ESTIMATING BICYCLE TRIP VOLUME FOR MIAMI-DADE COUNTY FROM STRAVA TRACKING DATA

Hartwig H. Hochmair
University of Florida, Geomatics Program
3205 College Avenue
Ft. Lauderdale, FL-33314
Tel: 954 577-6317 Fax: 954 475-4125; E-mail: hhochmair@ufl.edu

Eric Bardin
University of Florida, Geomatics Program
3205 College Avenue
Ft. Lauderdale, FL-33314
Tel: 954 577-6392 Fax: 954 475-4125; E-mail: bardin@ufl.edu

Ahmed Ahmouda
University of Florida, Geomatics Program
3205 College Avenue
Ft. Lauderdale, FL-33314
Tel: 954 577-6392 Fax: 954 475-4125; E-mail: ahmouda12@ufl.edu

Word Count: 6379 words text + 4 tables/figures x 250 words (each) = 7379 words

Revised version submitted: November 15, 2016
ABSTRACT
Over the past few years, sports and fitness apps on GPS enabled cell phones provided a significant growth in the volume of GPS tracking data. These crowd-sourced data can be analyzed to better understand the cycling behavior of a large user community. Using Strava activity tracking data from the Miami-Dade County area, the goal of this study is to identify which sociodemographic factors, network measures, in particular on-road bicycle facilities, and place specific characteristics influence bicycle ridership. For this purpose, a set of linear regression models are estimated to predict non-commuter and commuter bicycle kilometers travelled, as well as bicycle kilometers travelled on weekends and weekdays. Eigenvector spatial filtering is applied to explicitly model spatial autocorrelation and to avoid parameter estimation bias. Results suggest that Strava data, due to its high spatial and temporal resolution and coverage, can identify how the influence of explanatory variables on estimated bicycle trip volume varies between different trip purposes and days of the week. Strava tracking data can therefore be considered a useful supplement to other bicycle count systems to estimate cycling volume over large areas.

Keywords: bicycle volume, tracking data, Strava, eigenvector spatial filtering
INTRODUCTION
Large cycling volumes with sufficient spatial detail and temporal coverage are necessary to understand temporal fluctuations and well as spatial variation of bicycle ridership (1). The growth of fitness apps, such as Strava, Endomondo, or MapMyRide, which are operated on smart phones, smart watches, or fitness bands, have increased the sampling size of users to the millions (2). These apps provide data coverage for a large area on a continued basis, which makes app-based crowdsourcing methods particularly useful for transportation planning tasks, such as the identification of network locations that need construction of new bicycle infrastructure or improvement of existing bicycle infrastructure. While crowd-sourced data have their drawbacks, their sheer volume and coverage outweighs its disadvantages for many types of analyses if data are properly handled to minimize biases in final products and maps (3). Crowd-sourced tracking can supplement permanent count stations currently installed in the U.S that show long-term changes and trends in cycling frequency but lack spatial detail (4).

Using a set of linear regression models combined with eigenvector spatial filtering (ESF) to handle spatial autoregression in residuals, the research of this paper extends previous work presented in the literature by analyzing how explanatory variables of the road network, the built environment, and the natural environment affect non-commuter and commuter cycling volume, and how their role varies between the prediction of weekend and weekday cycling volume. For the purpose of this research 2015 Strava Metro data from Miami-Dade is used at different temporal summary levels, i.e. total annual rollup (for five months) and weekday/weekend data for one month.

LITERATURE REVIEW
The role of the road network structure, bicycle facilities, and the built environment on cycling demand has been studied for many years, considering various data sources that include household travel surveys, interviews, census data, and more lately GPS tracking data. A recent longitudinal analysis that used bicycle commuter data from the 2000 and 2010 census at the block group level identified a significantly larger increase in bicycle commuters in block groups that had on-road bicycle lanes installed during the study period compared to block groups with sharrow or no infrastructure installed. It showed also that the longitudinal increase in bicycle commuters for block groups that had only sharrors installed was not statistically significant, raising questions about the effectiveness of sharrow (5). Researchers also used a combination of network-based roadway characteristics along with contextual attributes of the neighborhood design and the surrounding built and natural environment to model expected cycling demand and travel choices, finding, for example, that block size, mixed land use, and bicycle friendly design are associated with the decision to ride a bike (6, 7). Several studies found that the availability of bike lanes was associated with more cycling, that higher levels of street connectivity were associated with more cycling for utilitarian trips (8), and that building new bike lanes increases cycling ridership (9, 10). A field study in Minnesota showed that bicyclists travel, on average, 67% longer in order to include a trail facility on their route (11).

