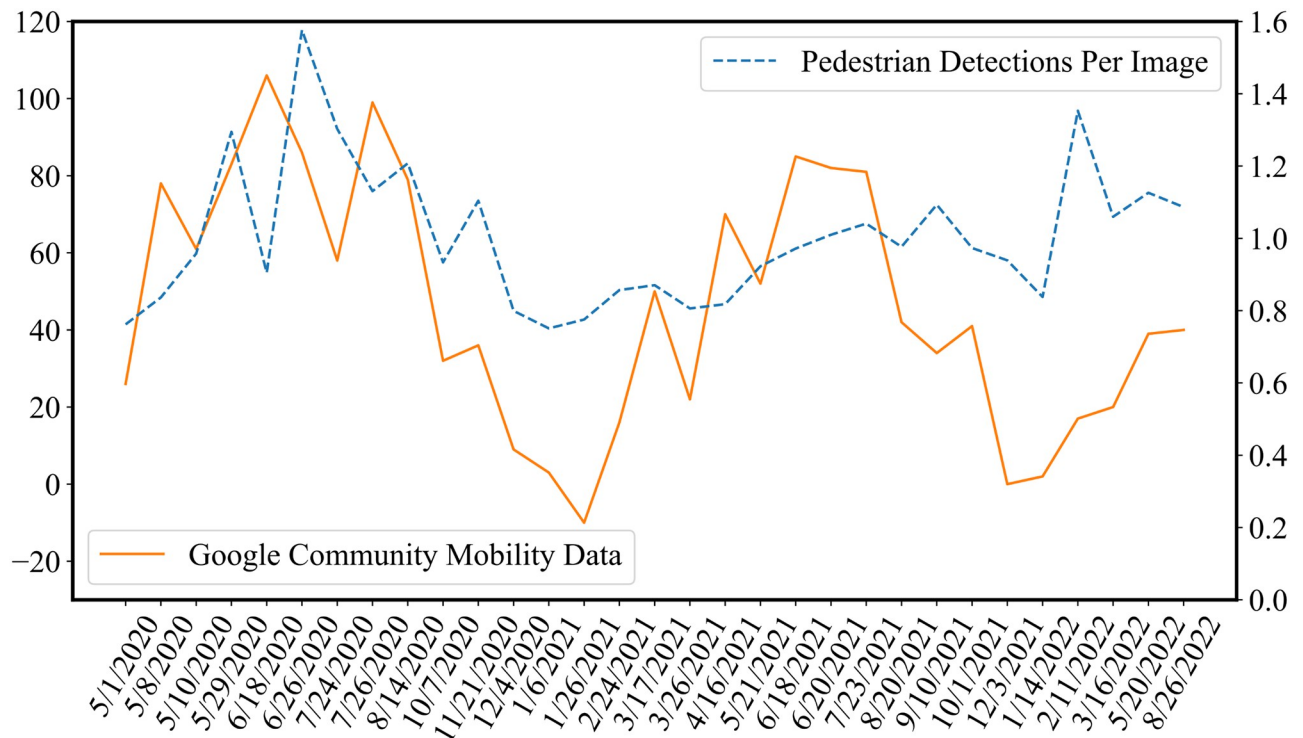
Fig 4. A depiction of our own detections per image data (blue, dashed; right axis) against Google Community Mobility data (orange, solid; left axis). The Pearson correlation between the two data sets is 0.387. The Google Community mobility data is aggregated at King County, WA, while our data covers a survey route within Seattle, which belongs to King County. As the dates of surveys were irregular (e.g., due to weather conditions), all dates are included in the figure.

https://doi.org/10.1371/journal.pone.0303180.g004

Implications, limitations, and extensions

Our results show that it is possible for researchers to collect and analyze longitudinal SVI data. The presented methods can be used to collect and process SVI data from 8 hours worth of video in a manner of days. This time will only further decrease with faster data processing infrastructure and methods. These methods will allow novel longitudinal SVI data to be collected for research in a variety of application areas.

