Using Urban Density Data to Predict Multipath GPS Error in NYC Taxi Data

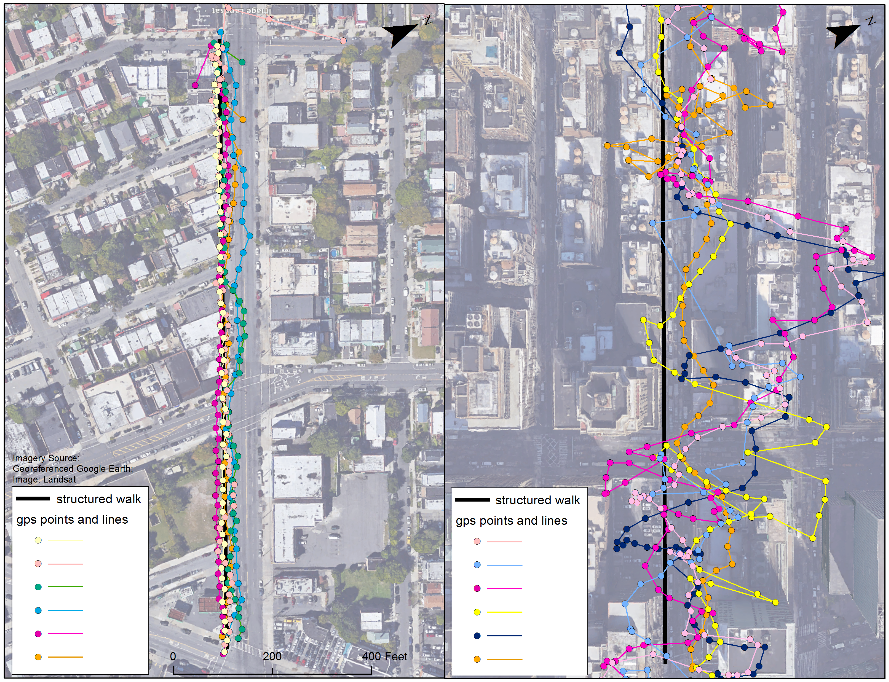
Danny, Emilie, Richie, Sankara

**Introduction**

In the past decade, renewed interest in the relationships between health and place and the development of smaller, cost-effective technologies, has increased the use of global positioning systems (GPS) in epidemiology and public health research. Location-based data collection is now applied to a broad range of health-related topics including environmental exposures to pollutants, physical activity tracking, infectious disease mapping, and built environment studies (refs). Mobile health applications are also beginning to use GPS in interventions that monitor health indicators and manage chronic disease. For example, the iWander app, an android based program, helps individuals suffering from Alzheimer’s disease and dementia who may become lost by providing automatic audible navigation and caregiver notification of coordinates ([Sposaro, Danielson, & Tyson, 2010](#_ENREF_4)).

Several factors contribute to the accuracy of GPS data including atmospheric effects, hardware quality, and satellite timing and availability. In many settings, close approximation of location is adequate, (i.e. geolocating a specific address), but in other contexts, more exact measurement may be necessary. Accuracy is especially important when GPS is being used to monitor activity data such as walking. Small errors in location, compounded over time/distance, may lead to significant overestimation of distance traveled (see Figure 1.)

Figure 1.



In dense urban environments, GPS measurements may be even more prone to poor accuracy as a result of multipath error. Multipath occurs when GPS signals bounce off buildings and other structures causing minute increases in the distance that the GPS signal must travel to reach the receiver. Because location is calculated from the length of time a GPS signal takes to reach a receiver, split second delays caused by signal reflections can have a significant impact. Multipath problems can also result in locations that seemingly wander or jump as a signal is dropped and regained, again contributing to overall measurement error.

Given the increasing utility of GPS in health-related applications and research, we sought to evaluate the impact of building density on GPS positional accuracy in New York City. Using a publically available dataset of approximately 175 million taxi trips occurred in New York City during 2013, we hypothesized that increasing building density would be positively associated with errors in GPS accuracy. To our knowledge, this is the first analysis to examine urban density as a factor in postional accuracy using a large dataset.

**Methods**

Data

To examine the relationship between density and location accuracy we combined several publically available New York City datasets. Taxi trip data were originally obtained through a Freedom of Information Act Law (FOIL) request by and subsequently posted on Google BigQuery for others to access ([Google BigQuery](#_ENREF_3); [Whong, 2014](#_ENREF_5)). The dataset included roughly 175 million observations, representing all of the yellow cab taxi rides occurring in New York City between January 1, 2013 and December 31, 2013. Variables included date, medallion, hack and vendor ID numbers, rate code, store and forward flag, pick-up and drop-off dates and times, passenger number, trip duration and distance, as well GPS-based latitude and longitude of pickup and drop off locations. Fare information for each trip was available in a separate dataset but not included in this evaluation.