A detailed review of the history and growth in data-driven research on cycling is provided in (2), dividing the review into three groups according to the nature of data collection, which are GPS data (collected via smart-phone and other GPS enabled mobile devices), live point data (collected at a particular location such as from a traffic camera), and journey data (providing origin and destination locations and times for an individual journey, but no route geometry). The paper provides also some user numbers for commonly used fitness apps that collect tracking data,
including Endomondo, Strava, and MapMyRide. It reports, for example, that 2.5 million GPS-tracked activities are uploaded to the Strava website every week. Data from sports tracking apps have recently been used to identify the effect of various characteristics of the built and natural environment on bicycle ridership numbers (12), identifying place based variables, such as the balance of job and housing, and plexus-based variables, such as bicycle lane length in meters, as significant predictors of bicycle kilometers travelled by census block group. The use of activity trackers can be motivated by a variety of tasks, including directive tracking (goal driven, e.g. to lose weight), documentary tracking (e.g. to document special walks), or collective rewards (e.g. score points) (13).

Although crowd-sourced tracking data provide an opportunity to obtain a large volume of data, as with all crowd-sourced data, they lack data quality assurance compared to data from traditional collection sources, such as governmental data (14). In this regard several papers address the issue of user bias with sport tracking and fitness apps. A comparison of the data obtained from two smartphone-based apps, namely Cycle Atlanta and Strava, for a selected area in midtown Atlanta, revealed that the skew for male users is greater for Strava (84%) than for Cycle Atlanta (76%) (15). The authors conclude that transportation planning and design analysis should carefully take into account the likely bias from the self-selected users of such apps. Similarly, another editorial paper points out that sports tracking data is affected by self-selection bias, and that census data provide a sampling framework against which the biases of Big Data can be estimated (16).

Several studies analyzed how well fitness app data represent actual ridership through comparison with manual cycling counts. Research conducted in Victoria, British Columbia, found moderate correlations ($r^2$ between 0.40 and 0.58) between crowdsourced data volumes from Strava and manual counts (1). Using a set of linear regression models another study compared flows obtained from Strava (provided in minute intervals) with those from the London Cycle Census (LCC) obtained at 164 survey sites (reported in 15-minute time periods). It concluded that that Strava flows correspond well to the LCC flows (17).

STUDY SETUP

Study area and data sources
The study area for this research comprises the Urbanized Area of Miami-Dade County as defined by the US Census Bureau, as well as some additional parts of the agricultural land and the built environment stretching into the south-western part of the county towards the Everglades National Park. Bicycle ridership is obtained from the Strava Metro product. These data come in a shapefile format and aggregate user rides for road segments, where data are sampled to HERE NAVSTREETS road geometries. Aggregated count data are available for different time periods (rollups). For the conducted study, three types of bicycle ridership summaries from Strava are used for the following time periods:

1. Year 2015 (data were available for January-May)
2. February 2015 Weekend
3. February 2015 Weekday

These ridership summary levels are further broken down into total trips per segment, and commuter trips per segment.
FIGURE 1a maps the Strava total trips count for January-May 2015 for the study area. According to the Strava Metro User Guide, Strava derives commuter data by three methods: (1) a commute flag; (2) an automated process that locates point-to-point cycling trips that are within duration and distance constraints; (3) fuzzy name matching from the activity titles. Strava counts per segment can then be multiplied by segment length and summed up to compute bicycle km travelled (BKT) in analyzed areal units (12).

FIGURE 1b shows the result of this process based on total trip counts between January-May 2015 for 1544 block groups in Miami-Dade County. These block groups denote the spatial extent for which Strava data were used to run different regression models described further below.

FIGURE 1 Strava January-May 2015 summary data: Total trip count on street segments (a) and derived bicycle kilometers travelled (BKT) for census block groups (b).

Sociodemographic data for the predictor variables were taken from the 5-year (2009-2014) summary American Community Survey (ACS) data at the block group level. Job data were obtained from the Census Bureau LEHD Workplace Area Characteristics at the census block level and then aggregated to census block groups. The HERE NAVSTREETS 2015 Quarter 1 dataset is used to aggregate network supply measures, such as the total length of local roads per block group. It is also used to derive topological or morphological network measures, such as betweenness centrality for network segments, which can also be aggregated for block groups. On-road bicycle facilities information, such as marked bicycle lanes, was originally obtained from the Miami-Dade
Metropolitan Planning Organization (MPO) and subsequently updated for an earlier research study that assessed the completeness of bicycle trail and lane features in OpenStreetMap (18). This facilities dataset was once more updated by the authors in spring 2015 through manual comparison of mapped facilities with Google Maps Aerial View and Google Street View imagery for the entire Miami-Dade County area.

