COMPARING THE SUITABILITY OF STRAVA AND ENDOMONDO GPS TRACKING DATA FOR BICYCLE TRAVEL PATTERN ANALYSIS

by

Dariia Strelnikova

BACHELOR THESIS 2

in fulfillment of the requirements for the degree of Bachelor of Science

Carinthia University of Applied Sciences
Geoinformation BSc Program

Supervisors

FH-Prof. Dr. Gernot Paulus, MSc. MAS, Carinthia University Of Applied Sciences

Hartwig Hochmair, Ph.D., University of Florida

Villach, February 2017

DECLARATION

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

Davie, FL, 28.05.2017  Dariia Strelnikova

ACKNOWLEDGEMENTS

I would like to thank my little son who managed to play on his own during the long hours that were invested in the creation of this work. I would like to thank my supervisors Dr. Hochmair and Dr. Paulus for their wisdom, support and patience. I would like to thank G-d for everything including the benefits of the modern-day world that made this work possible.

ABSTRACT

Over the past few years crowd-sourced data collected through GPS based bicycle tracking apps on mobile devices have become a widely used data source for the analysis of spatio-temporal travel patterns of cyclists around the world. These large-volume GPS data collections supplement travel information obtained from traditional travel household surveys and data from permanent count stations. Due to their high spatial and temporal resolution particularly in urban areas, GPS tracking data provide sufficient coverage to better understand spatio-temporal bicycle travel patterns. This research project compared coverage and segment based activity counts between two prominent bicycle tracking apps, namely Endomondo and Strava. The first goal was to determine the pros and cons of using the data from these two platforms for travel analysis, based on coverage, spatial and temporal resolution, and other aspects of spatial data quality. The second goal was to compare spatio-temporal usage patterns between the two platforms. Some of the research findings are that Endomondo is less used in South Florida than Strava for the analyzed time frame of three months, and hence provides less detail about travel patterns in South Florida. However, Endomondo tracks are observed on many more small roads and off-road tracks (e.g. in parks) than Strava, complementing Strava data in a meaningful way. There were also some differences in the temporal distribution of trips over the day between both platforms, indicating that the platforms are to some extent used for different trip purposes. This research is unique in the sense that it is the first work that compares Strava and Endomondo data at this level of detail.

ABBREVIATIONS

API	Application Programming Interface
CUAS	Carinthia University of Applied Sciences
ER	Entity-Relationship <diagram>
GI	Geographic Information
GIS	Geoinformation System
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GUI	Graphical User Interface
IDW	Inverse Distance Weighting
OSM	OpenStreetMap
RBF	Radial Basis Function
UAV	Unmanned Flying Vehicle
UF	The University of Florida, USA
URL	Universal Resource Locator
UTC	Universal Coordinate Time
VGI	Volunteer Geographic Information
XHR	XML HTTP Requests

Contents

1. INTRODUCTION

Negative environmental effects assotiated with traffic in urban areas arouse interest of authorities towards cycling. Funding for bicycle infrastructure is increasing and urban areas experience gains in cycling activity (Strava, 2014a). An effective investment in the cycling infrastructure development as well as impactful promotion of cycling is dependent on the understanding of bicycle travel patterns. Traditionally travel information is obtained from travel household surveys and permanent count stations. In the recent years crowdsourced data collected by means of mobile bicycle tracking applications is viewed as a promising data source for the analysis of spatio-temporal bike travel patterns (Griffin & Jiao, 2015; Haworth, 2016; Sileryte, Nourian, & Spek, 2016; Watkins, K., Ammanamanchi, R., LaMondia, J., and Dantec, C. A. L., 2016).

Tracking apps rely on the Global Navigation Satellite Systems (GNSSs). The Global Positioning System (GPS) is a GNSS most widely used. For many years GPS was the only GNSS available to civilian users as a component of their mobile devices. Another GNSS, GLONASS, became supported by mobile civilian devices in 2011. In 2016 that more then 150 smartphone models supported both GPS and GLONASS (Beebom Media, 2016). GLONASS did not substitute GPS but rather added to its coverage and accuracy. As historically GPS was the only wide spread GNSS and it is still a leading technology, the terms GPS and GNSS are often used interchangeably. The term *GPS data* used in this thesis includes data collected by means of all available GNSS.

Large volumes of GPS data can supplement traditional bike travel information. This data "holds the promise to illuminate social processes that were previously undersampled or poorly understood" (Romanillos, Zaltz Austwick, Ettema, & Kruijf, 2016, p. 114). As bicycle tracking GPS data often has a high spatial and temporal resolution, especially in urban areas, it has a potential of providing a better understanding of spatio-temporal bicycle travel patterns. It can reveal which routes are popular and which are avoided, what are the intersection waiting times and the peak travel times. It can also provide some of the characteristics of the cycling community in the area, such its distribution according to age or gender. Subsequently this data can be used to facilitate popularization of cycling and cycling infrastructure improvement.

1.1. Motivation

The number of mobile applications providing a possibility to track cycling activities has substantially grown in the last few years. The popular apps include Strava (released in 2009), Endomondo (2008), Garmin Connect (2008), MapMyRide (2009), CycleMap (2012), CycleMaps (2013), Runtastic Road Bike Tracker (2013) and Runtastic Mountain Bike Tracker (2013), Cycledroid (2016), BikeComputer

(2012), Urban Biker (2014), and Ride With GPS (2007). Some of the cities developed their own bike tracking apps, like CycleTracks (San Francisco, USA) or Radlkarte Salzburg (Salzburg, Austria).

Most of the popular bike tracking apps are available both for iOS (Apple Inc.) and Android (Google Inc.). Apple Store does not provide installation statistics. Google Play Store, however, displays the following installation statistics for some of the popular bike tracking apps (Table 1):

Mobile App	Installs	Rating
Strava	10,000,000 - 50,000,000	4,6
Endomondo	10,000,000 - 50,000,000	4,5
Garmin Connect	5,000,000 - 10,000,000	3,8
Map My Ride	1,000,000 - 5,000,000	4,4
Runtastic Road Bike Tracker	1,000,000 - 5,000,000	4,5
CycleDroid	500,000 - 1,000,000	4,4
Urban Biker	500,000 - 1,000,000	4,3
Bike Computer & Bike Computer Pro	100,000 - 500,000	4,5
Ride with GPS	100,000 - 500,000	4,1

Table 1. Google Store Installation Statistics for the Selected Bicycle Tracking Apps

Mobile bicycle tracking applications are available in different languages and have different features, their producers have different marketing policies. As a result, they vary in popularity, which leads to differences in user counts and location. Therefore, the data generated by users of these apps has different spatio-temporal extent and density. A detailed comparative analysis of such data helps to understand which app can provide more useful bike tracking information for a chosen region.

1.2. Problem Definition and Research Objectives

Romanillos et al. (2016) name Strava and Endomondo among the most widely used bike tracking apps. Endomondo ran past 20 million users in 2013, and more than half of miles logged with this app came from cyclists (Endomondo, 2013). MapMyRide had 9.5 million users in 2012, 3 million of which were cyclists (Velonews, 2012). Strava does not officially release its user counts (Romanillos et al., 2016; Velonews, 2012). But it is reported that "9.6 activities were shared on Strava every second in 2016" (Strava, 2016b).

The study described by this thesis was initiated by the University of Florida (UF) and aimed at the comparison of bike tracking dataset that represent activities of Strava and Endomondo users. The UF has access to some of the Strava Metro data, which is the proprietary cycling count data at the road segment level.

Strava Metro data is available for many regions worldwide. On the Strava website http://labs.strava.com/heatmap/ there is a global heatmap of Strava rides (Figure 1).

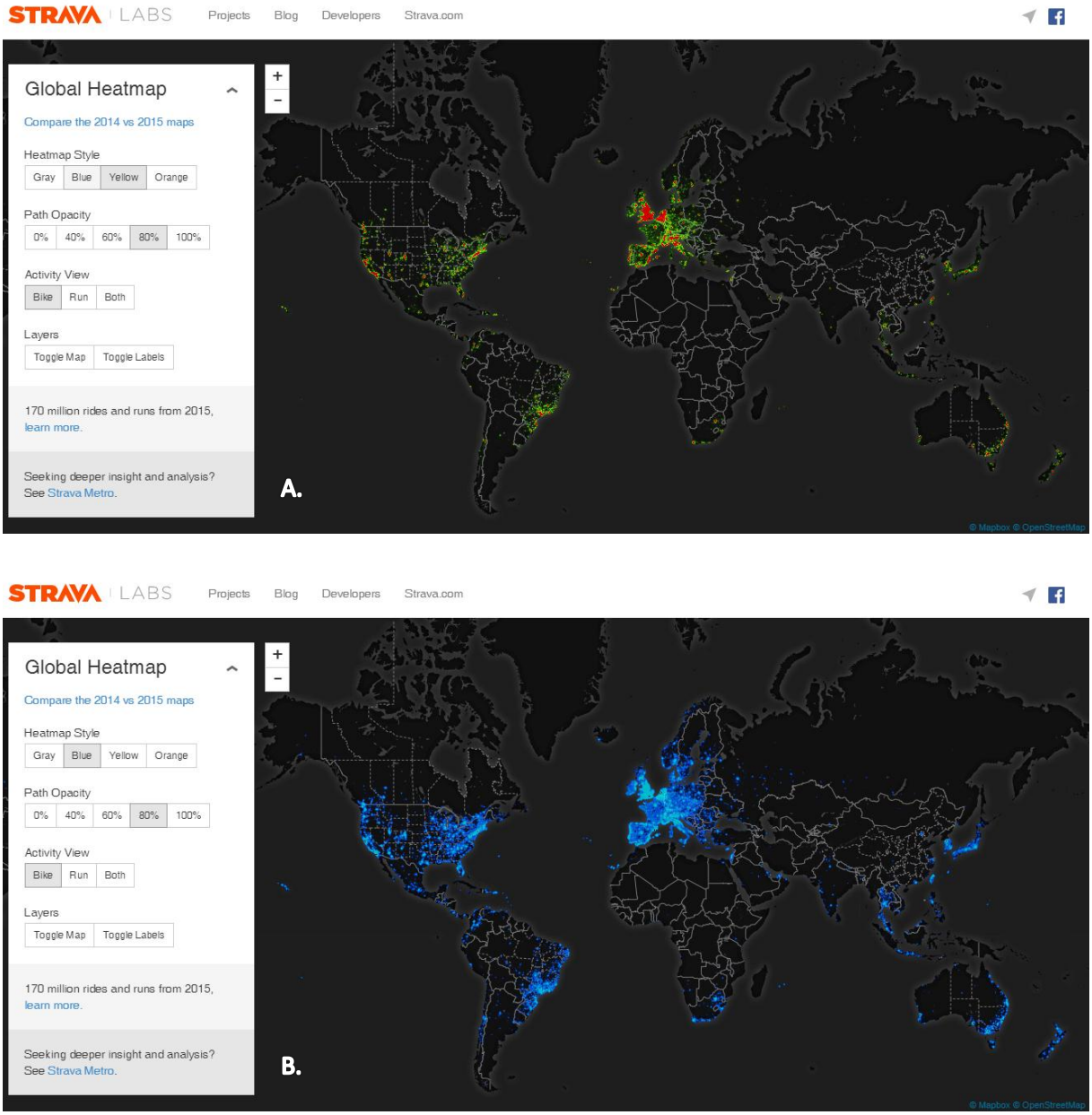


Figure 1. Strava Global Heatmap in Yellow (A) and Blue (B). Retrieved from http://labs.strava.com/heatmap/. Last Access Jun 10, 2017