Geo-referenced city basemaps containing roadbed data were provided by the Department of Information Technology & Telecommunications (DoITT) accessed in the NYC OpenData Portal ([Department of Information Technology & Telecommunications (DoITT)](#_ENREF_2)). New York City shapefiles containing 2010 census block data were obtained from the Department of City Planning (DCP) website ([Department of City Planning (DCP), 2014](#_ENREF_1)). The maps cover over 12 billion square feet of area and include over one million building footprints.

Geoprocessing

GPS accuracy was operationalized as a measure of distance from the roadbed using latitude and longitude of pick-up and drop-off mapped to the shapefile. Locations within 3 feet of the roadbed were not classified as GPS errors and were not included in the final models. Street centerlines typically bisect blocks and as a result 0.00007% of the location points were included in two census blocks. This misclassification was generally considered insignificant, however double-counted GPS points were only included in exploratory, rather than the final, analysis.

The relative building density, or the distributed building height (DBH), for each block was calculated by summing the volume for each census block and dividing by the area. Distributed building height can be considered in terms of uniform height of the block in feet. Shoreline? Edges? Census block with a distributed building height less than X ft., typically representing large parks, beaches, and cemeteries, were excluded from analysis. Something about size of block and how that would impact measure? Geoprocessing was conducted in python…?

Statistical Analysis

Preliminary data analysis examined the shape and distribution of the data. Summary statistics and weighted summary statistics were calculated for distributed building volume and distance to roadbed. Initial analysis explored the relationship between median/mean density and GPS error using the ggplot2 and BigVis packages in R. We also examined the shape and distribution of each variable using an interactive shiny application (). Each variable was inspected for data entry errors, missing values, outliers and unexpected patterns.

Continuous data on distance to roadbed was initially dichotomized to reflect any level of GPS error versus none. Binomial logistic regression models were created predict any GPS error as a function of building density. Linear regression analysis was also used to examine the linear relationship between building volume and distance from roadbed. We hypothesized that distance to the roadbed increases as additional units of density are include, regardless of whether the total density is low or high.

Because data pick-up and drop-off points tended to cluster by census block, we also explored spatial techniques to account for autocorrlation. The spatial distribution of GPS errors was evaluated at the census block level. The moran.test() function in the R spdep package was used to determine whether the pattern was clustered, dispersed or random. to examine and visualize the degree and range of autocorrlation. The lm.morantest function was used to adjust for spatial autocorrelation applied to regression residuals. Simple kriging, which treats clusters similar to simple points, was used to predict error from building density.

**Results**

Of the 38,716 census blocks examined, 33,562 (86.69%) had sufficient building height and count information to be included in the analysis. Blocks containing no buildings or missing building information (n=5,154) were removed. An average block was comprised of about 32 building with 6,974 blocks containing only one building. Less than 20% of blocks contained 100 or more buildings. To examine the relative density of block a measure of the percent of the block covered by the building footprint was created. Mean percent coverage was 28.3% indicating that on an average block, ???. I forget what this meant.