**Analysis methods**

This research explores the relationship between bicycle ridership (R) and road network characteristics (N), location specific characteristics (L), and sociodemographic variables (S) at the block group level. A stylized representation of this relationship in functional form can be given as

\[
R = f(N, L, S) \tag{1}
\]

where: R is bicycle kilometers travelled (BKT) in a census block group (or the difference in z-scores of bicycle kilometers travelled between two trip purposes or groups of days, respectively), N is measures of network characteristics within a block group (e.g. total length of arterial roads), L is other location specific characteristics of the block group (e.g. contains a park with dedicated bicycle tracks), and S is sociodemographic characteristics within a block group (e.g. median household income). A set of four linear regression models is used to explore these relationships. Whereas the set of predictor variables on the right side of Equation 1 stays unchanged between these four models to make them comparable, the variable that is to be predicted changes as follows:

- **Model 1:** This model uses the five-month summary of Strava ridership data (January-May 2015) to predict the BKT for non-commute trips at the census block group level
- **Model 2:** This model uses the same data set to predict the BKT for commute trips
- **Model 3:** This model predicts the difference between the non-commute BKT z-score and commute BKT z-score values at the block group level, and is also based on the five-month summary of Strava ridership data
- **Model 4:** This model predicts the difference between weekend BKT z-score and weekday BKT z-score values at the block group level, using February 2015 Weekend and Weekend summary count data. A separation into weekend and weekday ridership is not available for the annual Strava count data, hence monthly count data were used instead. February was selected since it has no major holidays and thus reflects a rather typical bicycle travel pattern.

A fundamental assumption of regression analysis is residual independence. Incorporating spatial data in nonspatial models typically results in residuals that exhibit spatial autocorrelation. Ignoring spatial dependence in spatial data can lead to coefficient estimation bias and biased standard errors, or both (19). Therefore, to obtain unbiased coefficient estimates and correct inference, spatial autocorrelation needs to be explicitly modeled in statistical analysis (20). Various types of spatial regression models exist that explicitly take into account the spatial structure of data, including spatial autoregressive modeling, conditional autoregressive modeling, or spatial filtering (21). In this case study, for each linear model an additional spatially filtered linear model is estimated which uses eigenvector spatial filtering (ESF) to model spatial autocorrelation. ESF utilizes eigenvector decomposition to extract a set of eigenvectors from the spatial weight matrix that is incorporated in the numerator of the Moran’s I coefficient (22). A spatial filter that comprises all relevant eigenvectors can then be used as a predictor variable in standard statistical techniques,
such as linear regression models. This combination of relevant eigenvectors is expected to explain a considerable part of the variance in the distribution of cycling volume across the study area. A detailed description of steps involved in ESF, as applied in this study, can be found in (23, 24).

**EXPLANATORY VARIABLES**

**Network measures**

Network measures can be broadly categorized into four categories, which are hierarchy, topology, morphology and scale (25, 26). Hierarchy measures the heterogeneity of a street network, such as number of changes in street hierarchy experienced along a fastest path from the trip origin to the trip destination. Topology provides measures of connectivity in the network that are based on elementary concepts of graph theory. Morphology describes the regularity of street networks as well as their shape and fragmentation. Scale captures the network supply in a particular area. This study applies measures from the last three categories and omits hierarchical measures since the network measurements are derived for block group units and do not use travel paths.

Network supply is measured as total road length per census block (in meters), separated into local, collector, and arterial roads. Road categories were derived from the NAVSTREETS road data set after removal of road segments that were not accessible to bicyclists (e.g. interstates, ramps, or closed port areas that showed no Strava counts). Roads that had a FUNC_CLASS attribute value of 2 or 3 in NAVSTREETS were classified as arterial roads, those with a FUNC_CLASS value of 4 as collector roads, and those with a FUNC_CLASS value of 5 as local roads. In addition to road lengths, all combinations of road classes (local, collector, arterial) and considered types of on-street facilities (marked bike lane, paved shoulder, shared lane marking (“sharrows”)) were included as supply variables as well, adding a total of 9 predictors to the regression models.

**FIGURE 2a** maps for part of the study area the location of on-road bicycle facilities. Off-road network supply was measured as bike/walk trails where motorized traffic was not permitted. These road segments could be extracted from the NAVSTREETS attributes through an "AR_PEDEST" = 'Y' AND "AR_AUTO" = 'N' query. Strava counts were often observed on off-roads paths that are typically designed for pedestrians, such as walking paths in parks, which justifies the use of bike/walk segments as a supply variable.