Strava Metro data is used by over 70 cities and organizations worldwide as a tool for cycling infrastructure improvement (Strava, 2014a). The cost of Strava Metro is 0.8 USD for a single Strava user data in a chosen region in a twelve month period (Strava, 2014b). Strava Metro data is anonymized and aggregated. It does not include personal user data such as age and gender, or distances and durations of individual activities. Haworth (2016) showed that Strava Metro data is a good predictor of the cycle flows in urban area as measured by automatic cycle counters, and Jestico,

Nelson, and Winters (2016) found out that there was a linear association between Strava data and manual cycling counts.

Endomondo is another prominent bike tracking app with the coverage comparable to the one of Strava (Figure 2). Unlike the US based Strava, Endomondo is originally Denmark based, and in 2015 about 80% of its users were located outside the US (Under Armour, 2015). In February 2015 Endomondo was acquired by Under Armour, Inc., an American manufacturer of sport apparel and footwear.

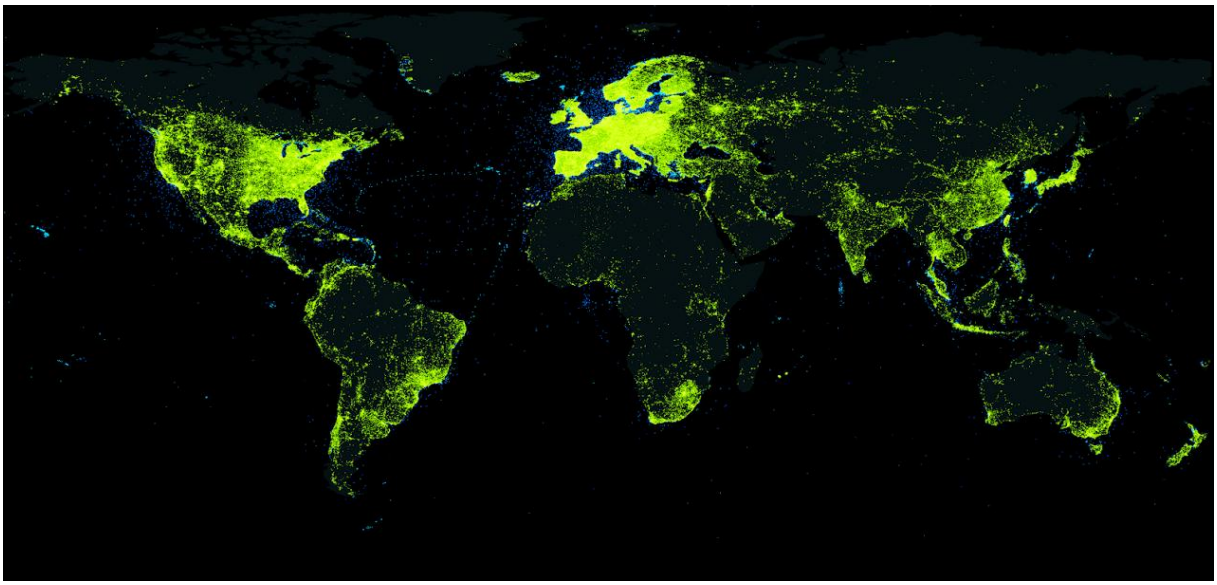


Figure 2. Endomondo Global Heatmap.
Retrieved from https://blog.endomondo.com/endomondo-reaches-20000000-users-the-world-is-going-green/.
Last Access Jun 10, 2017

Endomondo does not offer aggregated and anonymized cycling count data. The data submitted by Endomondo users is mostly public, unless the user decides otherwise. At the same time, this data is not easy to retrieve: "nor the data is available in a single click; neither application providers supply an interface for ready-made free access due to likely privacy issues" (Sileryte et al., 2016, p. 180).

In order to compare Strava data with Endomondo data, the later has to be extracted first. Extraction has to be followed by filtering and transformation into a format comparable with Strava Metro data, and the data analysis itself. All these tasks are time intensive and limit the spatio-temporal extent of study depending on its time frame. Our study had the following objectives:

1. To develop procedures of Endomondo data retrieval and to extract data comparable to the available Strava Metro data in its spatio-temporal extent.
2. To analyze the content of the retrieved Endomondo data in terms of its attributes, density and accuracy.

3. To perform the map-matching of raw Endomondo GPS point data to the Here Navstreets road network used in the Strava Metro dataset and to determine segments with biased counts in both datasets.
4. To perform the comparison of Endomondo and Strava Datasets at the segment and areal level, including the cycling temporal patterns and the usage intensity of different road segments according to their location and classification.
5. To carry out analysis of Endomondo data based on attributes not available in Strava Metro dataset.

The study was conducted from February to May 2017. It initially intended to address additional issues such as the comparison of the Endomondo data coverage between Florida and the selected Austrian cities. Another issue of interest was the evaluation of effect of newly added on-road bicycle facilities, such as bicycle lanes, on cyclist traffic. The study revealed that, as explained further in the section 0, the extraction of Endomondo data in volumes sufficient for such an analysis requires considerably more time than this study had at its disposal. For these reasons we focused on issues that could be adequately addressed within the given four-month time frame.

1.3. Expected Results

It was expected to see differences in the count distribution across different road types and across different sub-areas of cities between Strava and Endomondo as the later one is created to be "a free personal trainer in your pocket" (Endomondo, 2015, p. 1), whereas Strava is also designed, according to their developers, to capture commuter trips (Whitfield, Ussery, Riordan, & Wendel, 2016). It was also expected that in Florida cities Strava has a better coverage, i.e. more roads with any count value larger than 0, and higher overall cycling activity count numbers than Endomondo, since Strava was US based from its first release. It was, however, likely that raw Endomondo data covered off-road trails absent in Strava Metro, which is restricted to Here Navstreets road geometries.

The products of the study had to include a developed procedure and software for Endomondo data extraction, a database (DB) with the retrieved Endomondo data, and charts and maps representing analysis results.

It was expected that technical solutions developed in order to retrieve and compare bike riding data for the selected regions would provide the possibility to perform extraction and similar analyses of other Strava and Endomondo data, such as data for other regions or time intervals, or data representing running activities.

1.4. Thesis Structure

This thesis is structured as follows. Chapter 2 *Theoretical Background* provides information on the nature of bike tracking volunteer geographic information (VGI) and its application (section 2.1.) and explains the concept of map-matching (section 2.2.). Chapter 3 *Methodology and Data* is divided in four sections, which clarify the methodology of the study (section 3.1.), demonstrate its conceptual model (section 3.2.), give information about the chosen study area (section 3.3.), and provide overview of the data sources used in the study (section 3.4.), namely Strava and Endomondo bicycle tracking data (subsection 3.4.1.) and Here Navstreeets and OpenStreetMap (OSM) road network data (subsection 3.4.2.). Chapter 4 *Data Extraction and Preparation* shows how data was retrieved and prepared for analysis. Section 4.1. includes subsections on Endomondo data extraction (4.1.1.) and storage (4.1.2.). Section 4.2. describes the preparation of road network suitable for map-matching, and section 4.3. gives information on map-mathing itself. Section 4.4. explains the procedure of calculation of segment based statistics for Endomondo data. Section 4.5. describes the ways of calculation of the areal statistics.

Furher, Chapter 5 *Results* reveals the study outcomes. Section 5.1. is devoted to comparison between Strava and Endomondo datasets, whereas section 5.2. is based on additional analysis of Endomondo data based on attributes absent in Strava Metro data. Chapter 6 *Discussion* summarizes the identified differences in travel patterns (section 6.1.) and, consequently, the differences in application of the two data sources (section 6.2.). Chapter 7 discusses the data quality and the limitations of study results. Chapter 8 *Conclusions and Future Work* discuses the conclusions that can be drawn from the study outcomes and outlines the possible directions of future studies.

2. THEORETICAL BACKGROUND

2.1. Nature and Application of Bicycle Tracking Volunteer Geographic Information

In the recent years many studies explored the nature of GPS data generated by fitness apps tracking cycling activities. This data belongs to Volunteer Geographic Information (VGI), which is closely related to the concept crowdsourcing (Benkler & Nissenbaum, 2006). The term VGI was introduced by Goodchild (2007):

> "a remarkable phenomenon that has become evident in recent months: the widespread engagement of large numbers of private citizens [..] in the creation of geographic

> information, a function that for centuries has been reserved to official agencies. They are largely untrained and their actions are almost always voluntary, and the results may or may not be accurate. But collectively, they represent a dramatic innovation that will certainly have profound impacts on geographic information systems (GIS) and more generally on the discipline of geography and its relationship to the general public. I term this *volunteered geographic information* (VGI), a special case of the more general Web phenomenon of *user generated content*"

The nature of VGI does not presuppose an existance of stict quality control procedures. There exist research on development of VGI quality assessment procedures and methodology (Criscuolo et al., 2016) as well as on quality management (Bucher, Falquet, & Metral, 2016). Despite the fact that VGI does not guarantee the highest quality, it is "is a revolutionary source of information for increasing spatial and behavioural knowledge on different topics or phenomena in contemporary and everyday life" (Capineri, 2016, p. 28).

Mobile apps that are very popular in the recent years contribute to the exponential growth of GPS derived VGI. In a work devoted to Big Data in connection with cycling, Romanillos et al. (2016) mention with a reference to a global information company Nielsen that 46 million Americans use health or fitness apps. The number of studies exploring the potential of GPS tracking data from mobile fitness apps to be used for common good grows from year to year.