The results of the case study also bear further discussion. We demonstrated expected relationships between seasonal effects like day of week and weather on pedestrian traffic. Additionally, we showed that pedestrian traffic is inversely proportional to income, a known result during the COVID-19 pandemic, as lower income households are constrained in their capacity to work from home or take time off of work [30, 60]. Our results also showed that more white areas had higher on average pedestrian counts. This could be due to known trends, such as areas with larger non-white populations being more likely to stay home in response to government restrictions [61] and participate in other risk-reducing practices such as wearing a mask [62], or just due to local trends, as racial mobility trends tend to vary between cities [63]. These findings are consistent across all of our models, both looking at the entire data set, and the subsets examining pedestrians sharing an image. These results validate our method with respect to established literature, and provide a quantitative confirmation of results that had previously been found using cell phone data.

One new finding from our case study is that while overall pedestrian counts did not respond to vaccine availability, the subset of pedestrians who were in larger groups (4+ people

in an image) did. Likely, the reason we did not see a response to the vaccine in the aggregate data is because our data only captures people who are outdoors. There is data that shows that outdoor pedestrian activity varied across cities, frequently increasing at recreation locations like trails, during the early days of the pandemic [64, 65]. Given these increases at some locations, a return to 'normal' pedestrian traffic may not mean an increase, but rather a change in traffic patterns. Our data captures this by showing that there was a significant increase in larger groups of people after the vaccine became available. This implies that people were more willing to be near each other outdoors after they had been vaccinated.

While the data pipeline presented here does represent a method for generating a novel data product, there are implementation challenges worth further discussion. For data collection, in addition to the time required to drive the route limiting the places of interest the route could reach, there were also many tradeoffs that had to be made when designing the route itself [31]. Despite having our survey route carefully designed to assess a representative sample of the Seattle population, some bias in route design is unavoidable. Since the route design included data from the American Community Survey aggregated at the census tract level, there is an implicit assumption of spatial homogeneity of the population within each census tract. Such bias is a manifestation of the well-known modifiable areal unit problem [66]. Since the majority of the route was primarily based on locations of interest throughout the city, this concern is somewhat mitigated.

In terms of processing, the pre-trained model we used required a substantial amount of high-performance computing time, and at times the data product generated was so large as to be unwieldy. Given the challenge our data set represents, using a model designed to be generalizable is necessary to attain good detection results. As many state-of-the-art models perform substantially worse out of sample, we had to be careful to choose a model that was designed to perform well in this situation, at the cost of slower computing times [67]. Another unforeseen challenge was regular updates to the video camera's software to process and segment the video data into images. Consistent image formatting was vital for the data processing pipeline to function, so regular quality checks are necessary to make sure the images are processed properly.

The data product created, pedestrian detections, has some limitations as well. First, our method only captures pedestrians who are outdoors and near enough to the street to be captured via camera. This means that our data set does not include people who are indoors at these locations of interest, or who are too far from the street to be seen by camera. While the changes over time in pedestrian traffic we observed are still meaningful, it is important to recognize they don't capture everything. Similarly, our data cannot be interpreted as the actual number of pedestrians on the street. There is overlap in the image data, even when subset at 4 meter intervals and cropped during orthorectification. The orthorectified images only represent about 25% of the originals. However, this natural cropping is not enough to avoid the image overlap and further cropping would risk information loss. Pedestrians that appear in the foreground of one image may end up in the background of another. There are also several known instances of cyclists keeping relative pace with the street-view vehicle for several blocks, resulting in numerous detections. These issues are easy to circumvent in analysis by comparing the relative number of detections, although at the cost of interpretability.

Even with the above limitations, the data pipeline presented in this paper can be directly applied or adapted to be used in a number of contexts. Potential applications of longitudinal SVI data in assessing the built environment [14], broad urban research [1, 3, 68], and health research [8] have been well-documented, as the temporal instability of existing SVI data is discussed as a limitation in all of these fields. Beyond this, it is possible to estimate population demographics [4], and other neighborhood-level statistics [13, 69] using SVI data. As our ability to quickly and accurately parse scenes using computer vision improves [70], potential application areas will only increase in number.