Regarding network topology, the importance of an edge in a network can be characterized by its centrality, for which many indicators exist (27). A commonly used indicator, the betweenness centrality of an edge, is the number of the shortest paths that go through an edge in a graph or network. An edge with a high edge betweenness represents a bridge like connector between different part of the network (28). For this study the edge betweenness was computed on the NAVSTREETS dataset under consideration of one-way restrictions using the R-package igraph, where a maximum path length of 40 km was chosen. To avoid an underestimation of the edge-betweenness near the boundaries of the study area, this computation was performed on a road network that was extended beyond the selected Miami-Dade County study to a 40 km buffer around it.

**FIGURE 2b** visualizes for the Miami-Downtown, Miami Beach and Key Biscayne areas the obtained betweenness centrality values. In general, if a road is represented by split geometries in both driving directions, the betweenness centrality for each direction edge is smaller than if the road is represented as a single line geometry, everything else equal. This explains why the Venetian Causeway that crosses Biscayne Bay between Miami on the mainland and Miami Beach on the
barrier island (the second bridge from the top in the figure) and that is represented as a single line geometry for both driving directions is shown in darker color than other nearby bridges crossing Biscayne Bay with split lanes, such as the MacArthur Causeway to its south. For the regression models, the maximum edge betweenness value on any road segment within the block group polygon was used as the network measure for the corresponding block group polygon.

The shapefactor is a measure of network morphology which captures the generated impedance of the street network. Its computation involves identifying all polygons enclosed by the street network. For each polygon the shapefactor (P2A) can then be computed as (perimeter squared) / area, where higher values indicate more elongated or more winding polygon shapes, leading to greater impedance in circumnavigating the network. For this study for each block group polygon the maximum P2A value from all road polygons overlapping with the block group polygon was chosen as the shapefactor measure. Another morphological measure is based on the size of the street polygons overlapping with a block group polygon, where a large size means fewer route options to travel through the area and thus greater impedance. The area of the largest road polygon overlapping with the block group polygon of interest was chosen as a second morphology measure.

![FIGURE 2 On-road bicycle facilities (a) and betweenness centrality (b) for part of the analyzed road network.](image-url)
**Location specific variables**

This group of measures describes other control variables at the block group level that are expected to affect cycling ridership but are not directly related to the network structure or sociodemographic characteristics. Variables considered in regression models are:

- Presence of major park with dedicated bicycle facilities, i.e. bicycle trails (binary variable). Parks include both state and county parks.
- Located on a bridge that crosses Biscayne Bay (binary)
- Distance to bay or ocean
- Distance to Miami-Dade Central Business District (CBD)
- Number of jobs
- Mean NDVI (Normalized Difference Vegetation Index), derived from 30m resolution Landsat Thematic Mapper 5 satellite imagery. Only land was considered in the computation of the mean NDVI for block groups, meaning that bay area, ocean, and inland water was removed before processing.

**Sociodemographic variables**

The set of socioeconomic variables that are expected to potentially affect ridership volume at the block group level and therefore considered in the regression models include:

- Total population
- Median household income
- Median age
- Percent African American population
- Percent Hispanic population
- Percent male population

Block groups that were missing any of the sociodemographic variables (e.g. median household income), had zero population, or were lacking roads necessary to derive road polygon characteristics (e.g. shapefactor), were excluded. After this process a total of 1544 block groups was retained for regression analysis out of 1594 original block group polygons for the entire county.

**REGRESSION ANALYSIS**

**Processing steps**

To ensure that considered explanatory variables capture different aspects of block group characteristics in the network, location-specific and sociodemographic realm, Spearman’s rho correlation coefficient was computed between all explanatory variables. Percent African American and percent Hispanic were strongly negatively correlated ($r= -0.71, p<1 \times 10^{-15}$), and total length of local and collector streets were strongly positively correlated ($r= 0.71, p<1 \times 10^{-15}$). This led to the removal of percent Hispanic and total length of collector street variables. In addition to this check, collinearity diagnostics were run on all final models. The variance inflation factor was < 2 for each variable in each model, indicating that multi-collinearity was not a concern in this dataset.

As a first step to determine the final regression model design, nonspatial linear regression models (that is, without eigenvector spatial filtering) were built for the four predicted variables relating to BKT described earlier (i.e., non-commute BKT, commute BKT, z-score difference
between non-commute and commute BKT, and z-score difference between weekend and weekday BKT). Including all predictor variables at the beginning, variables that did not significantly contribute to the model fit (measured by adjusted R²) for any of the four predicted variables were removed one by one in a stepwise manual approach. In a second step, using the remaining subset of predictor variables, ESF was applied to all four linear models, rendering some predictor variables non-significant in all four ESF linear models. In the third step, the ESF linear models were re-run using only the same set of variables that were significant in at least one model from the previous step. While this approach does not necessarily maximize the adjusted R² for each model, it allows to compare the effect of predictor variables on the different predicted variables side by side. Finally, the nonspatial linear models were re-run using this final set of predictor variables, which allows to assess the effect of the spatial filter on the magnitude and the standard error of regression coefficients. For each model the Moran’s I coefficient together with its p-value is reported as a measure of autocorrelation in the regression residuals.