Cintia, Pappalardo, and Pedreschi (2013) studied Strava data to assess cyclists' training effort and performance. Hood, Sall, and Charlton (2013) estimated a route choice model for San Francisco with the help of CycleTracks app. The same app was used by Hudson, Duthie, Rathod, Larsen, and Meyer (2012) to analyze root choice decisions in Texas. Griffin and Jiao (2015) worked with Strava data to determine locations of health-related cycling activities. Cortes, Bonnaire, Marin, and Sens (2015) analyzed healthcare sensor data collected by means of Endomondo. Sileryte et al. (2016) used Endomondo data to analyze spatial patterns of outdoor physical activities. Jestico et al. (2016) also studied cycling route choices by analyzing Strava data of cyclists in British Columbia. Haworth (2016) examined the representativeness of Strava Metro data and assessed its ability to predict cycle flows in urban areas. Watkins, K., Ammanamanchi, R., LaMondia, J., and Dantec, C. A. L. (2016) compared datasets from Strava and Cycle Atlanta to identify the differences in usage of these two apps. Hochmair, Bardin, and Ahmouda (2017) applied Strava data to estimate the volumes of cycling trips in Miami-Dade county, Florida. In most of the studies the bike tracking VGI is used for the route choice modeling and cycle flows assessment. It also has potential to provide answers to healthcare related questions .

2.2. Map-Matching

Map-matching is a process of matching GPS points to existing map geometries. In most cases the term map-matching is used to describe the process of matching a series of ordered GPS points representing a track to a road network graph. This process is challenging because of possible GPS inaccuracies, especially in cities among tall buildings. Matching points to the nearest road segment is often incorrect, especially if it lies on an intersection or between two parallel streets that are very close to each other, e.g. a separate bike road along the car road or two parallel one-way streets.

Many algorithms were developed in order to provide accurate map-matching, the majority of which rely on probability models. Newson and Krumm (2009) developed a map-matching approach based on the Hidden Markov-Model (HMM). Bierlaire, Chen, and Newman (2013) and Jagadeesh and Srikanthan (2015) worked on probabilistic map-matching for smartphone data. Koller, Widhalm, Dragaschnig, and Graser (2015) optimized the HMM map-matching algorithm and reduced its runtime up to 45%. Schweizer and Rupi (2014) and Loidl (2016) offered simpler map-matching approaches that involve the shortest path calculation between unambiguously identified track segments. A paper of Quddus, Ochieng, and Noland (2007) provides an in-depth review of existing map-matching algorithms.

Real time map-matching is used to predict travel time or to calculate/recalculate a route. A need to minimize the output delay can make map-matching algorithms less accurate. Map-matching of previously captured and stored data in many cases can have higher accurasy because of the fewer temporal restrictions. Still if a big number of points has to be matched to a dense a road network, the processing can be computationally intensive and time consuming.

There is a number of open source map-matching projects, that allow to match GPS trajectories to the OSM road network, such as `graphhopper` (https://github.com/graphhopper/map-matching) or `mapillary` (https://github.com/mapillary/map_matching). Most of them process input data in GPS Exchange (GPX) Format. During our study we did not find any open source projects that would match GPS points to Here Navstreets data, which is expected considering the proprietary nature of this dataset.

3. METHODOLOGY AND DATA

3.1. Methodology

The methodology chosen for this study was based on two major tasks: data preparation and data analysis. Proprietary Strava data shapefiles were readily available for analysis. Endomondo data had to be retrieved. Initially it was planned to follow the steps of Endomondo data extraction described

by Sileryte et al. (2016), who relied on the works of Barsukov (2014a, 2014b). During the first stage of the study we found out that the described procedures could not be used as expected because coordinates of GPS points, in opposite to the precondition stated by Barsukov (2014b) and Sileryte et al. (2016), were not embedded in the HTML (Hyper Text Mark-up Language) source code of workout pages of the Endomondo website (www.endomondo.com/workouts/+workoutID). Thus we had to develop a new workflow for data extraction which is described in detail in the section 4.1. The main steps of data retrieval included requesting all workout IDs for a specified user within a chosen time interval and filtering workouts according to the fitness activity type. In case cycling activities were detected in the server response, we requested the user data and the workout data corresponding to the retrieved workout IDs. We saved the data in a JSON (Java Script Object Notation) format, and later parsed the JSON files to get user and workout data. Then we filtered and saved the parsed data into the PostGIS DB.

The following data filtering excluded workouts with metadata that was highly unrealistic for a cycling activity. Thus we did not further analyze workouts with the mean speed under 2 and over 40 kmh, with the distance under 500 m and with the duration under 2 minutes.

Proceding with map-matching was only possible after preparation of the underlying road network. As Strava Metro is based on Here Navstreets data, we used the Here Navstreet dataset as a foundation of the road network. Strava Metro uses a minimized version of the Here Navstreets dataset, which does not contain attributes necessary for routing, including the functional classes of street segments. The IDs of segments are also replaced with Strava IDs. We had access to two versions of the Here Navstreets dataset, from the years 2015 and 2016. The 2016 data, although having a smaller number of segments, had better coverage than the 2015 data. As initially we have planned to analyze the Strava Metro data for the first quarter 2016, we based the road network used in our study on the 2016 Here Navstreets dataset. To make the map-matching more accurate we supplied this dataset with road segments from OpenStreetMap and with the digitized segments.

The map-matching was performed with the use of a simple algorithm proposed by Loidl (2016) with some modifications. This algorithm is based on calculation of the shortest paths between the track points that can be unambiguously assigned to the underlying road network edges. The algorithm has limitations and produces inaccurate results if shortcuts or roads used by cyclists are not present in the underlying graphs, if GPS signal is highly distorted or if the road network is so dense that only a few points in the route can be assigned to edges unambiguously. The advantages of this approach include its simplicity and relatively high map-matching speed, which was especially important considering existing time limits. We used the pgRouting DB extension for the route calculation.

As a result of map-matching we acquired tracks for each workout, including the direction of travel on each road segment. Further, we calculated the distinct Endomondo workout counts per segment,

with respect to the direction of travel, in order to produce a dataset comparable with Strava Metro data. We calculated distinct Endomondo user counts at the segment level. User counts are also available in the Strava Metro dataset. We used this data in order to minimize the possible biasing effect arising from users repeatedly submitting the same route, as, according to Bergman and Oksanen (2016) reiterated submission of the same routes may skew the data even more than mass cycling events and group journeys.

As user and activity counts in Strava and Endomondo datasets substantially differed from each other, we used standard scores (z-scores) and relative values in percent for the comparative analysis of two datasets. We employed descriptive statistics to compare the data as well as to filter out the outliers. We explored correlation of workout durations and distances and calculated the Pearson's coefficient for this correlation. We presented the analysis results with the help of and thematic maps.

The statistical tests involved in data analysis within our study included the two-way analysis of variance (ANOVA) and the chi-square ($\chi 2$) test. ANOVA is the test aimed at the comparison of several population means in order to prove whether they are equal. The alternative hypothesis is that at least two means are different. The two-way ANOVA, unlike its one-way version, deals with populations classified in two ways, e.g. according to age and gender. ANOVA does not allow to detect which populations differ from each other, and in order to do that we ran run the Scheffe's post hoc test. The Scheffe's test requires larger differences between means than other post hoc tests comparing population means, but it is also flexible as it can be applied for different sample sizes. More information on the two-way ANOVA and the Scheffe's test can be found in the works of Moore and McCabe (2003) and Kerr, Hall, and Kozub (2002). The chi-square test deals with nominal data and compares differences between observed and expected counts. The null hypothesis in this test is the homogeniety of populations meaning that observed and expected category proportions are the same for all of them. Detailed description of the chi-square test can be found in Peck and Devore (2012). The software used for the statistical analysis included Microsoft Excel and IBM SPSS.

The geographic information systems (GIS) used in this study were QGIS and ArcGIS. All programming was done in Java programming language with the help of IntelliJ Idea development environment. The PostgreSQL DB server with the spatial PostGIS extension was deployed locally on a high performance workstation under a 64-bit Windows 10 Enterprise operating system, version 1607. We do not recommend this configuration as it proved to be highly unstable, installing updates and restarting or shutting down the workstation without a notice every 7 to 14 days. This behaviour repeatedly interrupted data retrieval and DB populating, leading to the DB inconsistency which required manual time intensive fixes. It also lead to a necessity to develop detailed logging and retrieve/load data in small portions in order to avoid data loss. The additional tasks that arose as a result of the operating

system behaviour took about 80 hours to fulfill. To speed up the data retrieval process we deployed the data extracting software on 4 additional workstations.

3.2. Conceptual Model

The high-level conceptual model of our study comprises the input datasets, the procedures of extraction and generation of other necessary data, the analysis procedures and the deliverables. It can be presented in the following way (Figure 3):

