Street Sense: Learning from Google Street View*

Suriyan Laohaprapanon† Kimberly Ortleb‡ Gaurav Sood§

July 9, 2018

Abstract

How good are the public services and the public infrastructure? Does their quality vary by income? These are vital questions—they shed light on how well the government is doing its job, the consequences of disparities in local funding, etc. But there is little good data on many of these questions. We fill this gap by describing a scalable method of getting data on one crucial piece of public infrastructure: roads. We assess the quality of roads and sidewalks by exploiting data from Google Street View. We randomly sample locations on major roads, query Google Street View images for those locations and code the images using Amazon's Mechanical Turk. We apply this method to assess the quality of roads in Bangkok, Jakarta, Lagos, and Wayne County, Michigan. Jakarta's roads have nearly four times the potholes than roads of any other city. Surprisingly, the proportion of road segments with potholes in Bangkok, Lagos, and Wayne is about the same, between .06 and .07. Using the data, we also estimate the relation between the condition of the roads and local income in Wayne, MI. We find that roads in more affluent census tracts have somewhat fewer potholes.

*Data and scripts behind the analysis presented here can be downloaded from https://github.com/geosensing/streetsense.

†Suriyan can be reached at: suriyant@gmail.com

‡Kimberly can be reached at: kim.ortleb@gmail.com

§Gaurav can be reached at: gsood07@gmail.com

The poorer the quality of the public infrastructure and public services, generally, the worse the quality of life. For instance, potholed roads mean that vehicles can't go as fast and the ride is bumpier. If sidewalks aren't paved, physically disabled have a tough time getting anywhere. If garbage isn't picked up regularly, foul smells and unsightliness are part of life, and the risk of disease is greater.

As these examples convey, the quality of public infrastructure and public services matters immensely. It sheds light on the quality of life, and on the resources and functioning of the government. So how good is the public infrastructure? And how good are the public services? More often than not, we have no good answer to these questions.

In this paper, we introduce a method to answer questions about the quality of one important piece of public infrastructure: roads. We capitalize on Google Street View to learn about the condition of the roads. We randomly sample locations on the roads, get Google Street View images for those locations, and crowdsource the coding of the images. To illustrate the method's utility, we apply the method to learn about the condition of roads in Wayne (Michigan), Bangkok, Lagos, and Jakarta, and to assess the association between local income and the condition of the roads in Wayne. We also discuss ways this labeled data can be augmented and used to build automated systems to answer these questions at scale.

Learning From Google Street View

Since 2007, Google has been working on regularly taking panoramic images of all the streets in the world. In the West, Google's efforts have been a success: Google's specially designed vehicles have traversed an overwhelming majority of the streets.[1] In the third-world, however, the coverage is patchy. For instance, as we show below, just about 24.6% of Dhaka's streets

[1] See https://en.wikipedia.org/wiki/Coverage_of_Google_Street_View

are covered by Google Street View.[2] But Google's coverage of some other big third-world cities isn't too shabby. For instance, it covers 99.9% of Bangkok's streets and 87.2% of Lagos' streets. In all, the coverage is good enough, especially in the West, that people can build a scalable measurement infrastructure on top of it.

But patchy coverage is not the only problem with Google Street View data. The other is that the data are not always current. A large chunk of the data is at least a few years old. But somewhat older data has its value, especially because we expect Google to map those areas again in the future. The data aren't perfect but they are rich and valuable.

But how do we efficiently capitalize on Google Street View data? We could download all the data for a city. But doing so is expensive. And it may not even be useful. Depending on the question, a large random sample can fill in nicely for a census. For learning the condition of the roads, that is precisely the case.

To efficiently learn about the condition of the streets, sidewalks, and such, from Google Street View data, we devise a new workflow. We start by downloading data on the kinds of roads we are interested from Open Street Map (OSM). We then chunk the roads into half a kilometer segments, and then randomly sample from the segments. (The open source Python package geo-sampling (Laohaprapanon and Sood 2017) implements this workflow.) We then take the starting latitude and longitude of the sampled segments and query the Google Street View API.

Application

To illustrate the utility of the method, we apply it to learn the condition of the roads, the condition of the sidewalks, and the presence of litter on the streets in four prominent third-world cities and one poor American county.

[2] Some of Google's estimates of its coverage are either wrong or have become outdated as the road network continues to grow.

To learn the condition of roads in Bangkok, Dhaka, Jakarta, Lagos, and Wayne, MI, in the latter half of 2017, we downloaded data on all the streets from OSM. We feared that in many of these cities, Google Street View's coverage of neighborhood roads would be patchy. So we decided to focus on primary, secondary, tertiary, and trunk roads. We used the geo-sampling package to take a random sample of primary, secondary, tertiary, and trunk road segments for each location (see Figures 1, 2, 3). (Figures SI 2.1, SI 2.2, SI 2.3, SI 2.4 plot the starting longitude and latitude without the surrounding detail of the sampled segments of Bangkok, Jakarta, Lagos, and Wayne, MI respectively.) For Bangkok, Dhaka, Jakarta, and Lagos, we drew a sample of 1,000 segments each. For Wayne, MI, we drew a sample of 5,000 segments. We drew a larger sample for Wayne, MI because we wanted to estimate the relationship between local income and road conditions there. (We chose an American county to estimate the relationship between local income and road conditions because data on local income is readily available for the US.)