The next two sections present a total of eight model estimations, which four nonspatial models with their corresponding spatially filtered linear models. The first section analyses BKT associated with non-commute trips (Model 1) and commute trips (Model 2), and the second section analyzes differences in BKT for different trip purposes (Model 3) and different days of the week (Model 4), respectively. Since the spatial models address autocorrelation their coefficients are less biased than those of nonspatial models, and therefore the discussion will mostly be on results of the spatial models.

**Analysis of cycling volume for non-commute and commute trips**

TABLE 1 tabulates the results of model 1 (M1) and model 2 (M2). Comparison of regression coefficients and their levels of significance between nonspatial and spatial models for M1 and M2 reveals clear differences in estimation outcomes due to inclusion of the spatial filter in the spatial models. The coefficient associated with the spatial filter is highly significant both in M1 and M2. Moran’s I coefficients and their p-values demonstrate that the spatial filter eliminates all of the unexplained residual spatial autocorrelation (p > 0.28) which occurs in the nonspatial regression models (p < 1E-15). The spatially filtered linear models satisfy therefore the fundamental assumption of residual independence in regression models. The adjusted R² values listed at the bottom row of TABLE 1 show that the spatially filtered models reach a better model fit than their nonspatial counter parts, demonstrating that a large part of the variance in bicycle trip volume can be explained by the eigenvectors.

Under the sociodemographic variables, M1 and M2 show that both non-commute BKT and commute BKT are negatively associated with percent male population. Given that the users signing up for Strava are mostly male, this result is somewhat unexpected. However, a closer look at the spatial distribution of this variable across the county shows that block groups with a high percentage of male residents are found at military bases (e.g., Homestead) and in housing near airports, industrial areas, and quarries, which are generally not attractive for cycling and may therefore lead to the negative coefficients. Analysis of location specific variables shows that block groups with bicycle parks attract more non-commute (e.g. leisure) cycling (M1) but not more commute cycling (M2). Location of a block group polygon along any of the Biscayne Bay bridges increases both non-commute and commute BKT, probably due to the nice scenery associated with travel over any of the bridges. Furthermore, both models indicate that non-commute and commute cycling occurs primarily in close proximity
to the bay area or the ocean, which matches the visual perception of the Strava trip count pattern on road segments shown in FIGURE 1a.

**TABLE 1 Predicting Non-commute and Commute Bicycle Kilometers Travelled (BKT) Using Nonspatial and Spatially Filtered Linear Models**

<table>
<thead>
<tr>
<th></th>
<th>BKT Non-Commute (M1)</th>
<th>BKT Commute (M2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Nonspatial</strong></td>
<td><strong>Spatial</strong></td>
</tr>
<tr>
<td><strong>Sociodemographic variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median household Income</td>
<td>6.680E-03***</td>
<td>2.239E-04</td>
</tr>
<tr>
<td>Percent male</td>
<td>-2.343E+01**</td>
<td>-2.442E+01***</td>
</tr>
<tr>
<td><strong>Location specific variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bicycle park</td>
<td>3.990E+03**</td>
<td>5.148E+03***</td>
</tr>
<tr>
<td>Bay bridge</td>
<td>2.971E+03***</td>
<td>1.818E+03***</td>
</tr>
<tr>
<td>Distance to bay or ocean</td>
<td>-9.148E-02***</td>
<td>-4.044E-02**</td>
</tr>
<tr>
<td>Jobs</td>
<td>3.055E-03</td>
<td>-1.875E-03</td>
</tr>
<tr>
<td><strong>Network measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum shapefactor (P2A)</td>
<td>1.832*</td>
<td>0.804</td>
</tr>
<tr>
<td>Maximum polygon area</td>
<td>1.106E-06</td>
<td>8.50E-07</td>
</tr>
<tr>
<td>Max. betweenness centrality</td>
<td>2.545E-07</td>
<td>1.291E-07</td>
</tr>
<tr>
<td>Local road length</td>
<td>6.938E-02***</td>
<td>4.755E-02***</td>
</tr>
<tr>
<td>Local road length w. bike lane</td>
<td>1.930***</td>
<td>1.367***</td>
</tr>
<tr>
<td>Collector rd. len. w. bike lane</td>
<td>-6.875E-02</td>
<td>-0.235</td>
</tr>
<tr>
<td>Coll. rd. len. w. paved shoulder</td>
<td>1.097**</td>
<td>0.841*</td>
</tr>
<tr>
<td>Coll. rd. len. w. sharrows</td>
<td>0.660*</td>
<td>0.549</td>
</tr>
<tr>
<td>Arterial road length</td>
<td>-0.157*</td>
<td>-0.156**</td>
</tr>
<tr>
<td>Arterial rd. len. w. sharrows</td>
<td>1.935</td>
<td>-2.095*</td>
</tr>
<tr>
<td>Walk/bike-only trail length</td>
<td>0.270***</td>
<td>0.254***</td>
</tr>
<tr>
<td><strong>Spatial filter</strong></td>
<td>-</td>
<td>4.056***</td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>8.597E+02*</td>
<td>1.352E+03***</td>
</tr>
<tr>
<td>Number of observations</td>
<td>1544</td>
<td>1544</td>
</tr>
<tr>
<td>Residuals Moran’s I</td>
<td>0.1654***</td>
<td>0.0067</td>
</tr>
<tr>
<td>Residuals Moran’s I (p-value)</td>
<td>&lt; 1E-15</td>
<td>0.287</td>
</tr>
<tr>
<td>R²</td>
<td>0.283</td>
<td>0.454</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.275</td>
<td>0.447</td>
</tr>
</tbody>
</table>