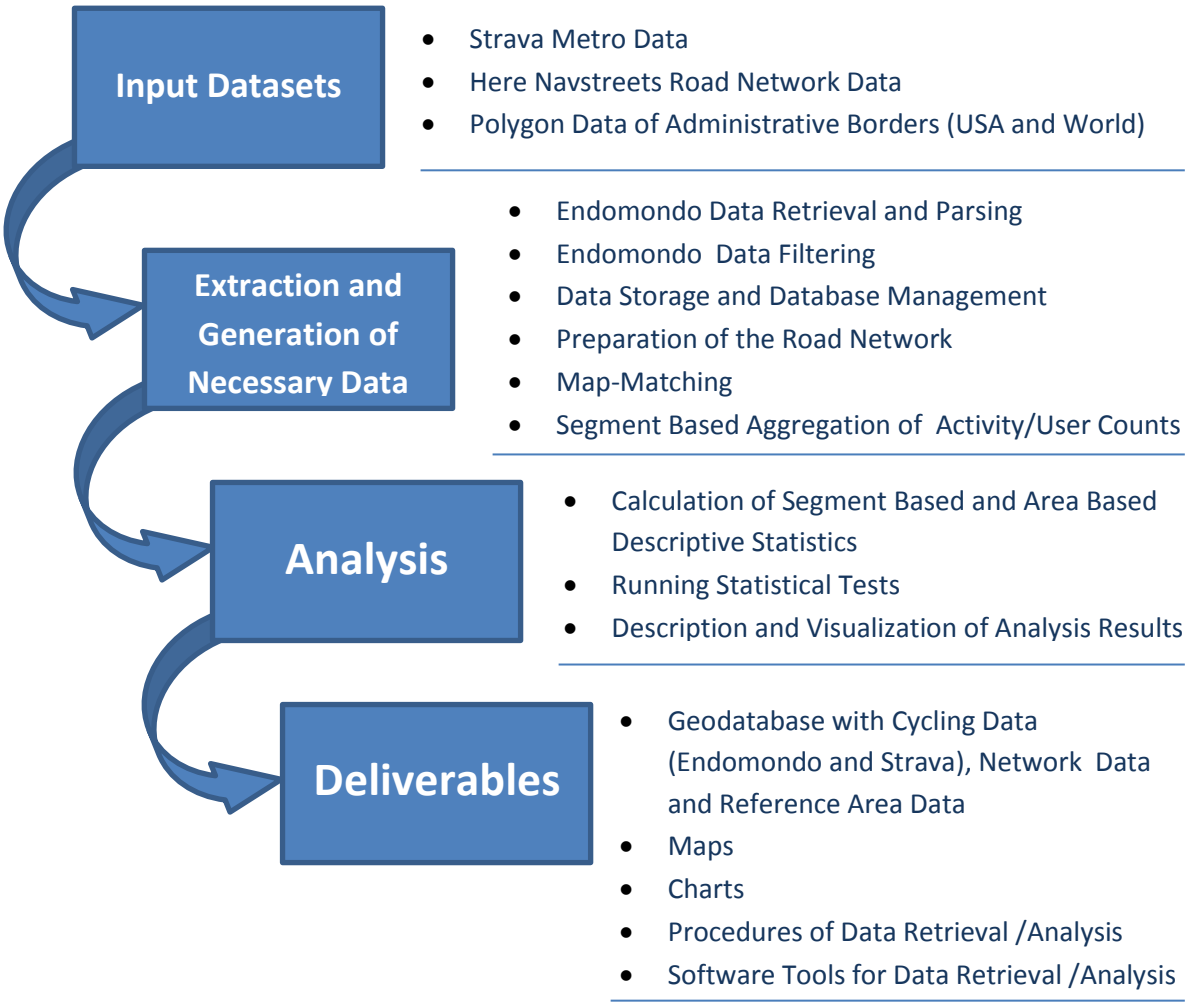


Figure 3. Conceptual Model of the Study

A detailed outlook of the study workflow is given by Figure 4. It shows the interaction between essential components of the study as well as the order of execution. The tasks given in light blue rectangles are fulfilled once. The tasks given in light green rectangles are executed repeatedly. Deliverables are displayed within ellipses, with orange ellipses presenting the deliverables created as intermediate steps in this study or as a result of analysis, and the indigo ellipse containing the study outcomes that facilitate its repeatability. All processes presented by this model are described in detail in further chapters.

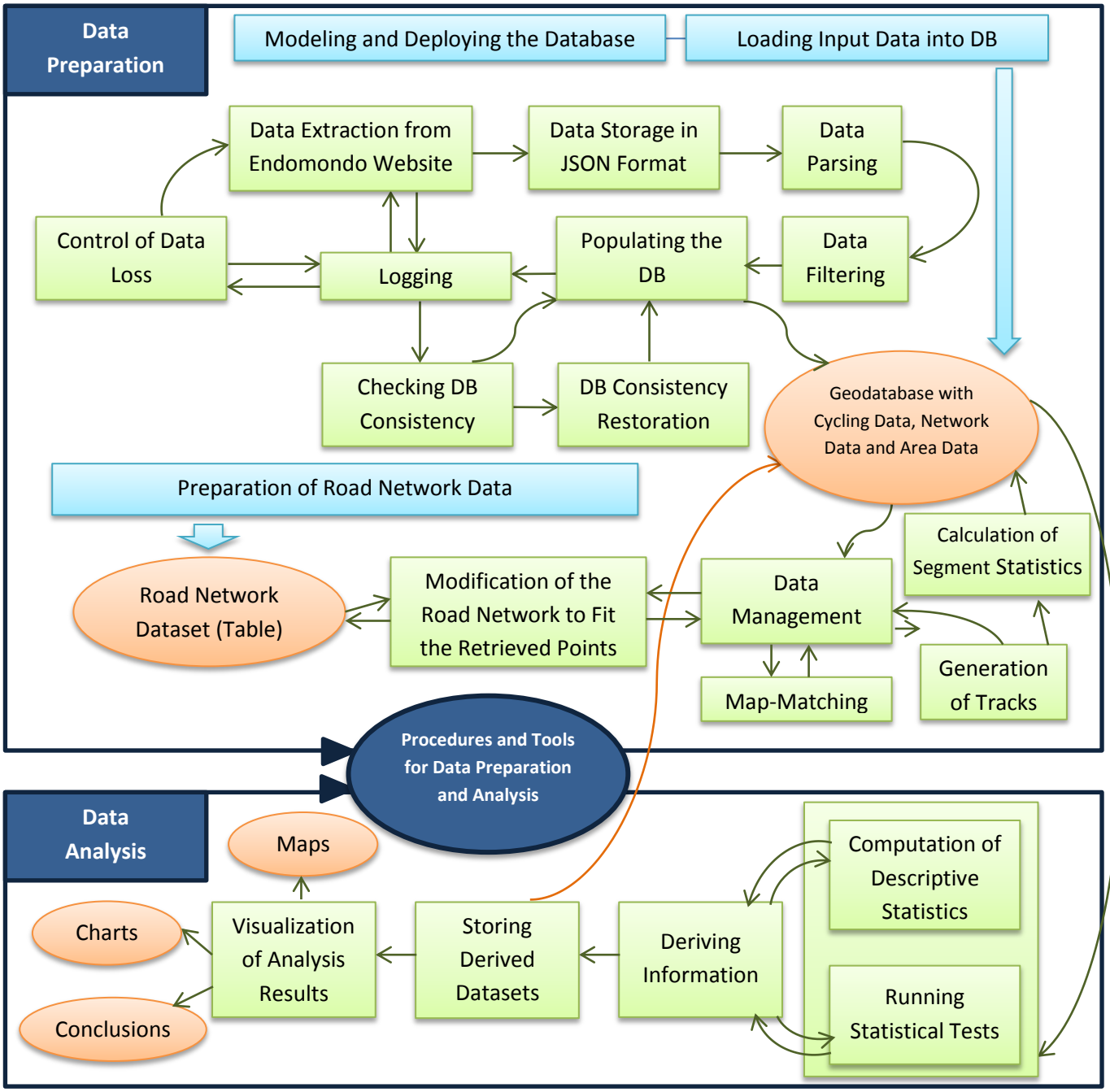


Figure 4. Workflow of the Study

3.3. Study Area

As University of Florida disposed of Strava Metro data for the state of Florida only, our comparative analysis was restricted to this region. Map-matching of all the Florida workouts would require the preparation of the underlying road network for the whole Florida. As cyclists often take shortcuts absent in both Here Navstreets and OSM, these connections had to be manually digitized in order for the chosen map-matching algorithm to function properly. The map-matching itself is a time-intensive process. It requires the intersection of points (413 per workout on average) with road segments and the subsequent route calculation. Further it has to prove whether there are loops or dead ends in the calculated route. Additionally, it has to calculate attribute values for the segments that are not assigned to any points based on the attributes of surrounding segments. The duration of the accomplishment of these tasks for the whole Florida was too large to fit into the time frame of the study. For this reason we limited the segment based data comparison to the area of Miami-Dade county that had the largest population counts, 2,496,457 according to the census data 2010, or 2,700,794 according to 2016 estimate (Bureau of Economic and Research, 2016) and one of the highest population densities (Rayer & Wang, 2014). We carried out analyses that do not require map-matching for four regions: Miami-Dade, Florida, the USA, and the World.

3.4. Overview of Used Data Sources

3.4.1. Strava and Endomondo Bicycle Tracking Data

Figure 5. Strava Logo. Retrieved from www.strava.com on 12.06.2017

Strava is a California based developer and owner of a mobile application and an online platform allowing users to track their fitness activities. Initially targeting running and cycling only, as of 2017 Strava app offers 31 possible activity types, including *Ride, Handcycle* and *E-Bike Ride*. Strava app is one of the most popular apps of this kind and has a strong and active community. Strava's marketing policy successfully contributes to the increase in user and activity counts. Strava treats its users as "athletes" and offers them to enter challenges and compete with other community members. Another attractive thing about Strava community is the so-called Strava Art, also named the GPS Doodles. It is the art created

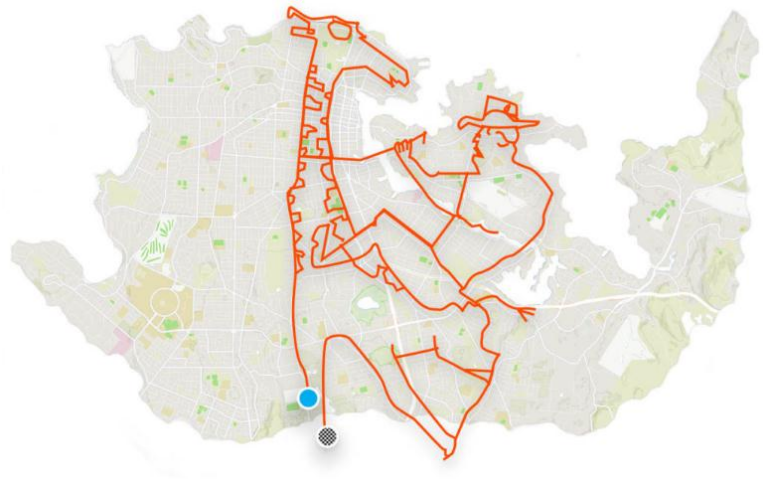


Figure 6. Strava Art by Steven Lund. Published in Strava (2016b), p. 10

by bike, when a cycling trajectory forms a certain artistic shape on the map. Figure 6 gives an example of such a Doodle published in Strava 2016 Statistics (Strava, 2016b). Strava reported that 304 million activities were uploaded in 2016; 26,9% of them were group activities (Strava, 2016b).

Strava offers an Application Programming Interface (API) for developers in several programming languages, including Java, Java Script, Ruby, Python and C/Objective C (Strava, 2016a). The Strava API website (Strava, n.d.) states the following:

1. The API allows access to the user's data only in case such access is authorized by the user
2. There are no privacy restrictions concerning segment data
3. The API usage has short-term (15 minutes) and long-term (1 day) limits, allowing no more than 600 requests for 15 minutes and no more than 30,000 requests per day.