Figure 1: Sampled Locations in Bangkok

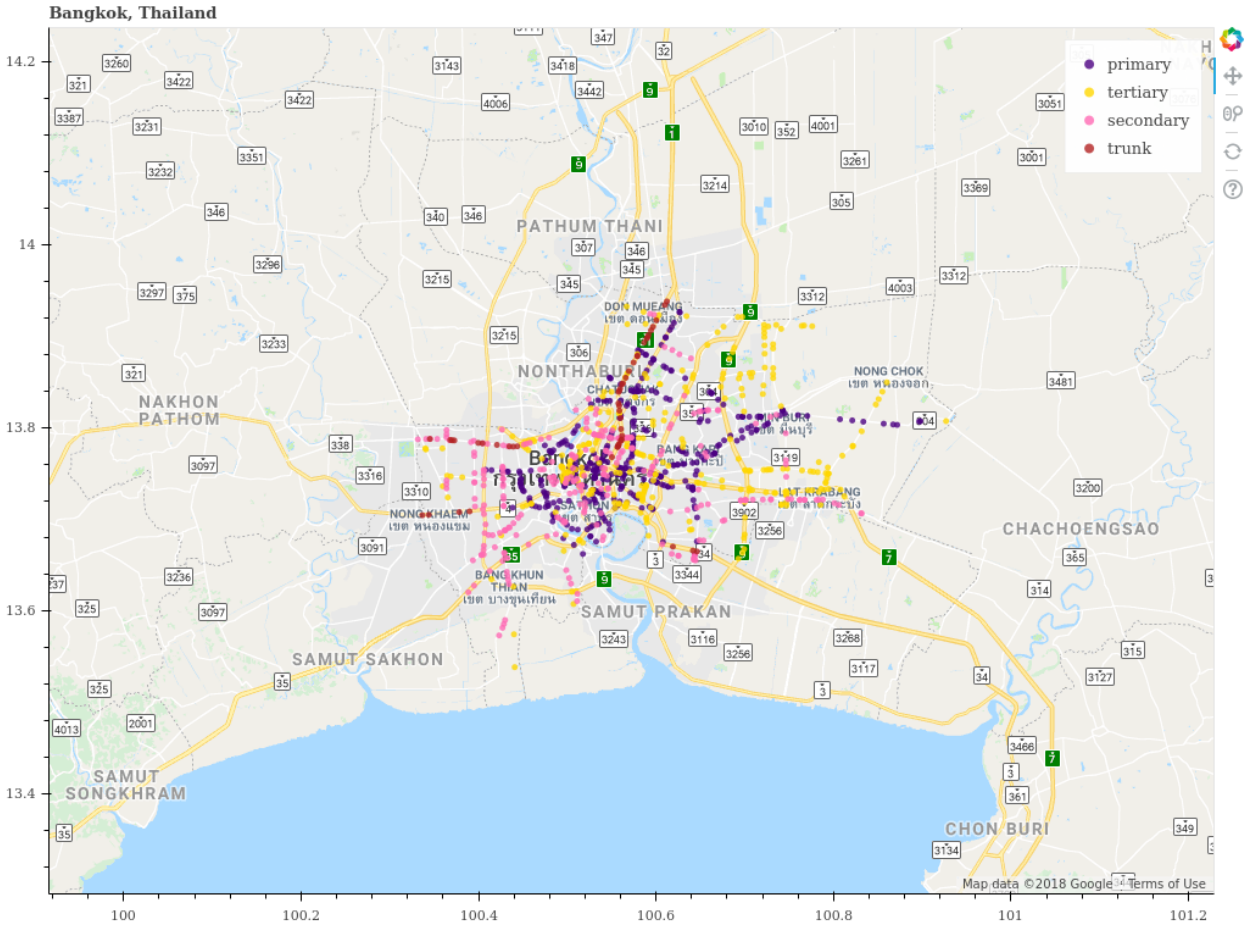


Figure 2: Sampled Locations in Jakarta