Note: *** p<0.001, ** p<0.01, * p<0.05

Regarding network supply M1 shows that non-commute cyclists avoid block groups dominated by long stretches of arterial roads. This is not the case for commuter cyclists (M2) who may be more sensitive to travel time to get to their job on time and therefore choose to not travel an extra distance in order to avoid arterial roads with higher traffic volumes. In addition, non-commuter cyclists also tend to avoid arterial roads even if they have shared lane markings (sharrows) painted on them, whereas bike commuters do not. A possible explanation is that sharrows mandate cyclists...
and motorized traffic to share the same lane (though not contemporaneously), which might decrease the subjective perception of cycling safety more for leisure riders than for bike commuters. An increase in total length of bike lanes on local roads and paved shoulders on collector roads is associated with an increased BKT both for non-commuters and commuters (M1, M2). In additional to this sharrows on collector roads are associated with an increase in bike commuter BKT. Other network measures considered in the model, including topological and morphological measures, were not significantly associated with cycling volume in either of the two models.

Analysis of differences in cycling volume between different trip purposes and travel days
Before this analysis, the BKT values for the four different ridership data at the block-group level were standardized using z-scores. This way absolute ridership numbers were removed from each data set and only the spatial variation of z-scores across the study would remain in each data set. Subsequently, differences in z-scores were computed by subtracting z-scores of commuter BKT from those of non-commuter BKT, and by subtracting z-scores of weekday BKT from weekend BKT. These differences were used as the predicted variable in model 3 (M3) and model 4 (M4) to study the nature of differences in cycling volume between these different ridership groups.

TABLE 2 tabulates the results of both models. In M3 a positive coefficient indicates a relative increase in non-commute BKT compared to commute BKT, whereas a negative coefficient indicates a relative increase in commute BKT compared to non-commute BKT. One of the previously unnoticed (when comparing M1 and M2 estimations), but expected, effects of location specific variables is that census blocks with more jobs also attract more commuter BKT, suggesting that commuter bike rides start or end in census blocks which provide job opportunities. Census blocks along bay bridges are also associated with a relative increase in commuter BKT, which might be explained by the special role of bridges in the network layout as they connect different components of the network. This leaves commuters only few or no alternative routes to choose from when planning to cycle from the mainland to the barrier island (or the other way round) across Biscayne Bay.

With regards to network measures, M3 marks a measure of network morphology, namely maximum polygon area, as a significant predictor of BKT differences. The positive coefficient indicates that non-commute BKT increases in block groups that enclose large street-based polygons, which occur in the bay area as well as in rural, agricultural stretches of land in the southwest of the county and also in sparsely populated natural environments close to mangrove forests and swamps in the south-east corner of the county. In terms of network supply M3 shows that selected bicycle facilities on higher-speed traffic roads (collector, arterial) are on average more used by bike commuters, whereas bicycle lanes on local roads are associated with an increase in non-commute cycling volume. Furthermore, trails for non-motorized traffic increase non-commute cycling more than bike commuting.