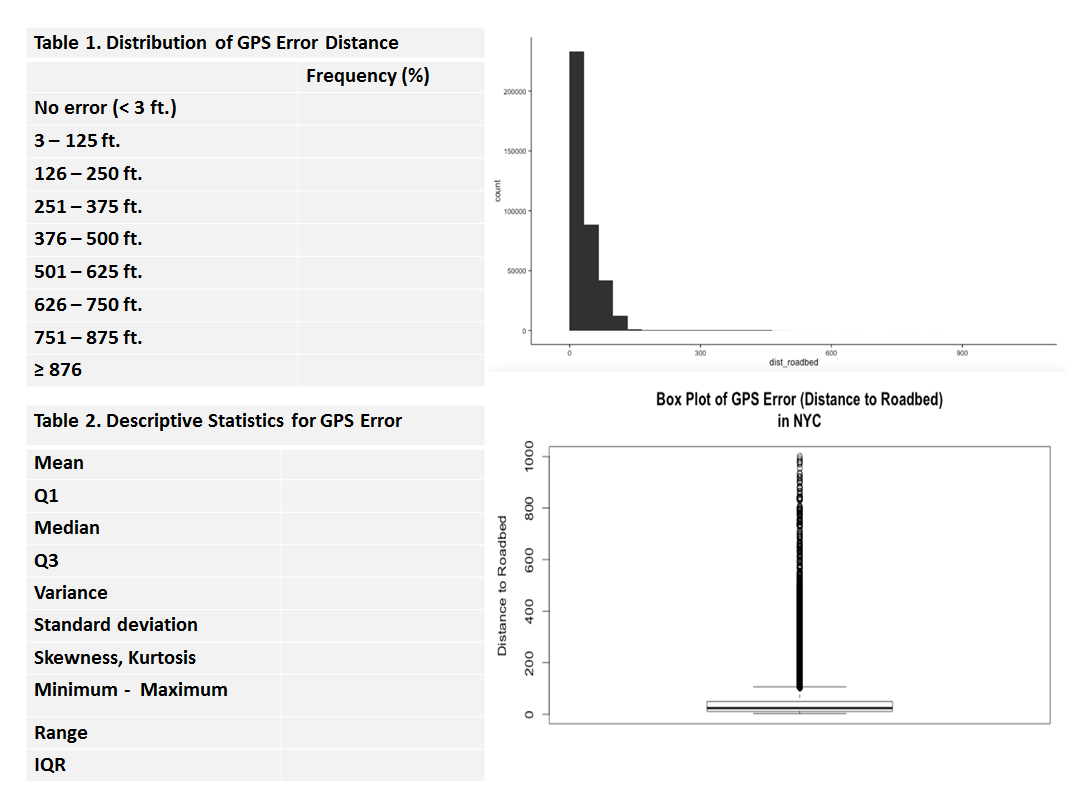
Descriptive statistics for distributed building height and average distance from the roadbed are presented in tables 1 and 2. The average number of pick up and drop off points per block was 8,668; the maximum was 5,654,000. The final spatial analytic data set consisted of 32,790 census blocks.

Results from the spatial analysis suggest a low degree of autocorrelation at the census block level. Adjusting for spatial variation, **every 1-unit increase in the distributed building height was associated with a 17% increase in distance from roadbed (95% CI = 1.15–1.19).** CHECK THIS – is it 117

Preliminary analyses were conducted to ensure no violation of assumptions. The spatially structured conditional autoregression term demonstrated areas of spatial patterning and similarity among census blocks.

**Discussion**

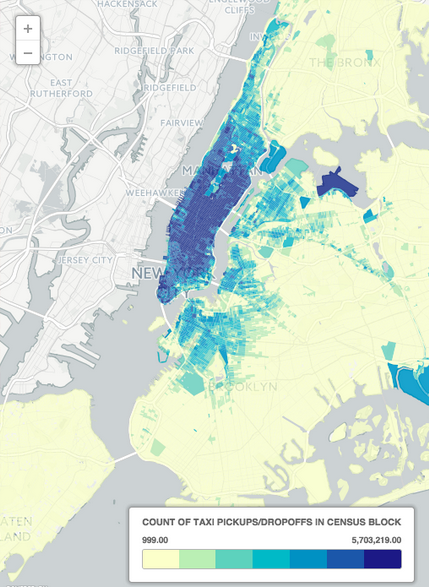
This analysis adds to these types of assessments by describing local patterns of GPS error in New York City at the census block level. There were areas in each borough of New York City (eg, lower Manhattan and Midtown, western Queens, and north Brooklyn) with high levels of distributed building volume—indicating that GPS errors may be significant even in locations not necessarily thought of as high density areas. This kind of analysis can inform geographically dependent public health research and interventions, by identifying areas that may require additional measurement or adjustment in the analysis.



Need Same for Distributed Building Height

Of the total XX pick-up and drop-off points, X% were located in Manhattan, followed by XYZ. The relative concentration of trip start and endpoints is shown in figure 2.

Figure 2. Density of Trip Start and End Point Counts by Block



Correlations

Regression models

Logit

Linear

Mult?

Geostatistical results

Moran’s and variogram

Spatial regression

Kriging

Discussion

References:

Department of City Planning (DCP). (2014). Retrieved April 21, 2015, from <http://www.nyc.gov/html/dcp/pdf/bytes/nybb_metadata.pdf?v=15a>

Department of Information Technology & Telecommunications (DoITT).). Roadbed. from https://data.cityofnewyork.us/City-Government/Roadbed/xgwd-7vhd

Google BigQuery.). Taxi. from https://bigquery.cloud.google.com/table/833682135931:nyctaxi.trip\_data

Sposaro, F., Danielson, J., & Tyson, G. (2010). iWander: An Android application for dementia patients. *Conf Proc IEEE Eng Med Biol Soc, 2010*, 3875-3878. doi: 10.1109/IEMBS.2010.5627669

Whong, Chris. (2014). Taxi TechBlog 1: Data Prep and Backend. from <http://chriswhong.com/data-visualization/taxitechblog1/>