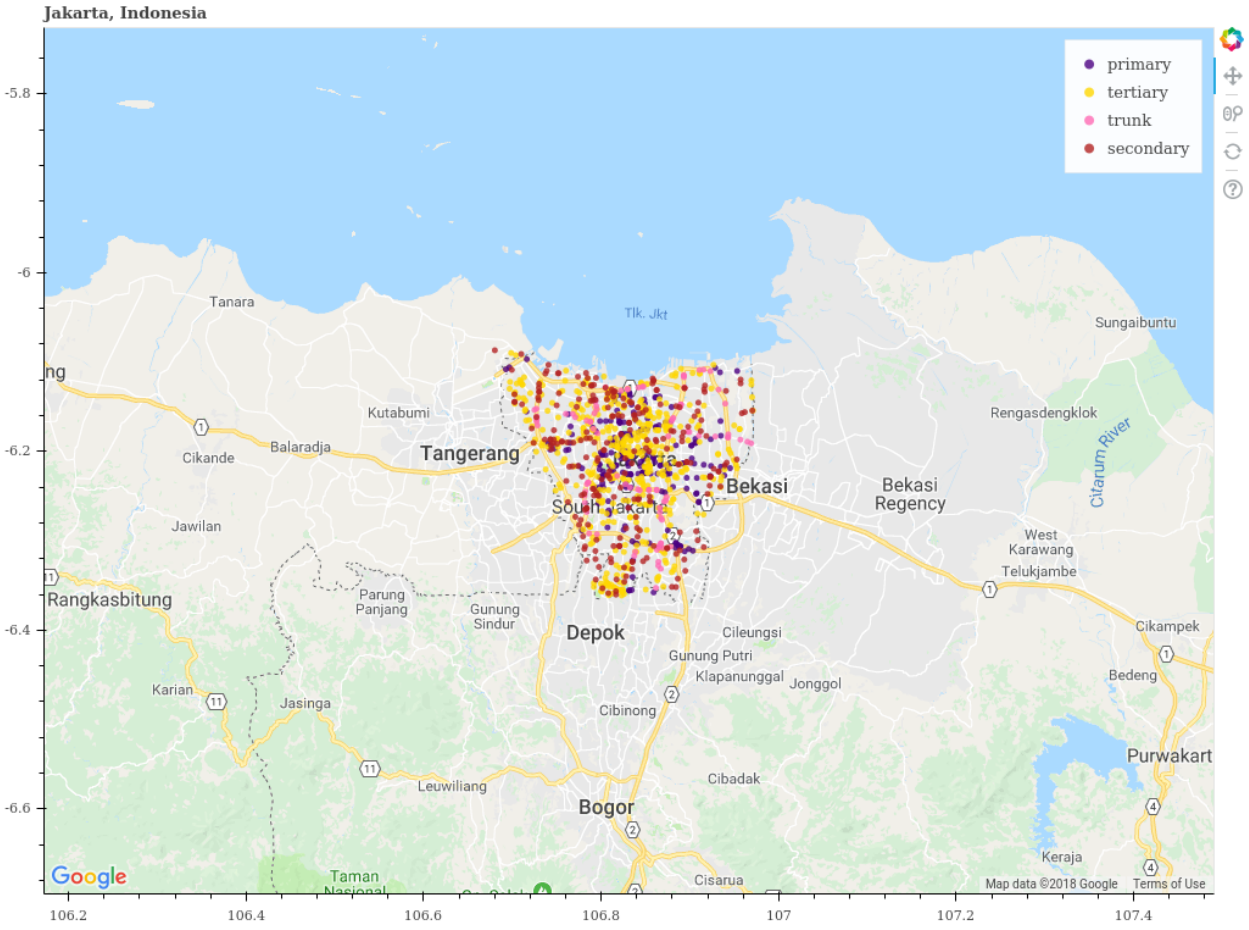


Figure 3: Sampled Locations in Lagos

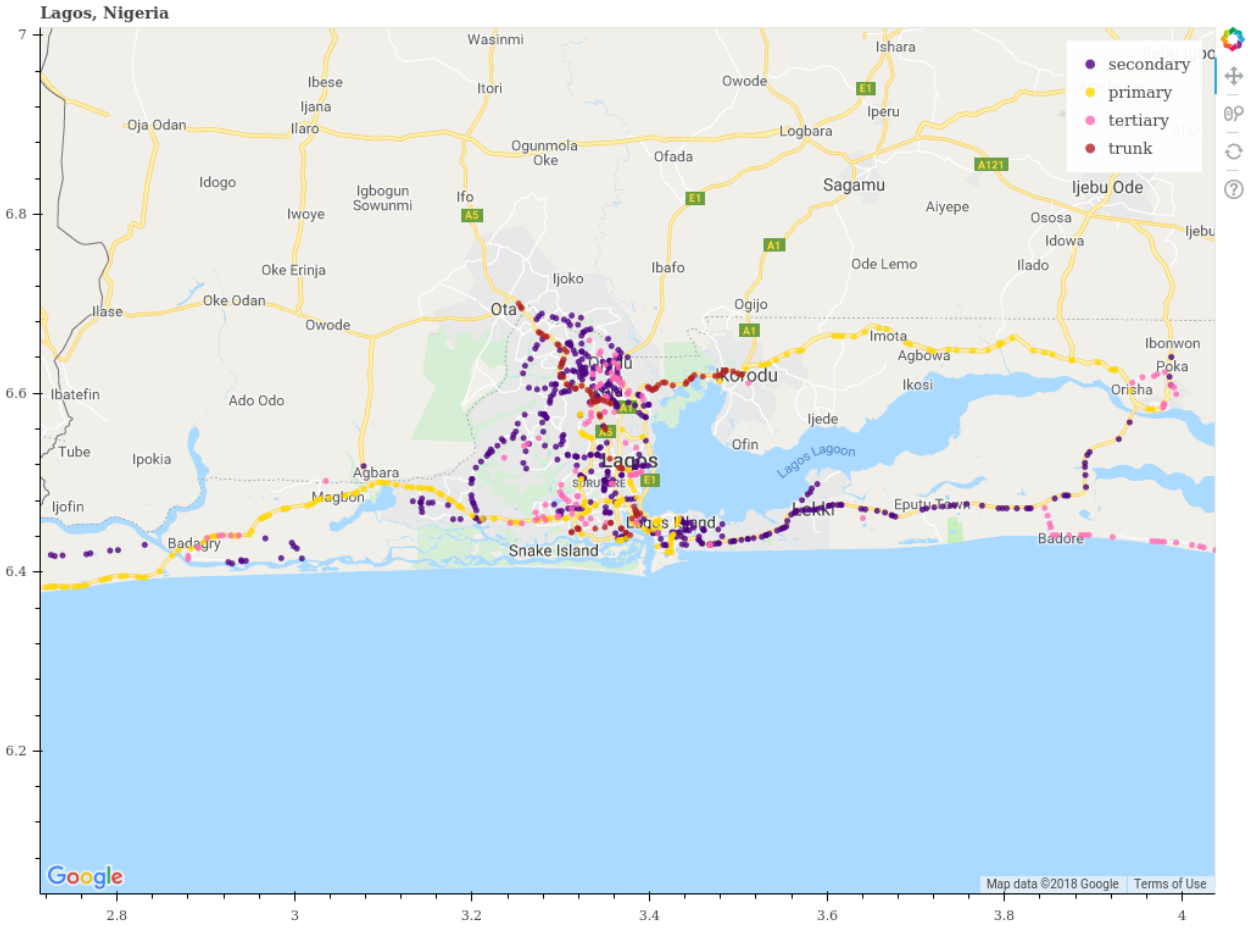
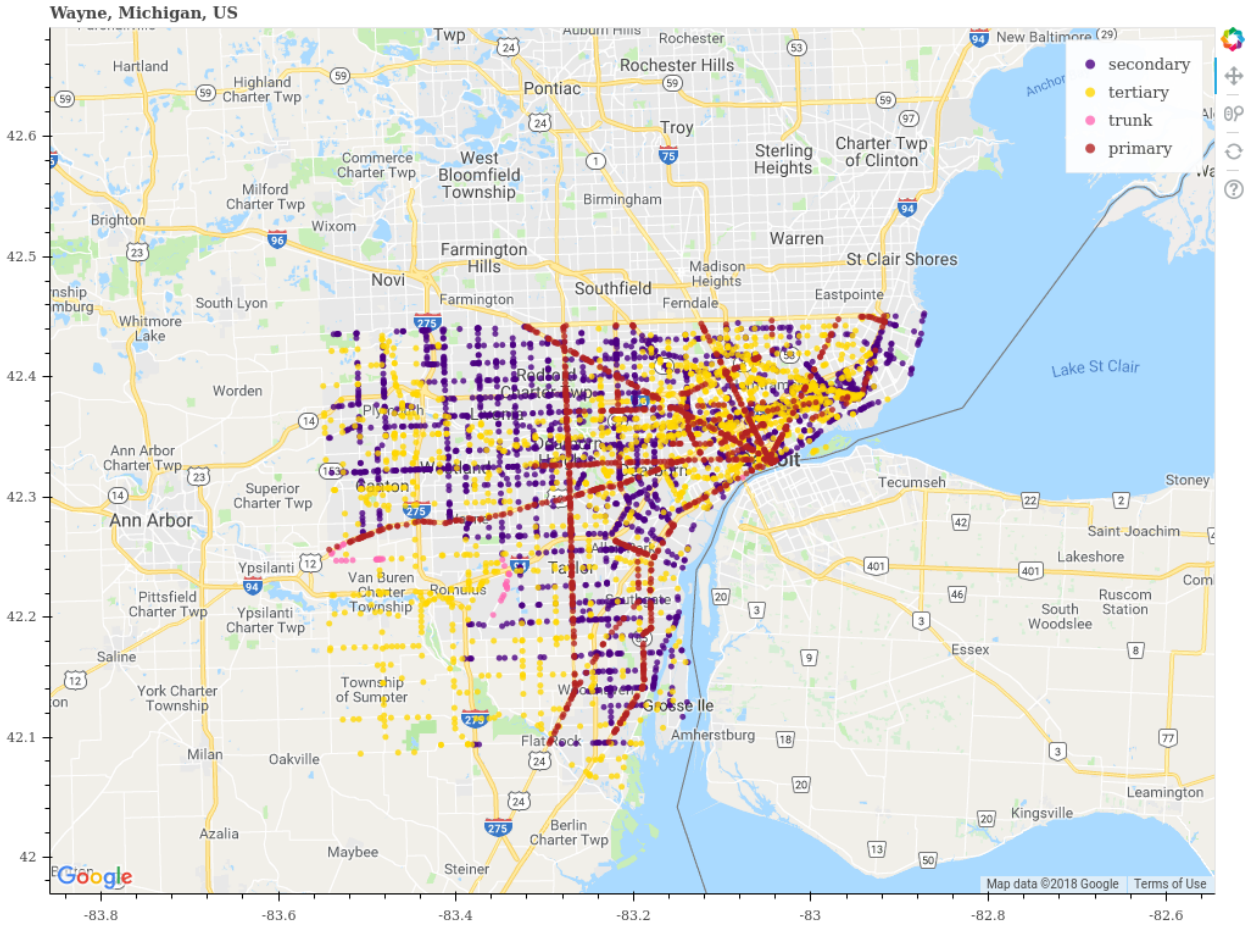