Another field where longitudinal SVI data could contribute a lot is disaster research. There is a substantial body of research dedicated to empirical methods for modeling various aspects of disaster recovery [71]. Our methods could be applied in this field to quantify recovery using pedestrian detections as a metric for community mobility, or another metric assessing the built environment as appropriate. Similar work has been done using repeat photography after Hurricane Katrina [15] but our methods represent a substantial increase in generated data, allowing for a wider range of analyses. Spatial video data collection for disaster reconnaissance has also been done [72], but involves manual assessment of the captured video. Our methods demonstrate that a fully-automated approach is possible, which would allow for more frequent data collection at a lower cost.

Conclusion

This article describes the creation of the first open-source SVI data pipeline for longitudinal analysis. Regression analysis based on the resulting longitudinal SVI data showed that pedestrian traffic patterns changed in response to the availability of the COVID-19 vaccine, thereby demonstrating the data pipeline's usefulness in research and practice. In particular, we showed that there were statistically significant increases in groups of people in proximity to each other after the vaccine became publicly available. Our data also captured expected trends in pedestrian traffic based on annual seasonality and socioeconomic factors. Our results demonstrate the feasibility and value in collecting SVI data as part of a longitudinal study. Longitudinal SVI data is capable of providing valuable insights in a variety of fields of study. Future work includes applications of our methods in broader public health research, disaster research, and other fields of study that can benefit from longitudinal SVI data. Potential methodological directions include study-specific route design process improvements and newer pedestrian detection approaches, as further progress is made in this area.

Supporting information

S1 Dataset. Full dataset used for obtaining regression results presented in this paper.
(CSV)

S1 File.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Nicole A. Errett, Scott B. Miles, Joseph Wartman, Youngjun Choe.

Data curation: Matthew Martell, Nick Terry, Ribhu Sengupta, Chris Salazar.

Formal analysis: Matthew Martell, Nick Terry, Ribhu Sengupta.

Funding acquisition: Nicole A. Errett, Scott B. Miles, Joseph Wartman, Youngjun Choe.

Investigation: Matthew Martell, Nick Terry.

Methodology: Matthew Martell, Nick Terry, Nicole A. Errett, Scott B. Miles, Joseph Wartman, Youngjun Choe.

Project administration: Nicole A. Errett, Scott B. Miles, Joseph Wartman, Youngjun Choe.

Resources: Joseph Wartman, Youngjun Choe.

Software: Matthew Martell, Nick Terry, Ribhu Sengupta, Chris Salazar.

Supervision: Youngjun Choe.

Validation: Matthew Martell, Nick Terry, Youngjun Choe.

Visualization: Matthew Martell, Nick Terry.

Writing – original draft: Matthew Martell, Nick Terry.

Writing – review & editing: Matthew Martell, Nick Terry, Nicole A. Errett, Scott B. Miles, Joseph Wartman, Youngjun Choe.

References

1. Cinnamon J, Jahiu L. Panoramic Street-Level Imagery in Data-Driven Urban Research: A Comprehensive Global Review of Applications, Techniques, and Practical Considerations. ISPRS International Journal of Geo-Information. 2021; 10(7):471. https://doi.org/10.3390/ijgi10070471

2. Gilge C. Google Street View and the Image as Experience. GeoHumanities. 2016; 2(2):469–484. https://doi.org/10.1080/2373566X.2016.1217741

3. Li Y, Peng L, Wu C, Zhang J. Street View Imagery (SVI) in the Built Environment: A Theoretical and Systematic Review. Buildings. 2022; 12(8):1167. https://doi.org/10.3390/buildings12081167

4. Gebru T, Krause J, Wang Y, Chen D, Deng J, Aiden EL, et al. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. Proceedings of the National Academy of Sciences—PNAS. 2017; 114(50):13108–13113. https://doi.org/10.1073/pnas.1700035114 PMID: 29183967

5. Badland HM, Opit S, Witten K, Kearns RA, Mavoa S. Can Virtual Streetscape Audits Reliably Replace Physical Streetscape Audits? Journal of Urban Health. 2010; 87(6):1007–1016. https://doi.org/10.1007/s11524-010-9505-x PMID: 21104331