As mentioned before, Strava aggregates and anonymizes the data submitted by its athletes and offers the resulting Strava Metro dataset for a fee. Strava Metro dataset contains road segments with attributes representing activity and athlete counts and mean activity duration for each segment. Along with total counts for the given time interval the dataset contains counts in and against the digitalization direction of the segment. Counts are available for five time-of-day ranges: 12 AM to 4:59 AM, 5 AM to 9:59 AM, 10 AM to 2:59 PM, 3 PM to 7:59 PM, and 8 PM to 11:59 PM. Data also provides distinction between commuter and non-commuter cyclist trips.

Figure 7. Endomondo Logo Before Acquisition by Under Armor. Retrieved from Endomondo Mobile App on 10.06.2017.

Endomondo had been Denmark based until 2015 when it was acquired by Under Armor. Historically the majority of Endomondo users originated from outside the US. One of the studies (Cortes et al., 2015) has shown that 60.84% of all users had workouts on the territory of Europe, 24.28% - on the territory of the USA, and 11,5% - in Asia. Like Strava, Endomondo is a website and a mobile app used by a large community. It is, however, less focused on group activities. As of 2017 it offers submitting activities in 70 categories, including 4 cycling activity types: *Indoor*, *Sport*, *Transport* and *Mountain Biking*.

Endomondo does not offer an official API accessible to developers. The Endomondo support website communicates the following:

> If you want to back up all your workouts from the Endomondo servers you can use the site www.tapiriik.com. It's created by a Collin Fair from Canada, who has done a great job of integrating various fitness trackers and Dropbox. As an exception, we've given him access to our API, so we can do the syncing properly. (Endomondo, n.d., p. 1)

There are several unofficial Endomondo API versions created by independent developers, like the ones found under https://github.com/fabulator/endomondo, https://github.com/adasq/endomondo or https://www.npmjs.com/package/endomondo-unofficial-api. All of them allow access to the workout data (metadata and track) and the user personal data only after the user authorizes this access submitting the password. As our goal was to retrieve the data for hundreds and thousands of users and workouts, we could not use any of these APIs.


```
{
    "expand": "full",
    "id": [redacted]
    "sport": 2,
    "start_time": "2016-05-15T16:12:03.000Z",
    "local_start_time": "2016-05-15T19:12:03.000+03:00",
    "distance": 19.16200065612793,
    "duration": 2794,
    "speed_avg": 24.68976462493219,
    "speed_max": 43.6211,
    "altitude_min": -16.5,
    "altitude_max": 52.7,
    "ascent": 148.7,
    "descent": 105.7,
    "pb_count": 0,
    "calories": 648.37,
    "is_live": false,
    "include_in_stats": true,
    "author": { … }, // 9 items
    "is_peptalk_allowed": false,
    "can_copy": false,
    "weather": {
        "type": 1
    },
    "feed_id": [redacted]
    "laps": { … }, // 2 items
    "records": [ … ], // 1 item
    "hashtags": [],
    "tagged_users": [],
    "pictures": [],
    "points": { … }, // 3 items
    "show_map": 1,
    "show_workout": 0,
    "admin_rejected": false,
    "hydration": 0.85,
    "personal_bests": []
}
```


Figure 8. Example of Endomondo Workout Data in JSON Format

Endomondo workout data contains location data and metadata, including start time (Coordinate Universal Time (UTC)) and local start time, workout distance, duration, average and maximum speed, minimum and maximum altitudes, calories, weather, author data and some other attributes. The important author data includes user id, name and gender. The attribute "show_workout" determines whether this workout is public (value 0) or private (value 1). An example of Endomondo workout data in JSON format is given by Figure 8.

Workout location data can be displayed as points, as laps, as both of the above, or can be absent. There can be maximum 500 points per workout. Distances between points usually vary between 5 and 60 meters. Points, in addition to coordinates, contain some other attributes like timestamp, offset duration in milliseconds, offset distance in kilometers, and sensor data if available (Figure 9).


```
{
    "time": "2016-05-15T16:12:12.000Z",
    "instruction": 0,
    "latitude": 59.5000473,
    "longitude": 24.8580043,
    "distance": 0.005,
    "duration": 9000,
    "sensor_data": {}
},
```


Figure 9. Example of Endomondo Workout Point in JSON Format

Sometimes points do not contain coordinates. An example is given by Figure 10. In most cases points without coordinates have zero values of distance, duration and speed attributes.


```
{
    "time": "2016-01-02T09:24:42.000Z",
    "instruction": 2,
    "distance": 0,
    "duration": 0,
    "sensor_data": {}
},
{
    "time": "2016-01-02T09:24:44.000Z",
    "instruction": 0,
    "distance": 0,
    "speed": 0,
    "duration": 2000,
    "sensor_data": {
        "speed": 0
    }
},
{ … }, // 6 items
```


Figure 10. Example of Endomondo Workout Points With No Coordinates

Lap is a one-kilometer piece of the workout track that is characterized by its start and end coordinates, average pace and time spent on the lap. The distance is always equal to 1 kilometer, except of the last lap that varies in length according to the workout distance. It may or may not contain the exact shape of the track. This is defined by an attribute “small_encoded_polyline”. An example of a workout with laps presenting the exact workout geometry is given by Figure 11.


```
"laps": {
  "metric": [
    {
      "average_pace": 334.109,
      "distance": 1,
      "duration": 334109,
      "begin_latitude": 59.433352,
      "begin_longitude": 24.732012,
      "end_latitude": 59.4404529540814,
      "end_longitude": 24.738598129102613,
      "small_encoded_polyline": "mawiJan}uC[^y@Bm@CqAe@qA[MeAOa@_@LmBVg@Se@CmAPEPA?SkAaBqAm@o@c@Qu@s@{A{A{BkEa@wAe@_Ak@y@]wACe@g@aCi@mAk@u@a@}@GB"
    },
    {
      "average_pace": 260.62,
      "distance": 1,
      "duration": 260620,
      "begin_latitude": 59.4404529540814,
      "begin_longitude": 24.738598129102613,
      "end_latitude": 59.44617870289883,
      "end_longitude": 24.73746857994293,
      "small_encoded_polyline": "ymxiJew~uCYHe@`@k@FOqASF]qCq@sDuByC_@qAy@cDQa@kBiBi@c@e@Ho@Ns@o@s@][j@Ux@c@t@WlBi@hCuAvI]dBKnAiAnG"
    },
    { … }, // 8 items
    { … }, // 8 items
    {
      "average_pace": 413.522282509639,
      "distance": 0.3917999267578125,
      "duration": 162018,
      "begin_latitude": 59.431720658364185,
      "begin_longitude": 24.736835966004712,
      "end_latitude": 59.433346,
      "end_longitude": 24.73182,
      "small_encoded_polyline": "gwviJel~uCI^StAY~A[tAg@xAmA`Ao@`@mAbAGvALrBHxBJxAWZ"
    }
  ],
  "imperial": [
    {
      "average_pace": 323.70704969987014,
      "distance": 1.6093440055847168,
      "duration": 520956,
      "begin_latitude": 59.433352,
      "begin_longitude": 24.732012,
      "end_latitude": 59.44445935131603,
      "end_longitude": 24.74339051638305,
      "small_encoded_polyline": "mawiJan}uC[^y@Bm@CqAe@qA[MeAOa@_@LmBVg@Se@CmAPEPA?
      SkAaBqAm@o@c@Qu@s@{A{A{BkEa@wAe@_Ak@y@]wACe@g@aCi@mAk@u@a@}@a@Le@`@k@FOqASF]qCq@sDuByC_@qAy@cDQa@kBiBi@c@e@Ho@Ns@o@s@]IN"
```

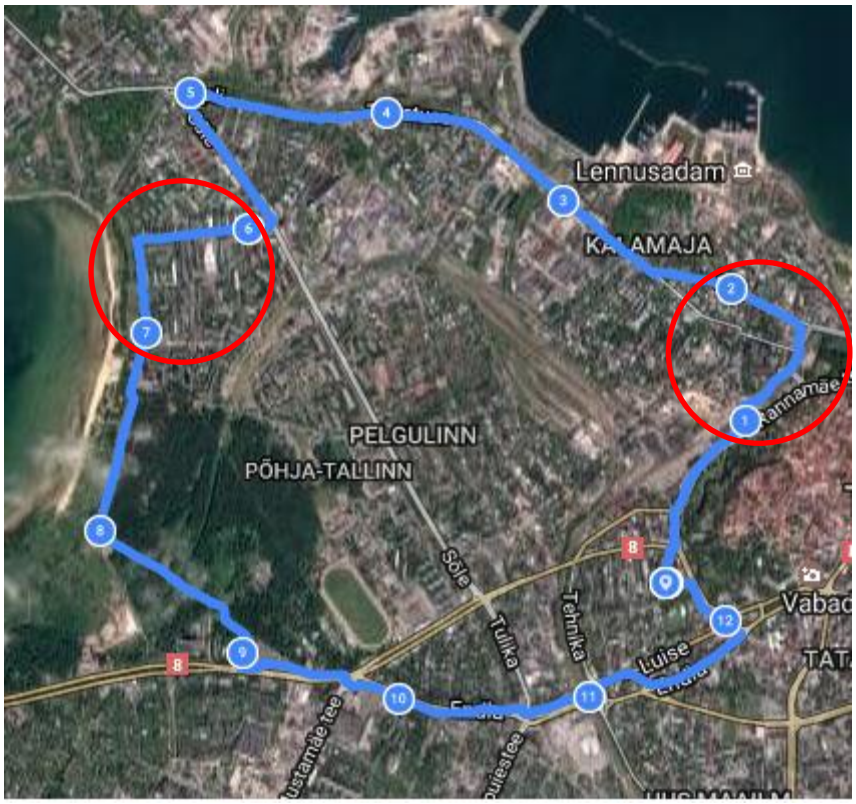

Figure 11. Example of Endomondo Workout Laps and the Exact Geometries of the Laps

The polylines contained in laps are the sets of latitude-longitude pairs enciphered according to a Google algorithm of polyline encoding (Google, n.d.). If laps contain no polylines, which is mostly the case, there is no possibility to restore the original geometry of the workout from laps alone. Figure 12 gives an Example of an Endomondo workout where polyline is not a part of the lap content.

Figure 13 compares an actual track geometry (image A.) with the one restored from laps that contain no polylines (image B.). The encircled numbers (image A.) indicate endpoints of the laps. Red circles point out the most obvious differences between the two tracks.

