MindRider Geographic Analysis: Summary

March 2017; Updated August 2017

By Jennifer Sta. Ines, Arlene Ducao, Ilias Koen

team@multimerdata.com

Our first pilot in 2014 (http://book.multimerdata.com) was conducted with ten Data Collectors, who, for the purposes of controlled wide-area coverage, all collected data by riding bicycles rather than by driving or walking. Each Data Collector signed up to ride one or more of Manhattan's major streets one time, ensuring one-time coverage of the area. This data collection was focused on the utility of MindRider data for the analysis and evaluation of the current cycling infrastructure. The pilot demonstrated promising results based on a simple proximity analysis, in which areas of urban infrastructure, traffic and safety concern were evaluated against the aggregate of EEG data points within 100 feet of those areas. For the purposes of simplicity, we isolated *Hotspots*, which are locations of EEG values indicating high concentration levels (a possible indicator for perception of danger), and then we separately isolated *Sweetspots*, which are locations EEG values indicating high relaxation levels (a possible indicator for perception of safety). For the areas of traffic and safety concern:

- Bike collisions: Hotspot counts were 72% higher than Sweetspot counts within 100 feet of where a bike collision took place in 2014.
- Fatal collisions: Hotspot counts were 300% higher than Sweetspot counts within 100 feet of where a fatal collision took place in 2014.
- Truck/HGV (Heavy-Goods Vehicle) Routes: Hotspot counts were 27% higher than Sweetspot counts within 100 feet of an HGV route.
- Traffic complaints: Hotspot counts were 333% higher than Sweetspot counts within 100 feet of where a traffic complaint took place in 2014.

Our second pilot in 2016 was conducted with a larger sample size of 50 pedestrians and cyclists. Driver data was also experimentally collected outside of the pilot project. It covered a smaller geographic area than the 2014 pilot, but it expanded the collected data's spatiotemporal resolution with data repeatedly collected from major streets. The second pilot tested the feasibility of the fleet deployment model, which involved a larger and less controlled sample size: Data Collectors were allowed to ride streets most convenient to them so as to collect more data more structure in terms of community meetings, data management, safety training, compensation/incentivization and informed consent. iterative user experience design to account for very long sessions/rides of data collection.

The analysis of the 2016 data involved conducting spatial regression and correlation tests with the MindRider data (set as dependent variables) against data from the City of New York (set as explanatory variables). From a preliminary analysis:

- Using the Spatial Autocorrelation (Global Moran's I) tool in ArcGIS, the z-score (standard score) for Hotspot areas was calculated as 8.16 indicating that there is less than a 1%

chance (99% confidence level) that this clustered pattern is the result of random chance in our study area of Manhattan.
- Using the Spatial Autocorrelation (Global Moran's I) tool in ArcGIS, the z-score for areas of Sweetspots was calculated as 2.18, with a p-value (probability value) of 0.03. This indicates that there is a less than 5% likelihood (95% confidence level) that this clustered pattern is the result of random chance in the same study area.
- Ordinary Least Square (OLS) Regression analysis was run to help explain areas of Hotspots and Sweetspots in the study area. The explanatory variables used to model these phenomena included some of the same variables that were analyzed during the 2014 pilot: locations of collisions*, traffic complaints, 311 calls logged to NYPD, 311 calls logged to DOT*, truck/HGV routes*, and bike routes*. Variables marked with * an asterisk were found to have statistically significant p-values, indicating that these variables might help the model.

All of the statistically significant explanatory variables had a negative relationship to the dependent variable, with the exception of 311 calls logged to DOT (positive). This was the case for both Hotspots and Sweetspots. Further exploration will occur during Phase I.
The adjusted R-squared value was calculated as 0.209 when OLS regression was run to explain the areas of Hotspots. When the model was run to explain areas of Sweetspots, the adjusted R-squared value was calculated as 0.164. The Spatial Autocorrelation tool was run on both OLS regression residuals to confirm that the models are not biased--this is also indicated by the non-statistically significant Jarque-Bera value.

A few considerations could be taken into consideration to increase model performance:
- Adding more explanatory variables, such as primary land use of adjacent properties
- Refining existing variables, such as filtering the complaint type of 311 calls to DOT to focus on street conditions, or traffic signal conditions.
- Fitting a new OLS regression into a smaller focus area within the study area, such as blocks with the highest building heights

The 2016 pilot dataset, which was more than thirty times that of the dataset from 2014, was the base for our feasibility study for our analytical product platform, Multimer. Like in our 2014 pilot, we isolated EEG values indicating Hotspot levels and Sweetspot levels.