6. Ringland J, Bohm M, Baek SR, Eichhorn M. Automated survey of selected common plant species in Thai homegardens using Google Street View imagery and a deep neural network. Earth Science Informatics. 2021; 14(1):179–191. https://doi.org/10.1007/s12145-020-00557-3

7. Yin L, Cheng Q, Wang Z, Shao Z. 'Big data' for pedestrian volume: Exploring the use of Google Street View images for pedestrian counts. Applied Geography. 2015; 63:337–345. https://doi.org/10.1016/j.apgeog.2015.07.010

8. Rzotkiewicz A, Pearson AL, Dougherty BV, Shortridge A, Wilson N. Systematic review of the use of Google Street View in health research: Major themes, strengths, weaknesses and possibilities for future research. Health & Place. 2018; 52:240–246. https://doi.org/10.1016/j.healthplace.2018.07.001 PMID: 30015181

9. Wang R, Yuan Y, Liu Y, Zhang J, Liu P, Lu Y, et al. Using street view data and machine learning to assess how perception of neighborhood safety influences urban residents' mental health. Health & Place. 2019; 59:102186. https://doi.org/10.1016/j.healthplace.2019.102186 PMID: 31400645

10. Novack T, Vorbeck L, Lorei H, Zipf A. Towards Detecting Building Facades with Graffiti Artwork Based on Street View Images. ISPRS International Journal of Geo-Information. 2020; 9(2). https://doi.org/10.3390/ijgi9020098

11. Mapillary. Mapillary; 2013. [Cited 2020 July 10]. Available from: https://www.mapillary.com.

12. Mahabir R, Schuchard R, Crooks A, Croitoru A, Stefanidis A. Crowdsourcing Street View Imagery: A Comparison of Mapillary and OpenStreetCam. ISPRS International Journal of Geo-Information. 2020; 9 (6). https://doi.org/10.3390/ijgi9060341

13. Suel E, Bhatt S, Brauer M, Flaxman S, Ezzati M. Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. Remote Sensing of Environment. 2021; 257:112339. https://doi.org/10.1016/j.rse.2021.112339 PMID: 33941991

14. Smith CM, Kaufman JD, Mooney SJ. 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; 72:102701. https://doi.org/10.1016/j.healthplace.2021.102701

15. Burton C, Mitchell JT, Cutter SL. Evaluating post-Katrina recovery in Mississippi using repeat photography. Disasters. 2011; 35(3):488–509. https://doi.org/10.1111/j.1467-7717.2010.01227.x PMID: 21272057

16. Depauw L, Blondeel H, De Lombaerde E, De Pauw K, Landuyt D, Lorer E, et al. The use of photos to investigate ecological change. Journal of Ecology. 2022; 110(6):1220–1236. https://doi.org/10.1111/1365-2745.13876

17. Hao X, Lyu H, Wang Z, Fu S, Zhang C. Estimating the spatial-temporal distribution of urban street ponding levels from surveillance videos based on computer vision. Water Resources Management. 2022; 36 (6):1799–1812. https://doi.org/10.1007/s11269-022-03107-2

18. Velastin SA, Fernández R, Espinosa JE, Bay A. Detecting, Tracking and Counting People Getting On/ Off a Metropolitan Train Using a Standard Video Camera. Sensors (Basel, Switzerland). 2020; 20 (21):6251. https://doi.org/10.3390/s20216251 PMID: 33147784

19. Bureau USC. Understanding Geographic Identifiers (GEOIDs); 2021. [Cited 2021 October 20]. Available from: https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html.

20. Hasan I, Liao S, Li J, Akram SU, Shao L. Pedestrian Detection: Domain Generalization, CNNs, Transformers and Beyond. arXiv preprint arXiv:220103176. 2022;.

21. Shao S, Zhao Z, Li B, Xiao T, Yu G, Zhang X, et al. CrowdHuman: A Benchmark for Detecting Human in a Crowd. arXiv preprint arXiv:180500123. 2018;.