In M4, a positive coefficient indicates a relative increase in weekend BKT compared to weekday BKT, whereas a negative coefficient indicates a relative increase in weekday BKT. In the location specific variables, M4 shows that presence of a bicycle park is associated with a relative increase in BKT during weekdays, indicated by the negative coefficient. Although this association is unexpected (one would typically expect more cyclists on weekends in these parks), this can at least be partially explained by parking fees that are charged on weekends, but not weekdays, for a major mountain bike park that is located in Amelia Earhart Park. These fees may deter cyclists from further away who need to use a car to reach the park to do so on weekends.
Block groups further away from the bay or ocean are associated with a relative increase in weekday ridership, meaning that weekend rides take place closer to these open water areas. Also, block groups with higher job numbers show a relative increase in weekday ridership which can be expected since many jobs categories (e.g. governmental, teaching) are weekday jobs. Block groups along bay bridges experience a relative increase in weekday ridership, which can be associated with many commuters traveling across these bridges, as discussed in connection with M3.

**TABLE 2 Predicting Differences in Bicycle Kilometers Travelled (BKT) Between Different Trip Purposes and Travel Days Using Nonspatial and Spatially Filtered Linear Models**

<table>
<thead>
<tr>
<th>Difference BKT Non-Commute minus Commute (M3)</th>
<th>Difference BKT Weekend minus Weekday (M4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonspatial</td>
<td>Spatial</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sociodemographic variables</th>
<th>Nonspatial</th>
<th>Spatial</th>
<th>Nonspatial</th>
<th>Spatial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median household Income</td>
<td>-1.010E-06***</td>
<td>-4.314E-07*</td>
<td>1.501E-08</td>
<td>4.100E-08</td>
</tr>
<tr>
<td>Percent male</td>
<td>-3.043E-03**</td>
<td>-5.683E-04</td>
<td>5.330E-03**</td>
<td>3.552E-03*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Location specific variables</th>
<th>Nonspatial</th>
<th>Spatial</th>
<th>Nonspatial</th>
<th>Spatial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicycle park</td>
<td>1.838***</td>
<td>1.466***</td>
<td>-0.660</td>
<td>-0.790*</td>
</tr>
<tr>
<td>Bay bridge</td>
<td>-0.328***</td>
<td>-0.117*</td>
<td>-0.4.81***</td>
<td>-0.307*</td>
</tr>
<tr>
<td>Distance to bay or ocean</td>
<td>3.323E-06</td>
<td>-1.104E-07</td>
<td>-2.346E-05***</td>
<td>-1.170E-05***</td>
</tr>
<tr>
<td>Jobs</td>
<td>-1.535E-05***</td>
<td>-1.138E-05***</td>
<td>-2.092E-05***</td>
<td>-1.190E-05*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network measures</th>
<th>Nonspatial</th>
<th>Spatial</th>
<th>Nonspatial</th>
<th>Spatial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum shapefactor (P2A)</td>
<td>2.106E-04</td>
<td>3.824E-05</td>
<td>8.388E-04***</td>
<td>5.653E-04**</td>
</tr>
<tr>
<td>Maximum polygon area</td>
<td>2.960E-09***</td>
<td>1.918E-09***</td>
<td>-1.491E-09</td>
<td>-1.549E-09</td>
</tr>
<tr>
<td>Max. betweenness centrality</td>
<td>-5.486E-11</td>
<td>8.994E-12</td>
<td>-2.576E-10**</td>
<td>-1.511E-10*</td>
</tr>
<tr>
<td>Local road length</td>
<td>-3.450E-06**</td>
<td>-2.118E-06*</td>
<td>3.053E-05***</td>
<td>1.713E-05***</td>
</tr>
<tr>
<td>Local road length w. bike lane</td>
<td>6.973E-05***</td>
<td>-5.171E-04***</td>
<td>-3.335E-04***</td>
<td></td>
</tr>
<tr>
<td>Collector rd. len. w. bike lane</td>
<td>-1.584E-04***</td>
<td>-1.065E-04***</td>
<td>1.643E-04***</td>
<td>8.199E-05*</td>
</tr>
<tr>
<td>Coll. rd. len. w. paved shoulder</td>
<td>-1.292E-05</td>
<td>-3.713E-05</td>
<td>6.489E-04***</td>
<td>4.720E-04***</td>
</tr>
<tr>
<td>Coll. rd. len. w. sharrows</td>
<td>-3.147E-04***</td>
<td>-2.428E-04***</td>
<td>8.353E-05</td>
<td>5.115E-05</td>
</tr>
<tr>
<td>Arterial road length</td>
<td>-5.728E-05***</td>
<td>-4.502E-05***</td>
<td>-7.244E-05***</td>
<td>-4.577E-05***</td>
</tr>
<tr>
<td>Arterial rd. len. w. sharrows</td>
<td>-6.642E-04***</td>
<td>-4.975E-04***</td>
<td>3.943E-04</td>
<td>7.907E-04**</td>
</tr>
<tr>
<td>Walk/bike-only trail length</td>
<td>3.687E-05***</td>
<td>2.690E-05***</td>
<td>-3.689E-05*</td>
<td>-8.814E-06</td>
</tr>
<tr>
<td>Spatial filter</td>
<td>-0.841***</td>
<td>-0.316**</td>
<td>0.795***</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-0.235***</td>
<td>8.820E-02*</td>
<td>-0.316**</td>
<td>-0.219*</td>
</tr>
</tbody>
</table>