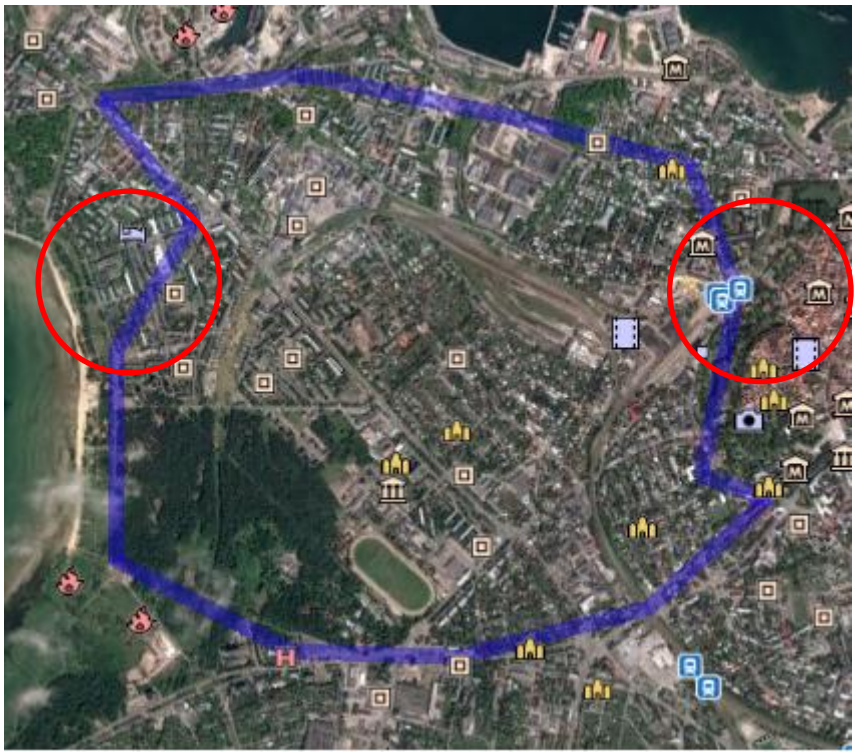
Figure 12. Example of Endomondo Laps without Exact Geometries

Endomondo data model changes over time. A US based study from 2014 (Fang, 2014), for instance, named 52 possible activity types, which is about 15% than Endomondo offers in 2017. One can expect that with the new activity types the new attributes come in place. Another study (Cortes et al., 2015) reported that user profiles contained postal codes, which could not be discovered in data extracted within this study. Thus the characteristics of Endomondo data considered is this section are not comprehensive and reflect only the state of the data at the moment of retrieval.



A

A. Original geometry of the workout restored from the contained points.



B

B. Workout geometry restored from laps with no polylines distorts the actual shape of the workout track

Figure 13. Comparison of an Actual Track with the One Restored From Laps with no Exact Geometries

Among the important attributes of Endomondo user data are user ID and gender, date of birth, date of registration, the number of all workouts since registration, their total duration and length, personal bests and the so-called summary by sport. This summary (Figure 14) provides information on the workout count in each activity category, as well as total duration, distance and calories burned in these workouts.


```
{
    "sport": 1,
    "count": 24,
    "total_distance": 352.8739974498749,
    "total_duration": 102758.56999999999,
    "total_calories": 13657.2
},
```


Figure 14. Example of Summary by Sport

The example of summary given by Figure 14 shows that a user has 24 workouts in the category "Bicycle Transport", their total distance is approximately 353 km and duration is approximately 102,759 seconds which is about 28,5 hours.

Endomondo data contains information absent in Strava Metro dataset and allows to carry out analyses based on the additional attributes, such as workout duration and distance, user age and gender, or sensor data.

3.4.2. Here Navstreets and OpenStreetMap Road Network Data

Here Navstreets is a high-quality proprietary road network dataset. Formerly known as Navteq Navstreets, it is one of the most frequently used road networks in the world. The dataset has 91 attributes that aim at comprehensive routing support. Detailed information about the features and different versions of Here Navstreets data can be found under http://navmart.com/here-navstreets-comparison/. Strava Metro dataset is based on Here Navstreets geometries but includes no attributes from an original Here Navstreets dataset.

OpenStreetMap (OSM), as its name implies, is open data. It is a map of the world, resulting from a collaborative work of people around the globe who contribute to it by adding and maintaining the data. OSM contains much more than just road network. It provides download tools for all its components. The data can be extracted all at once, or for chosen areas. Some APIs allow to select specific tags or features (OpenStreetMap, 2017). The standard format of data is .OSM, which is an Extensible Markup Language (XML) formatted text file. More detail on the OSM data download can be found online under http://wiki.openstreetmap.org/wiki/Downloading_data.

The German company Geofabrik GmbH offers free up-to-date data extracts from OSM by region in the shapefile format (http://download.geofabrik.de/). We used the Geofabrik download service to

retrieve the OSM road network data for Florida. Out of multiple available segment attributes we used only "highway" which is equivalent to the road functional class of Here Navstreets.

4. DATA EXTRACTION AND PREPARATION

We had physical to access 7 ready Strava Metro datasets for Florida:

- One dataset with counts for the first six months of 2016 with no differentiation between weekday and weekend
- Six datasets with counts for the first three months of 2015, with weekday and weekend counts in two separate files for each month from January to June 2015.

We have selected the 2016 counts as a basis for comparison because the information was more recent and the dataset had a bigger coverage. We requested more detailed data for the first six months of 2016 that would include the weekend / weekday distinction.

To be able to compare Endomondo data with the chosen Strava dataset, we had to retrieve the all Endomondo workouts within the same spatio-temporal extension. The data retrieval had to result in all Endomondo cycling data (besides indoor cycling) within Florida border that was generated between January 1st, 2016 and June 30th, 2016.

4.1. Extraction and Storage of Endomondo Data

4.1.1. Data Retrieval

There are several studies and projects that reported Endomondo data extraction (Barsukov, 2014b; Cortes et al., 2015; Sileryte et al., 2016). The data retrieval described in these studies took place from 2014 up to May 2015. The described process of data extraction was based on the presence of GPS data in a JSON format in the source of the publicly available Endomondo website pages.

Figure 15 shows that as of 2017 Endomondo workout webpage sources do not contain embedded JSON data. The body of the document contains a few element most of which are empty in their default state. The content is generated by Java Script functions. The container classified as "layout-content js-main" outlined with a red rectangle in the Figure 15 is filled with workout data dynamically by means of XML HTTP Requests (XHRs). Debugging the webpage uncovered a JSON API that pulled data from certain web addresses.

```
</head>
<body>
  <!-- Google Tag Manager -->
  <noscript><iframe src="//www.googletagmanager.com/ns.html?id=GTM-MK7BQB"
  height="0" width="0" style="display:none;visibility:hidden"></iframe></noscript>
  <script>(function(w,d,s,l,i){w[l]=w[l]||[];w[l].push({'gtm.start':
  new Date().getTime(),event:'gtm.js'});var f=d.getElementsByTagName(s)[0],
  j=d.createElement(s),dl=l!='dataLayer'?'&l='+l:'';j.async=true;j.src=
  '//www.googletagmanager.com/gtm.js?id='+i+dl;f.parentNode.insertBefore(j,f);
  })(window,document,'script','dataLayer','GTM-MK7BQB');</script>
  <!-- End Google Tag Manager -->

  <div ng-include="'tpl://ui/form/validation/validationMessages.en.html'" style="display:none;"></div>
  <div class="layout-header" ng-include="'tpl://view/layout/header/header.en.html'"></div>
  <div class="js-layout-cta" style="display:none;"></div>
  <div class="js-layout-cta-placeholder"></div>
  <div class="layout-header-placeholder"></div>
  <div class="layout-cookies" ng-include="'tpl://view/layout/cookiesWarning/cookiesWarning.en.html'"></div>

  <div class="layout-content js-main" ui-view="" autoscroll="false" style="width:100%;height:100%;">
    <div style="position:fixed;top:6px;left:6px;">Loading...</div>
  </div>
  <div class="layout-footer" ng-include="'tpl://view/layout/footer/footer.en.html'" style="display:none;"></div>

  <script src="/assets/app.a026be7904bb77f4816499e110d7461e.js"></script>

</body>
</html>
```

Figure 15. Example of an Endomondo Workout Page Source

The actual URL (Uniform Resource Locator, also known as a web address) to the workout data was not www.endomondo.com/workouts/+workoutID (URL-1) mentioned in other studies but https://www.endomondo.com/rest/v1/users/userID/workouts/workoutID (URL-2).

To access the data under the URL-2 both a workout ID and the corresponding user ID must be known. A call to the URL-1 gets automatically redirected to a URL containing the user ID: https://www.endomondo.com/users/userID/workouts/workoutID. This is not the case for the URL-2. It is, however, possible to first call the URL-1 and obtain the user ID from the link to which the call is redirected, and then to send the second a request to the URL-2.

All the previous studies that we could find proceeded with data extraction by downloading workout data by looping through subsequent workout IDs. The numbers representing workouts IDs tend to increase with the time. A workout that gets logged later is likely to have an ID number exceeding those logged earlier. A workout ID is just an order number, thus there is no way to determine the workout location or the contributing user ID solely from the workout ID. Also, the relationship between the workout ID and the workout date is not exactly straightforward, as activities may be synchronized with the Endomondo DB after the actual activity date. By the middle of 2016 there existed more than 700,000,000 workouts submitted by Endomondo users. Some of the workouts belonging to the time interval of interest were synchronized with the Endomondo DB later, thus a few of the workouts had IDs past 700,000,000 up to approximately 925,000,000. A preliminary analysis of a manually collected sample consisting of 300 workouts has shown that more than 93% of workouts logged in the first six months of 2016 have IDs between 650,000,000 and 757,000,000. Processing this ID interval would require sending the following number of requests:

- 107,000,000 requests to determine user IDs corresponding to workout IDs
- 107,000,000 requests to check whether a workout is classified as cycling and retrieve the actual workout data
- 1 request per each unique user with a cycling activity. According to Cortes et al. (2015), the mean number of workouts per user in five months in 2014 was 22, and 22,87% of all workouts were cycling activities. These proportions could change in five years, but if we proceed from them the expected number of distinct cyclists would be 107,000,000 x 0,2278 / 22 = 1,112,314.