Figure 4: Sampled Locations in Wayne



Next, we used the Google Street View API to download images at the starting point of each of the random road segment. Sometimes the Google API came back empty. We take the proportion of failed queries as an estimate of Google Street View coverage of the primary, secondary, tertiary, and trunk roads in the respective city. In Dhaka, for instance, just about 24.6% of queries were successful. (Figure SI 1.1 plots the sampled locations.) Given the low coverage of Dhaka, we dropped Dhaka. In all, we have images of 978 locations for Bangkok, 872 for Jakarta, 999 for Lagos, and 4,828 for Wayne. Each photo captures a small segment of the road. (All the photos are available on Harvard Dataverse.)

Next, we recruited workers on Amazon's Mechanical Turk (MTurk) to code the images for the condition of the roads. To ensure quality, we only recruited 'master' workers. We asked

them if the segment of the road in the image had any 1) cracks, and 2) potholes. We also asked them, "if there are any road markings on the road, are they clear?" Lastly, we asked them, if there any litter and if the sidewalks were paved. The final survey for Bangkok, Jakarta, and Wayne, MI was the same (see SI 3.1).[3] Lagos' survey differed in very minor ways from Bangkok, Jakarta, and Wayne's (see SI 3.2). We paid MTurkers 5 cents for answering the short survey for each image. To ensure quality, we also checked a few images at random to see if the coding was reasonable. We found one instance where one worker's judgments seemed really off and decided to reject those HITs.

Results

Lest the readers miss an obvious point, before we present the results, we would like to draw their attention to it. Differences in the quality of roads across cities do not by default capture the extent of the road network. The extent of road network is easy to compute and regularly cited. Our contribution is measurement of quality of roads, sidewalks, and litter on the streets efficiently.

The proportion of road segments with potholes is Jakarta is an astonishing .23. The commensurate number for Bangkok, Lagos, and Wayne is between .06–.07. But what does that mean? As we mentioned above, each image captures a small segment of the street. If we assume that a photo captures .5km, the expected number of potholes on a 10 km journey in Jakarta would be 2.3. That would make for a somewhat of a rough ride.

When it comes to cracks in the road, Wayne takes the top spot—the proportion of segments in Wayne with cracks is .62 followed by .44 for Jakarta and .20 and .24 for Bangkok and Lagos respectively. The high proportion is not particularly noteworthy for Wayne given its latitude, but it is noteworthy for Jakarta.

[3] We initially got Jakarta's images coded using alternate instrumentation (see SI 3.3). But we were concerned that this would lead to incommensurability. So we did another round of data collection with the same instrument.

Jakarta is also the dirtiest of the 4 cities with .21 of the segments containing litter. Lagos comes second with .15 of the segments with litter. Lagos also takes the bottom spot for paved sidewalks—just .30 of the segments have a paved sidewalk.