22. Hou X, Gao S, Li Q, Kang Y, Chen N, Chen K, et al. Intracounty modeling of COVID-19 infection with human mobility: Assessing spatial heterogeneity with business traffic, age, and race. Proceedings of the National Academy of Sciences. 2021; 118(24):e2020524118. https://doi.org/10.1073/pnas.2020524118 PMID: 34049993

23. Schlosser F, Maier BF, Jack O, Hinrichs D, Zachariae A, Brockmann D. COVID-19 lockdown induces disease-mitigating structural changes in mobility networks. Proceedings of the National Academy of Sciences. 2020; 117(52):32883–32890. https://doi.org/10.1073/pnas.2012326117 PMID: 33273120

24. Google COVID-19 Community Mobility Reports; 2020. [Cited 2020 April 15] Available from: https://www.google.com/covid19/mobility/.

25. Apple COVID-19 Mobility Trends Reports; 2020. [Cited 2020 April 15] Available from: https://covid19.apple.com/mobility.

26. Paez A. Using Google Community Mobility Reports to investigate the incidence of COVID-19 in the United States. Findings. 2020. https://doi.org/10.32866/001c.12976

27. Loewenthal G, Abadi S, Avram O, Halabi K, Ecker N, Nagar N, et al. COVID-19 pandemic-related lockdown: response time is more important than its strictness. EMBO Molecular Medicine. 2020; 12(11): e13171. https://doi.org/10.15252/emmm.202013171 PMID: 33073919

28. Jacobsen GD, Jacobsen KH. Statewide COVID-19 Stay-at-Home Orders and Population Mobility in the United States. World Medical & Health Policy. 2020; 12(4):347–356. https://doi.org/10.1002/wmh3.350 PMID: 32837774

29. Finazzi F. Replacing discontinued Big Tech mobility reports: a penetration-based analysis. Scientific Reports. 2023; 13(1):935–935. https://doi.org/10.1038/s41598-023-28137-7 PMID: 36650298

30. Weill JA, Stigler M, Deschenes O, Springborn MR. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. Proceedings of the National Academy of Sciences. 2020; 117(33):19658–19660. https://doi.org/10.1073/pnas.2009412117 PMID: 32727905

31. Errett NA, Wartman J, Miles SB, Silver B, Martell M, Choe Y. Street View Data Collection Design for Disaster Reconnaissance. arXiv preprint arXiv:230806284. 2023;.

32. Miles SB. Foundations of community disaster resilience: Well-being, identity, services, and capitals. Environmental Hazards. 2015; 14(2):103–121. https://doi.org/10.1080/17477891.2014.999018

33. Rathje EM, Dawson C, Padgett JE, Pinelli JP, Stanzione D, Adair A, et al. DesignSafe: New Cyberinfrastructure for Natural Hazards Engineering. Natural hazards review. 2017; 18(3). https://doi.org/10.1061/(ASCE)NH.1527-6996.0000246

34. Stanzione D, West J, Evans RT, Minyard T, Ghattas O, Panda DK. In: Frontera: The Evolution of Leadership Computing at the National Science Foundation. New York, NY, USA: Association for Computing Machinery; 2020. p. 106–111.

35. Chard K, Foster I, Tuecke S. Globus: Research Data Management as Service and Platform. PEARC17. New York, NY, USA: Association for Computing Machinery; 2017.

36. Yang W, Qian Y, Kamarainen JK, Cricri F, Fan L. Object Detection in Equirectangular Panorama. 24th International Conference on Pattern Recognition. IEEE; 2018. p. 2190–2195

37. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.

38. Seabold S, Perktold J. statsmodels: Econometric and statistical modeling with python. 9th Python in Science Conference. 2010.

39. Baker PS, Bodner EV, Allman RM. Measuring life-space mobility in community-dwelling older adults. Journal of the American Geriatrics Society. 2003; 51(11):1610–1614. https://doi.org/10.1046/j.1532-5415.2003.51512.x PMID: 14687391

40. Petersen J, Austin D, Mattek N, Kaye J. Time out-of-home and cognitive, physical, and emotional well-being of older adults: a longitudinal mixed effects model. PloS one. 2015; 10(10):e0139643. https://doi.org/10.1371/journal.pone.0139643 PMID: 26437228

41. Zhu J, Fan Y. Daily travel behavior and emotional well-being: Effects of trip mode, duration, purpose, and companionship. Transportation Research Part A: Policy and Practice. 2018; 118:360–373.