| Number of observations      | 1544       | 1544    | 1544       | 1544    |
| Residuals Moran’s I         | 0.2254     | -0.0171 | 0.1898     | -0.0030 |
| Residuals Moran’s I (p-value) | < 1E-15   | 0.878   | < 1E-15   | 0.566   |
| R²                          | 0.266      | 0.498   | 0.273      | 0.467   |
| Adjusted R²                 | 0.258      | 0.492   | 0.265      | 0.461   |

Note: *** p<0.001, ** p<0.01, * p<0.05

Regarding network measures, M4 shows that a higher maximum shapefactor is positively associated with an increased relative level of weekend trips. In earlier studies a higher shapefactor
(P2A) of network polygons along a travelled route was found to lead to an overestimation of travel
time compared to measured travel time for commute trips (29). This overestimation might be less
pronounced for non-commute trips due to fewer time constraints, leading to the increase in
weekend ridership in block groups with high P2A values. M4 identifies higher maximum
betweenness centrality values to be associated with a relative increase of weekday BKT. Edges
with high betweenness centrality are located along shortest-path corridors in the network
connecting different components of the network. It is plausible that census block groups with such
corridors are primarily used by bike commuters (and during weekdays), which leads to the negative
coefficient of this variable in model 4. Other regression coefficients of M4 show an increase of
weekend BKT with local road supply in census block groups, whereas an increase in arterial road
length supply in a census block group is associated with an increase in weekday BKT. This
suggests that weekend and weekday cyclists have different preferences for roads types. Analysis
of on-road bicycle facilities does not reveal a clear preference pattern of facility types for either
weekday or weekend travel. Furthermore, supply of walk/bike segments in census block groups is
not a significant predictor of difference in cycling volume between weekend and weekday,
suggesting that off-road paths play an equally important role for weekend and weekday cycling
activities.

SUMMARY AND FUTURE WORK
This study used Strava tracking data to model the effect of sociodemographic variables, place
specific characteristics, and network related measures to model the expected cycling volume for
non-commuter and commuter trips in the greater Urbanized Area of Miami-Dade County. This
research complements existing research on Strava usage patterns by introducing additional
topological and morphological network measures previously underexplored in the context of GPS
based bicycle tracking data analysis. These measures include shapefactor, size of road network
polygons, and betweenness centrality. It is well known that Strava travel data, like ridership data
from other fitness apps, are heavily skewed towards male and younger riders (1, 15). However,
fitness apps have the advantage of an extensive coverage which allows modelers to predict
categories of ridership and reveal spatial variation of ridership. This abundance of data has been
used in this study to model expected differences in relative cycling volume between non-commuter
and commuter cycling volume, as well as between weekend and weekday cycling volume. Most
of the coefficients that were found in the different linear regression models had the expected
arithmetic sign. For example, analyzing the difference between non-commute and commute
cycling volume shows that census tracts with jobs tend to attract commuter cycling volume more
than non-commuter cycling volume, whereas walk and bike trails are associated with a relative
increase of non-commuter cyclist volume. Selected on-road cycling facilities for slower-traffic
roads (local, arterial) are associated with an increase in both non-commuter and commuter BKT,
whereas arterial roads with sharrows see a decrease in non-commuter cycling volume.

Residuals of all nonspatial linear models exhibited significant autocorrelation which can
cause biases of estimated coefficients and their standard errors, risking erroneous conclusions (30).
In this study spatial autocorrelation was explicitly modeled through eigenvector spatial filtering to
obtain a well-defined regression model that assures model assumptions. For each linear model two
versions were estimated, that is, one with and one without spatial filter. Comparison of regression
coefficients demonstrated the importance of addressing autocorrelation in econometric models to
prevent parameter bias.
The effects of network and other variables identified in this study are based on aggregated feature counts per census block group, e.g. total length of local roads with bike lanes. Similarly, predicted quantities were aggregated to a BKT measure (and not computed at the segment level). In order to use segment level measurements (e.g. availability of a bicycle lane, trip count) and combine it with parameters that vary at other levels, e.g. block level, and block group level, multilevel analysis needs to be applied (31). A goal for future work is therefore to refine the presented model to a multilevel model approach with segment counts at the lowest level, which will produce a larger sample size and can be expected to produce more refined analysis results. Future work will also aim to analyze if and how various factors effect ridership differently throughout the day, e.g. during morning, noon and evening hours.

REFERENCES