According to this calculation, the total number of requests would be more than 215,000,000. The server response time allowed to process 8 requests from one thread in a second, so data retrieval by one workstation in one thread would take about 311 days. Multithreading could not be applied as requests sent from in many threads from the same IP were not processed faster than consecutive requests from one thread. As we had 5 workstations available for the project, the best case scenario was getting the data in 62 days if all workstations were employed non-stop. In practice, the number of requests to a resource (the Endomondo DB in our case) is always limited. If Endomondo had the same retrieval limit as Strava (30,000 requests daily), making 215 million calls would take about 20 years. Non-stop requests sent from multiple systems simultaneously are considered to be a Distributed Denial of Service Attack (DDoS attack), and the requesting systems' access to the resource is denied as soon as an attack has been detected. To guarantee that the necessary data is retrieved and analyzed within the project time frame we had to substantially decrease the number of calls to Endomondo DB.

First, we decided to decrease the time interval of interest to the first quarter 2016. The majority of Endomondo users are located in Europe, and winter months are months of low cycling activity in many European countries because of the weather conditions. Thus among workouts logged in January and February the proportion of those located in Europe is likely to be less than among those logged in May or June.

Next, processing user IDs instead of workout IDs had to lead to a substantial decrease in a number of necessary calls. The Endomondo user IDs are assigned in the same way as workout IDs, namely in the order of registration. We have identified that by the end of the first quarter 2016 the maximum user ID was approximately 27,290,000. If we could determine which workouts were logged by each user within the given time interval (January 1st, 2016 to March 31, 2016) and choose only those categorized as cycling, we could request workout data based on already known workout-user pairs and avoid double requests for each workout ID. Also the number of calls for workout data would be reduced to the calls for cycling activities only. Additionally this approach would help to capture the workouts with IDs outside the estimated ID interval.

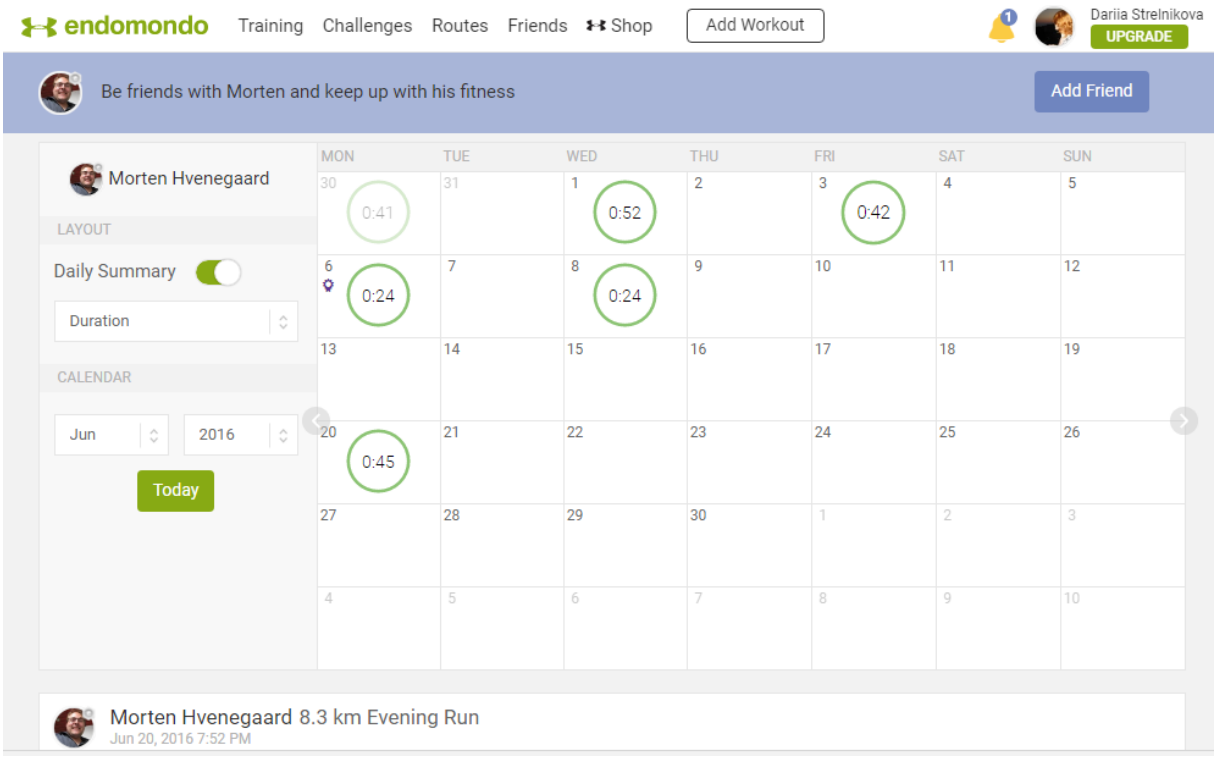


Figure 16. A Typical Workout Webpage on Endomondo Website

Each workout page on the Endomondo website (Figure 16) displays a calendar with activities for a month. Therefore there must exist an XHP requests that loads data for a selected time interval. We have determined that such requests are sent to the URL

https://www.endomondo.com/rest/v1/users/userID/workouts?before=END&after=START

where END and START are timestamps written according to the ISO 8601 standard, e.g.

2016-01-01T00:3A00:3A00.000Z

which get slightly modified when used as a parameter in a HTTP request:

2016-01-01T00%3A00%3A00.000Z

An example of a server response to a request described above is given by Figure 17. The response provides comprehensive workout metadata including activity types. Only activity types with the following values of the attribute “sport” need to be processed further:

1 = Bicycle Transport

2 = Bicycle Sport

3 = Mountain Biking.

```
[
  {
    "expand": "abs",
    "id": 651346884,
    "sport": 1,
    "start_time": "2016-01-02T13:39:52.000Z",
    "local_start_time": "2016-01-02T14:39:52.000+01:00",
    "distance": 9.72599983215332,
    "duration": 3106.25,
    "speed_avg": 11.271983708893988,
    "speed_max": 33.8641,
    "altitude_min": 24.1,
    "altitude_max": 62,
    "ascent": 44.1,
    "descent": 53,
    "pb_count": 0,
    "calories": 369.28,
    "is_live": false,
    "include_in_stats": true
  },
  {
    "expand": "abs",
    "id": 651337420,
    "sport": 1,
    "start_time": "2016-01-02T13:23:58.000Z",
    "local_start_time": "2016-01-02T14:23:58.000+01:00",
    "distance": 3.306999921798706,
    "duration": 849.66,
    "speed_avg": 14.011722004655207,
    "speed_max": 28.4173,
    "altitude_min": 36,
    "altitude_max": 52,
    "ascent": 33.8,
    "descent": 22.2,
    "pb_count": 0,
    "calories": 124.12,
    "is_live": false,
    "include_in_stats": true
  }
]
```


Figure 17. Example of a List of Workouts Logged Within a Defined Time Interval

We developed a simple command line tool that looped through user IDs retrieving workout-user pairs. The application of this tool has shown that there are restrictions concerning the number of requests sent to Endomondo: after each 100 to 120 requests the server returned the status code 429 “Too Many Requests”. It appeared to be possible to continue requesting the database in 12.5 seconds after the code 429 was returned. However, after about 19 hours of workout-user pairs retrieval the Endomondo server returned the status code 403 “The server understood the request but refuses to authorize it”. After that it did not processed any requests sent from the same IP address during the next 72 hours.

After several modifications of the retrieval tool we came to the following approach. Data extraction is done in iterations. An iteration size is adjustable. The size from 1,000 to 2,000 users guarantees minimal data loss in case retrieval process is interrupted. Progress is logged upon completion of every iteration. If iteration is interrupted, the loss of statistic data is minimal, and it is easy to restart the process from the same point or very close to it. All user data is saved in files "userID.json" in the user folder. The workout data is stored as "workoutID.json" in the workout folder. The workflow of the retrieval tool is organized as follows:

1. Request IDs of all workout logged in the first quarter 2016 by a given user
    1.1. If the call is rejected (code 429), wait 12.5 seconds and repeat the call
        1.1.1. If the call is rejected again, store the rejected id to repeat the call at the end of the current iteration
        1.1.2. Else proceed according to 1.4.
    1.2. Else if the server returns the HTTP status code 404 "Not Found" or 500 "Internal Server Error", consider the user invalid
    1.3. Else if the server returns the HTTP status code 403, write the log and terminate the retrieval immediately to avoid long-term blockage
    1.4. Else If there is at least one outdoor cycling activity among the logged workouts:
        1.4.1. Request personal user data: gender, date of birth, date of registration, country of registration, summaries by sport
            1.4.1.1. If the server returns the code 429, 404, 500 or 403, proceed as described in steps 1.1. – 1.3.
            1.4.1.2. Else store the user data
        1.4.2. For each logged workout request workout data for each workout ID retrieved in step 1.
            1.4.2.1. If the server returns the code 429, 404, 500 or 403, proceed as described in steps 1.1. – 1.3.
            1.4.2.2. Else store the workout data
            1.4.2.3. Wait 2 seconds until requesting data for the following workout
2. Assess number of processed IDs:
    2.1. Iteration size is not reached:
        2.1.1. Repeat for the user next in turn.
    2.2. Else:
        2.2.1. Repeat calls to all rejected user IDs and workout IDs
            2.2.1.1. If it returns data, save this data
            2.2.1.2. If it is rejected, increase the count of rejected calls

2.2.2. Log the start and end user IDs processed in the current iteration and its duration, as well as number of made calls, rejected calls, retrieved workout count, retrieved user count, and invalid user count.

3. Proceed with the next iteration (steps 1-2).

The data extraction part of this study resulted in the following numbers:

Parameter	Absolute Value	Relative Value
Processed user IDs	27,288,924	
Number of invalid users	1,963,298	7.19%
Number of sent requests	32,513,380	
Number of rejected requests	623,967	0.70%
Number of retrieved users worldwide	551,052	
Total number of retrieved workouts	4,009,663	
Retrieved data volume	≈200Gb	
Total duration of data retrieval* *(data synchronization and processing of logs are not included)*	6,029h 17m (251 days)	
Number of requests:		
per second:	≈1.5**	
per day:	≈129,535***	

* Sum of durations for all five employed workstations

** ≈1.5 from one workstation, ≈7.5 from all five

*** ≈129,535 from one workstation, ≈647,675 from all five

Table 2. Statistics of Endomondo Data Extraction

4.1.2. Database Management

Simultaneously with data retrieval we parsed the already extracted data. As we expected a high number of points most of which would not be required in the scope of this study, we initially divided point data in multiple tables according to their location. Proceeding from the distribution of users among the countries of registration we have created point tables corresponding to 19 regions according to their location: 17 distinct countries (including the USA without Florida), Florida, and one table for the rest of locations. A tool we have developed for parsing data and populating the database intersected the first workout point with the polygons representing the areas of interest. As a result, the location of the workout (and all of its points) was determined by the region its first point belonged to. In this way we could minimize the number of intersections needed to assign each point to a certain region. It is possible that some of the points were not assigned correctly if the workout

took place on the border of two regions. During the further data analysis, however, we did not run into incorrectly assigned points.