42. Polku H, Mikkola TM, Portegijs E, Rantakokko M, Kokko K, Kauppinen M, et al. Life-space mobility and dimensions of depressive symptoms among community-dwelling older adults. Aging & Mental Health. 2015; 19(9):781–789. https://doi.org/10.1080/13607863.2014.977768 PMID: 25376479

43. Vallée J, Cadot E, Roustit C, Parizot I, Chauvin P. The role of daily mobility in mental health inequalities: the interactive influence of activity space and neighbourhood of residence on depression. Social science & medicine. 2011; 73(8):1133–1144. https://doi.org/10.1016/j.socscimed.2011.08.009 PMID: 21903318

44. Mulry C, Rivera G, Musini K, Rankin A, Astorini M. The relationship between community mobility, health status, and quality of life (QOL) in older adults. The American Journal of Occupational Therapy. 2019; 73(4_Supplement_1):7311505078p1–7311505078p1. https://doi.org/10.5014/ajot.2019.73S1-PO1017

45. Bergstad CJ, Gamble A, Gärling T, Hagman O, Polk M, Ettema D, et al. Subjective well-being related to satisfaction with daily travel. Transportation. 2011; 38(1):1–15. https://doi.org/10.1007/s11116-010-9283-z

46. Ossimetha A, Ossimetha A, Kosar CM, Rahman M. Socioeconomic disparities in community mobility reduction and COVID-19 growth. In: Mayo Clinic Proceedings. vol. 96. Elsevier; 2021. p. 78–85.

47. Zhai W, Liu M, Peng ZR. Social distancing and inequality in the United States amid COVID-19 outbreak. Environment and Planning A: Economy and Space. 2021; 53(1):3–5. https://doi.org/10.1177/0308518X20932576

48. Hong B, Bonczak BJ, Gupta A, Kontokosta CE. Measuring inequality in community resilience to natural disasters using large-scale mobility data. Nature communications. 2021; 12(1):1–9. https://doi.org/10.1038/s41467-021-22160-w PMID: 33767142

49. Lewis J. Exposures in the City: Looking for Socioeconomic Patterns for the Urban Exposome. Environmental Health Perspectives. 2019; 127(4):044003. https://doi.org/10.1289/EHP4807 PMID: 31013143

50. Stockham N, Washingon P, Chrisman B, Paskov K, Jung JY, Wall DP. Causal Modeling to Mitigate Selection Bias and Unmeasured Confounding in Internet-Based Epidemiology of COVID-19: Model Development and Validation. JMIR public health and surveillance. 2022; 8(7):e31306–e31306. https://doi.org/10.2196/31306 PMID: 35605128

51. Klingwort Jonas, Schnell Rainer. Critical Limitations of Digital Epidemiology. Survey research methods. 2020; 14(2). https://doi.org/10.18148/srm/2020.v14i2.7726

52. Liu Z, Maneekul P, Pendergrast C, Doubleday A, Miles SB, Errett NA, et al. Physical activity monitoring data following disasters. Sustainable Cities and Society. 2022; 81:103814. https://doi.org/10.1016/j.scs.2022.103814

53. Roy A, Nelson TA, Fotheringham AS, Winters M. Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists. Urban Science. 2019; 3(2). https://doi.org/10.3390/urbansci3020062

54. Milusheva S, Bjorkegren D, Viotti L. Assessing Bias in Smartphone Mobility Estimates in Low Income Countries. In: ACM SIGCAS Conference on Computing and Sustainable Societies; 2021. p. 364–378.

55. Aleta A, Martín-Corral D, Bakker MA, y Piontti AP, Ajelli M, Litvinova M, et al. Quantifying the importance and location of SARS-CoV-2 transmission events in large metropolitan areas. Proceedings of the National Academy of Sciences. 2022; 119(26):e2112182119. https://doi.org/10.1073/pnas.2112182119 PMID: 35696558

56. Birenboim A, Shoval N. Mobility Research in the Age of the Smartphone. Annals of the American Association of Geographers. 2016; 106(2):283–291.