Table 1: Condition of the Roads in Different Places.

city	potholes	cracks	clear road markings	roads w/ markings	litter	paved sidewalk
bangkok	.06	.24	.81	.98	.06	.51
jakarta	.23	.44	.34	.97	.21	.49
lagos	.06	.20	.23	.95	.15	.30
wayne	.07	.62	.60	.90	.09	.67

Given there are differences across cities in the proportion of trunk, primary, secondary, and tertiary roads in the road network, we checked if cross-city comparisons are mostly capturing differences in road types than differences in conditions within each type of road. To examine this, we regressed the appropriate variable (whether or not there is a pothole, a crack) on the type of the road and city. Compared to tertiary roads, potholes are more common on primary roads (Diff. $= .03$), secondary roads (Diff. $= .01$), and trunk roads (Diff. $= .05$). But adjusting for the type of road doesn't change the across-city estimates much. For instance, the difference in the proportion of segments with potholes between Wayne and Jakarta is still .16.

Moving to cracks in the road, compared to tertiary roads, primary, secondary, and trunk roads have fewer cracks with differences being -.05, -.03, and -.09 respectively. Like with potholes, adjusting for the kind of roads doesn't seem to make much of a difference for inferences from raw data for cross-city comparison.

Next, we analyzed the relationship between the condition of the roads and local income. To do that, we used the AskGeo API to get information on per capita income the census tract in which the lat/long lay. And we regressed whether a segment had a crack (or a pothole) or not on income split into quintiles.

Before we present the results, a caveat. Given that we expect the largest relationship between quality of neighborhood roads and local income, we expect that our subsetting on

primary, secondary, tertiary, and trunk roads to lead to smaller coefficients.

Compared to road segments in tracts with per capita income less than 12k, the proportion of road segments with potholes in tracts with per capita income between 12k and 17k was -.01 less. The proportion of road segments in tracts with per capita income of 17k to 23k was -.02 less. For tracts with 23k and 29k, it was -.03 fewer segments with potholes, and for tracts with income between 29k and 83k, -.02 fewer segments had potholes. The relationship between local income and the proportion of segments with cracks was more uneven. The highest quintile had the fewest cracks but roads in the second and third income quintile areas had roughly the same number of cracks.

Discussion

What is the condition of the streets? Are the streets paved? Do the streets have proper traffic signs and road markings? Is there litter on the streets? What proportion of vehicles on the streets is two-wheeled? And what proportion is man-powered, e.g., rickshaws? These are some of many the questions we can answer with Google Street View. In this paper, we provide a scalable way to answer such questions. We capitalize on Google Street View, pairing it with an open source Python package to randomly sample locations on the streets and crowdsourcing, to learn a host of compelling facts.

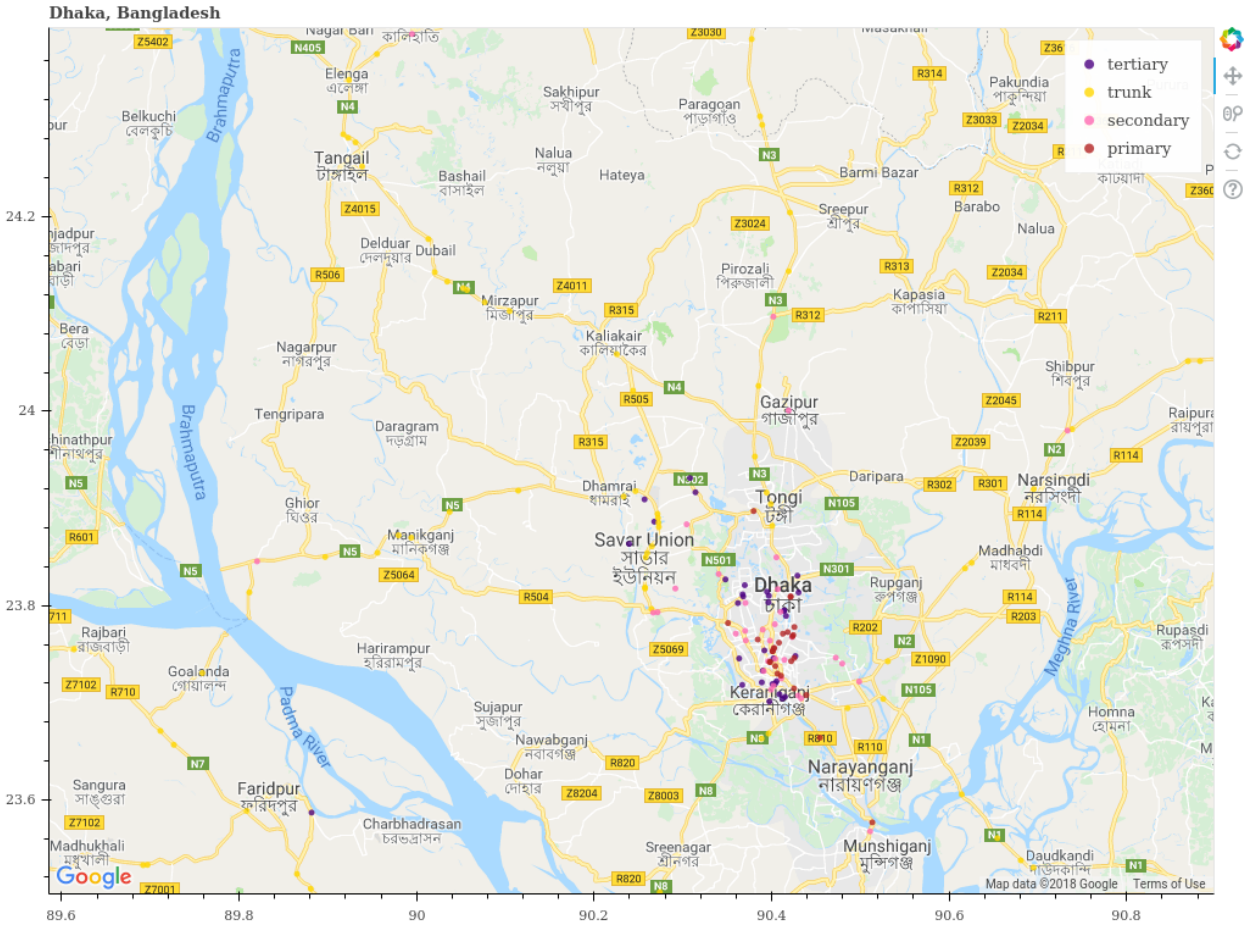
The method that we describe here can be easily extended to automate the production of answers. Given that we are technically building a large labeled dataset, an obvious next step is to build a supervised machine learning infrastructure on top of it. Such an infrastructure can then provide automated estimates on many of these questions, along with useful caveats around coverage.

References

Laohaprapanon, Suriyan and Gaurav Sood. 2017. "geosampling: Sampling Locations on the Streets.". https://github.com/geosensing/geo_sampling.

SI 1 Sampled Locations in Dhaka

Figure SI 1.1: Sampled Locations in Wayne



SI 2 Plots of Sampled Locations

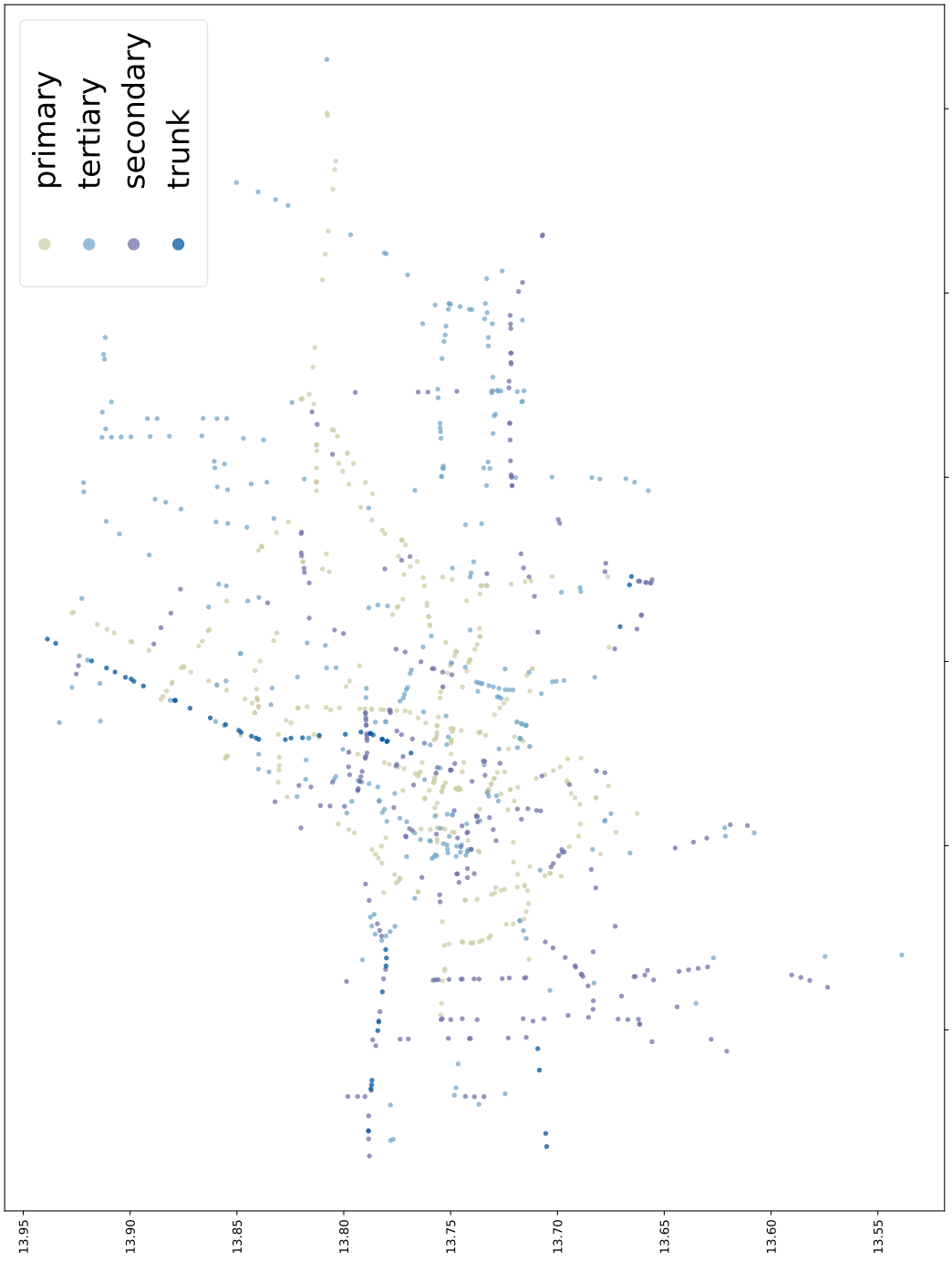


Figure SI 2.1: Sampled Locations in Bangkok

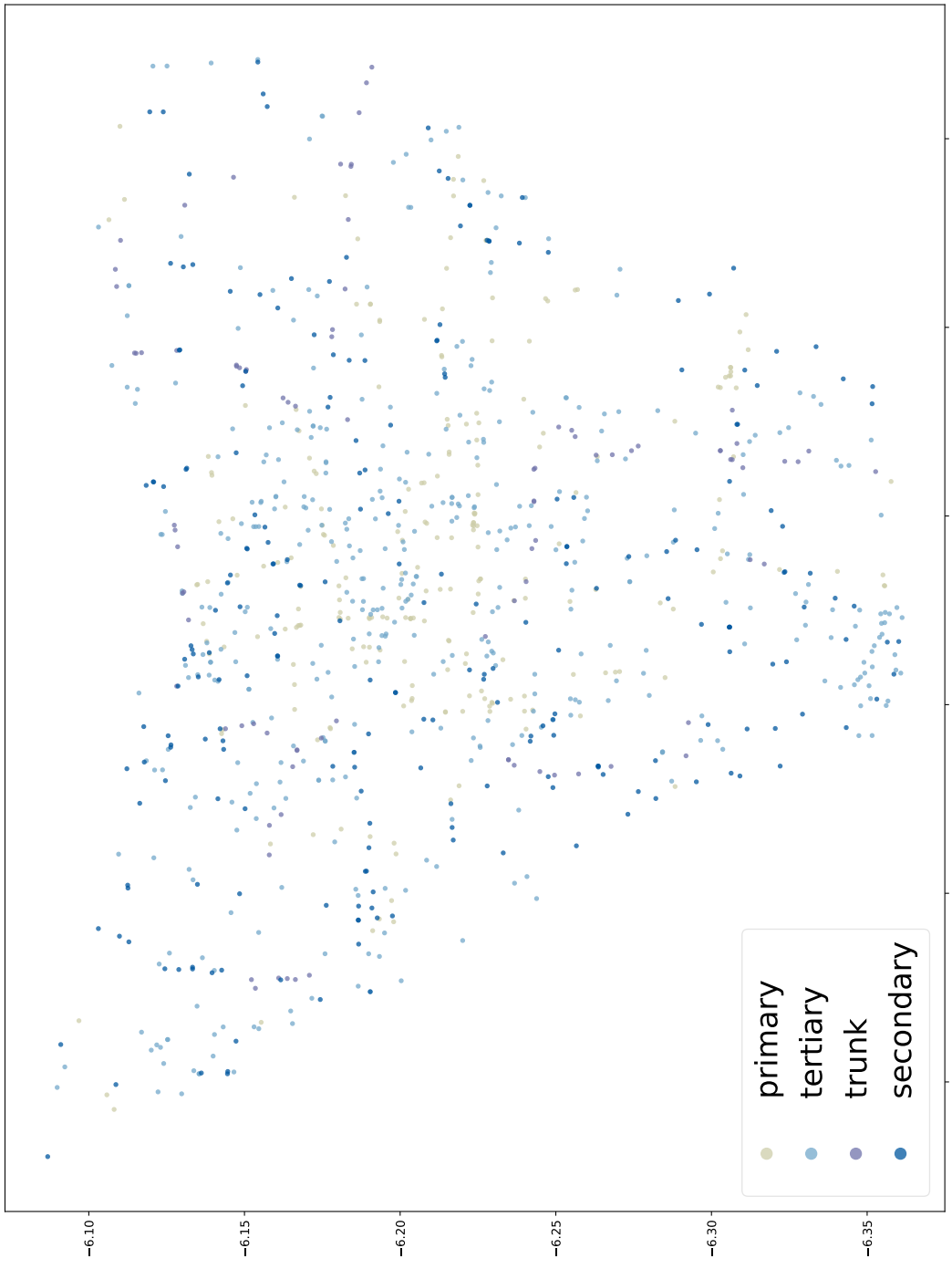


Figure SI 2.2: Sampled Locations in Jakarta

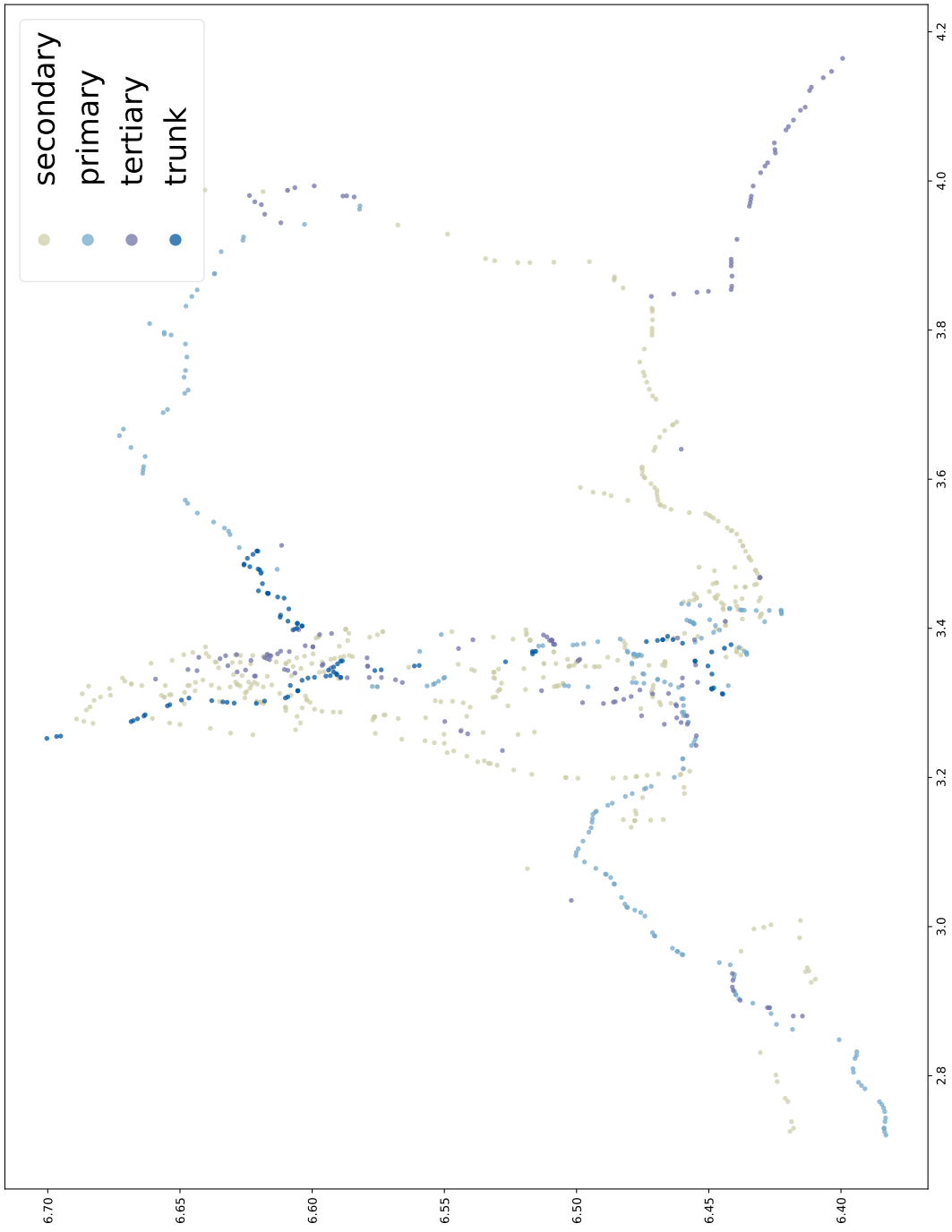
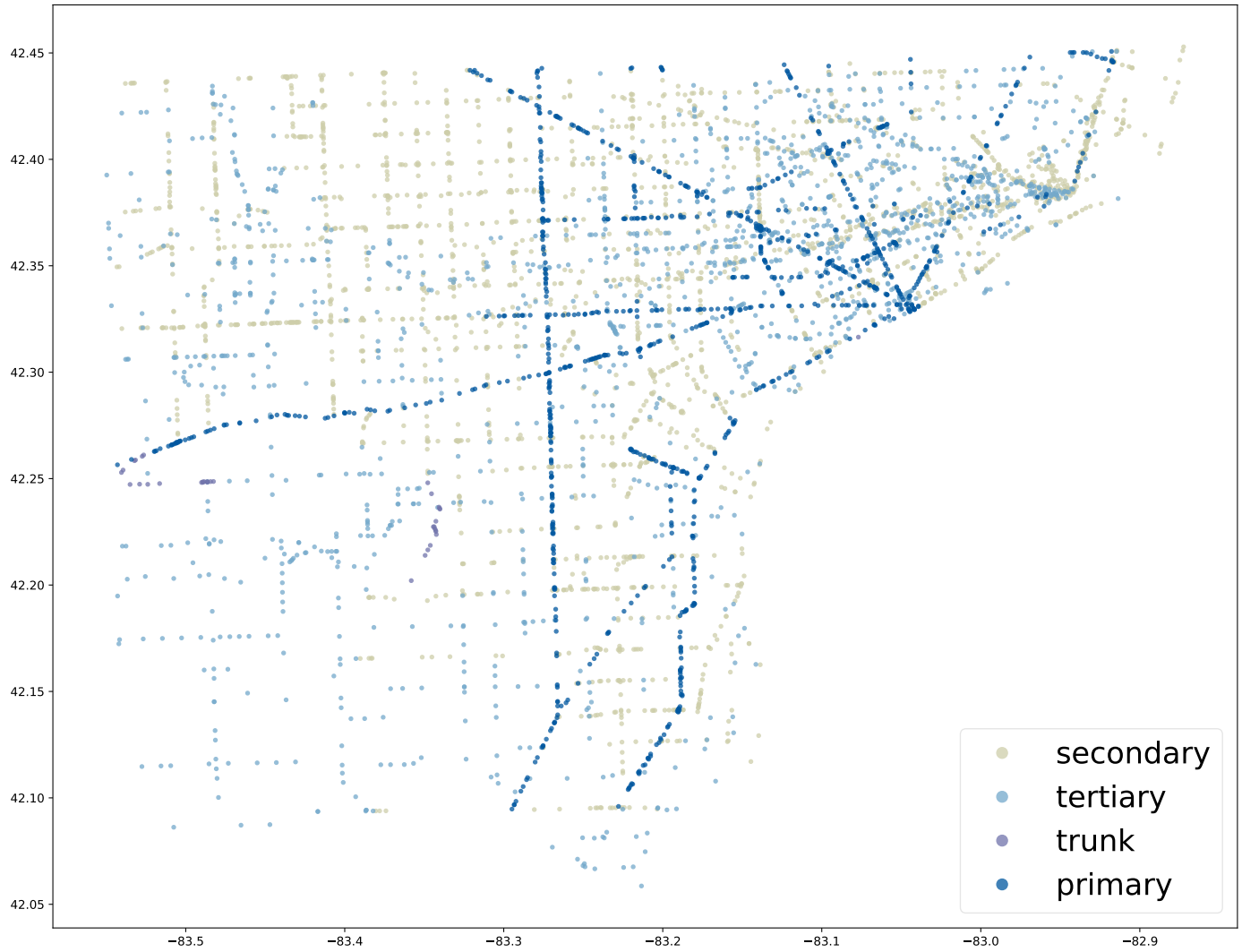


Figure SI 2.3: Sampled Locations in Lagos

Figure SI 2.4: Sampled Locations in Wayne, MI



SI 3 Mturk Surveys