57. Thimm T, Seepold R. Past, present and future of tourist tracking. Journal of Tourism Futures. 2016; 2 (1):43–55. https://doi.org/10.1108/JTF-10-2015-0045

58. Barry V, Dasgupta S, Weller DL, Kriss JL, Cadwell BL, Rose C, et al.. Patterns in COVID-19 Vaccination Coverage, by Social Vulnerability and Urbanicity—United States, December 14, 2020–May 1, 2021; 2021.

59. Center PR. Mobile Fact Sheet; 2021. [Cited 2022 January 8]. Available from: https://www.pewresearch.org/internet/fact-sheet/mobile/.

60. Elarde J, Kim JS, Kavak H, Züfle A, Anderson T. Change of human mobility during COVID-19: A United States case study. PLOS ONE. 2021; 16(11):e0259031. https://doi.org/10.1371/journal.pone.0259031 PMID: 34727103

61. Singh S, Shaikh M, Hauck K, Miraldo M. Impacts of introducing and lifting nonpharmaceutical interventions on COVID-19 daily growth rate and compliance in the United States. Proceedings of the National Academy of Sciences. 2021; 118(12):e2021359118. https://doi.org/10.1073/pnas.2021359118 PMID: 33658331

62. Hearne BN. and Niño MD. Understanding How Race, Ethnicity, and Gender Shape Mask-Wearing Adherence During the COVID-19 Pandemic: Evidence from the COVID Impact Survey Journal of Racial and Ethnic Health Disparities. 2022; 9(1):176–183. https://doi.org/10.1007/s40615-020-00941-1 PMID: 33469866

63. Chang S, Pierson E, Koh PW, Gerardin J, Redbird B, Grusky D, et al. Mobility network models of COVID-19 explain inequities and inform reopening. Nature. 2021; 589(7840):82–87. https://doi.org/10.1038/s41586-020-2923-3 PMID: 33171481

64. Kraus S, Koch N. Provisional COVID-19 infrastructure induces large, rapid increases in cycling. Proceedings of the National Academy of Sciences. 2021; 118(15):e2024399118. https://doi.org/10.1073/pnas.2024399118 PMID: 33782111

65. Doubleday A, Choe Y, Busch Isaksen T, Miles S, Errett NA. How did outdoor biking and walking change during COVID-19?: A case study of three U.S. cities. PloS one. 2021; 16(1):e0245514. https://doi.org/10.1371/journal.pone.0245514 PMID: 33471858

66. Fotheringham AS, Wong DW. The modifiable areal unit problem in multivariate statistical analysis. Environment and planning A. 1991; 23(7):1025–1044. https://doi.org/10.1068/a231025

67. Hasan I, Liao S, Li J, Akram SU, Shao L. Generalizable pedestrian detection: The elephant in the room. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 11328–11337.

68. Biljecki F, Ito K. Street view imagery in urban analytics and GIS: A review. Landscape and Urban Planning. 2021; 215:104217. https://doi.org/10.1016/j.landurbplan.2021.104217

69. Gullon P, Fry D, Plascak JJ, Mooney SJ, Lovasi GS. Measuring changes in neighborhood disorder using Google Street View longitudinal imagery: a feasibility study. Cities & Health. 2023; 7(5):823–829. https://doi.org/10.1080/23748834.2023.2207931 PMID: 37850028

70. Dong S, Zhou W, Xu C, Yan W. EGFNet: Edge-Aware Guidance Fusion Network for RGB–Thermal Urban Scene Parsing. IEEE Transactions on Intelligent Transportation Systems. 2023; p. 1–13. https://doi.org/10.1109/TITS.2023.3306368

71. Martell M, Miles SB, Choe Y. Review of Empirical Quantitative Data Use in Lifeline Infrastructure Restoration Modeling. Natural hazards review. 2021; 22(4). https://doi.org/10.1061/(ASCE)NH.1527-6996.0000514

72. Curtis A, Mills JW. Spatial video data collection in a post-disaster landscape: The Tuscaloosa Tornado of April 27th 2011. Applied Geography. 2012; 32(2):393–400. https://doi.org/10.1016/j.apgeog.2011.06.002