Figure SI 3.1: Screenshot of the Bangkok, Jakarta, and Wayne Questionnaire

Does the road have any...?

- ☐ Cracks
- ☐ Potholes

Do you see any of the following?

- ☐ Litter
- ☐ A paved sidewalk

Road markings are painted marks on the road to guide traffic, e.g. pedestrian crossing, lane dividers, etc.

If there are any road markings, are they clear?

- ○ Yes
- ○ No

Submit

Figure SI 3.2: Screenshot of the Lagos Questionnaire

Does the road have any...?

- ☐ Cracks
- ☐ Potholes

Do you see any of the following?

- ☐ Litter
- ☐ A paved sidewalk

The photo is from google street veiw. Portions of the image may be blurred to preserve privacy. Road markings are painted marks on the road to guide traffic, e.g. pedestrian crossing, lane dividers, etc.

If there are any road markings, are they clear?

- ○ Yes
- ○ No

The photo is from Google Street Veiw. Google blurs portions of the image to preserve privacy.

Submit

Figure SI 3.3: Screenshot of the First Jakarta Questionnaire

Road markings are painted marks on the road to guide traffic, e.g. pedestrian crossing, lane dividers, etc.

Are the road markings clear?

- Yes
- No
- No road markings

Are there any potholes in the road?

- Yes
- No

Are there any cracks in the road?

- Yes
- No

Submit