We populated the DB with the parsed data. The Entity-Relationship (ER) diagram of the DB deployed for this study is given by Figure 19. The core of the Endomondo data model is given by Figure 18.

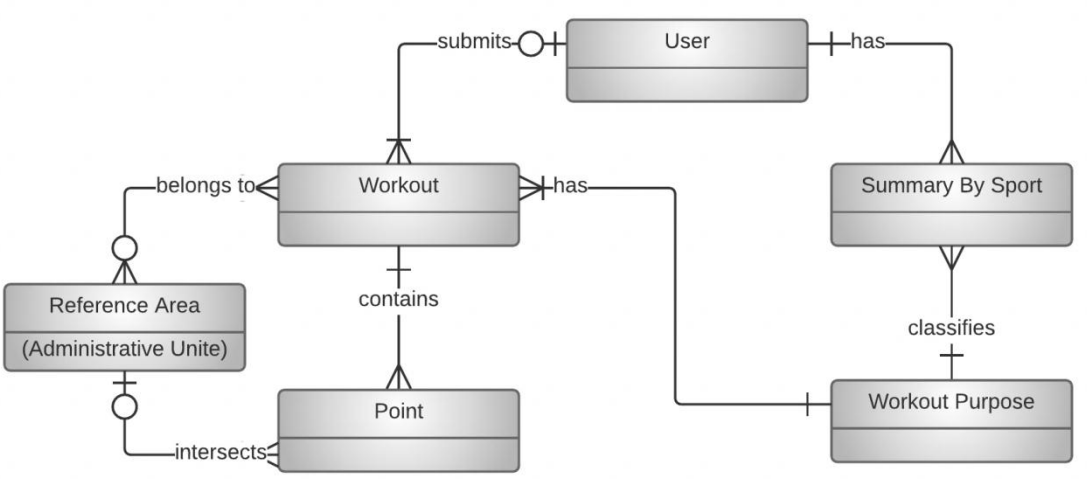


Figure 18. Core of Endomondo Data Model

The data belonging to different data categories has unequal volumes (Table 3, column *Count*). As a result, the time of populating the DB with this data varies greatly depending on the data category. A naive approach of filling in the tables with simple INSERT statements is inefficient but still applicable for user, summary and workout data. However, when it comes to the point data, the duration of populating the DB with all retrieved points by means of the simple INSERT statements would be approximately 4,741 hours, or 198 days.

Category	Count	Number Of Columns	INSERT speed, rows per second	COPY speed, rows per second	INSERT duration, hours	COPY duration, hours
User	551,052	8	79	2,208	2	0
Summary	844,625	5	44	1,227	5	0
Workout	4,009,663	11	97	2,699	11	0
Point	1,354,704,199	8	79	2,208	4,741	170

Table 3. Retrieved Endomondo Data in Numbers by Category

Before the point tables were populated, we did not create indices which decelerate the data loading process. As we had only one physical hard drive for the whole database, using parallel connections to populate the DB could not lead to an increase in speed. The use of transactions of 100 inserts each sped up the insertion to about 115 rows per second, but the estimated total duration of insertion was still out of the study limits. The use of the COPY statement allowed to significantly improve the

processing times. With the speed of approximately 2,208 points per second, populating the DB with all points took approximately 170 hours or a little more than 7 days.

We organized data parsing and populating the DB iteratively within a total of 196 iterations to minimize data loss in case of the workstation restart. Each iteration dealt with data corresponding to a certain interval of user IDs. The parsing tool converted the Endomondo JSON files into text files formatted and named as target DB tables, and these files were then loaded into DB by means of COPY. At the time of parsing the points were distributed among files representing regional tables, e.g. `point_fl`, `point_us`, `point_de` etc. The tool logged progress upon each iteration and compared input data counts with the counts queried from the DB to validate data completeness. At the end we archived all the interim text files as backup data.

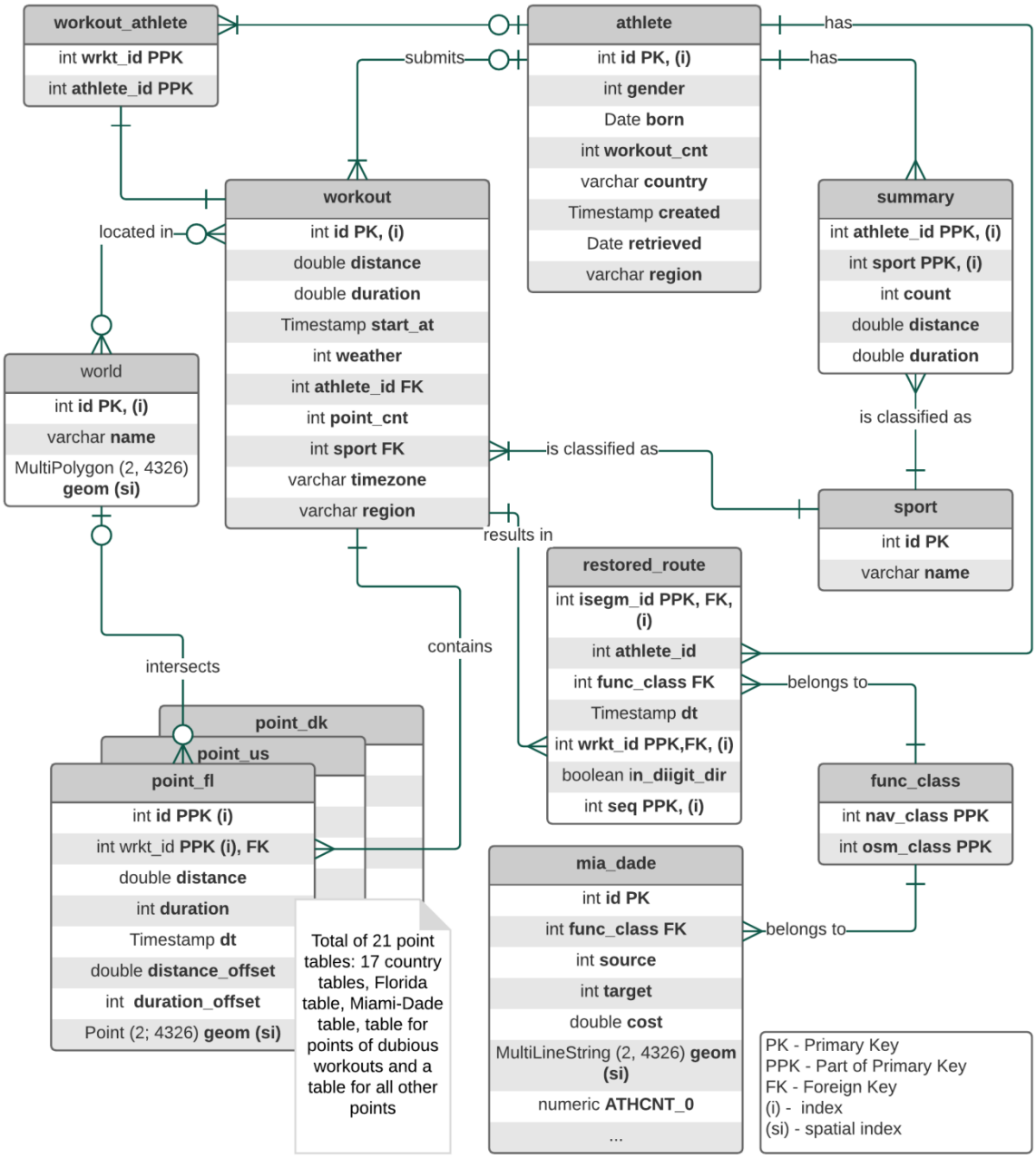


Figure 19. Entity-Relationship Diagram of a Database for Endomondo Data Storage

The database structure presented by Figure 19 contains the database elements involved in data storage. For data analysis we used additional tables and views not mentioned in this ER diagram.

Tables `athlete, workout, point, summary, and sport` model basic Endomondo data. Table `world` contains reference polygons representing administrative regions. Table `workout_athlete` contains workout-user pairs. In some cases user summary can be extracted, but the corresponding user profile is unavailable or invalid. For such users there is no data in the `athlete` table. To be able to match each workout with user id we use an additional table. The foreign key `athlete_id` references the `workout_athlete` table.

The `mia_dade` table contains the road network data for Miami-Dade County. This data combines Here Navstreets and OSM geometries and digitized segment geometries with Strava Metro counts. The network data is discussed in detail in the next section. The table has all the attributes of the Strava Metro dataset that are explained in detail in Appendix 1. These attributes are absent in the ER diagram for the sake of readability. The additional attributes present in the `mia_dade` table but absent in the original road network datasets are `source`, `target` and `cost`, used for the route calculation.

The table `func_class` contains the matched functional road classes of Here Navstreets and OSM as the two data sources have different standards. The table `restored_route` stores road segments corresponding to workout's route computed by map-matching. A redundant attribute `athlete_id` is stored for each segment to speed up further data analysis. The attribute `seq` is the sequential number of the road segment in the route and `in_digit_direction` shows whether the activity along the segment was in digitation direction.

4.2. Road Network Preparation: Integration of Navstreets and OpenStreetMap Data

Strava Metro data contains the Here Navstreets geometries but the IDs of the road segments are replaced with Strava own IDs. Strava Metro contains no information of road functional classes. To obtain them from a corresponding Here Navstreets dataset we calculated a 0.5 m buffer around each segment of Strava Metro. Then we assigned to each Strava Metro segment the class of the Here Navstreets segment that was completely within the buffer. Thus we obtained functional classes for all Strava Metro segments.

Here Navstreets data has a good coverage. At the same time, it is mostly used for car navigation and does not contain some of the streets or shortcuts that can be used only by cyclists. We analyzed location of the retrieved points and found multiple cases where points were located along roads absent in Strava Metro (Figure 20) or along shortcuts (Figure 21).

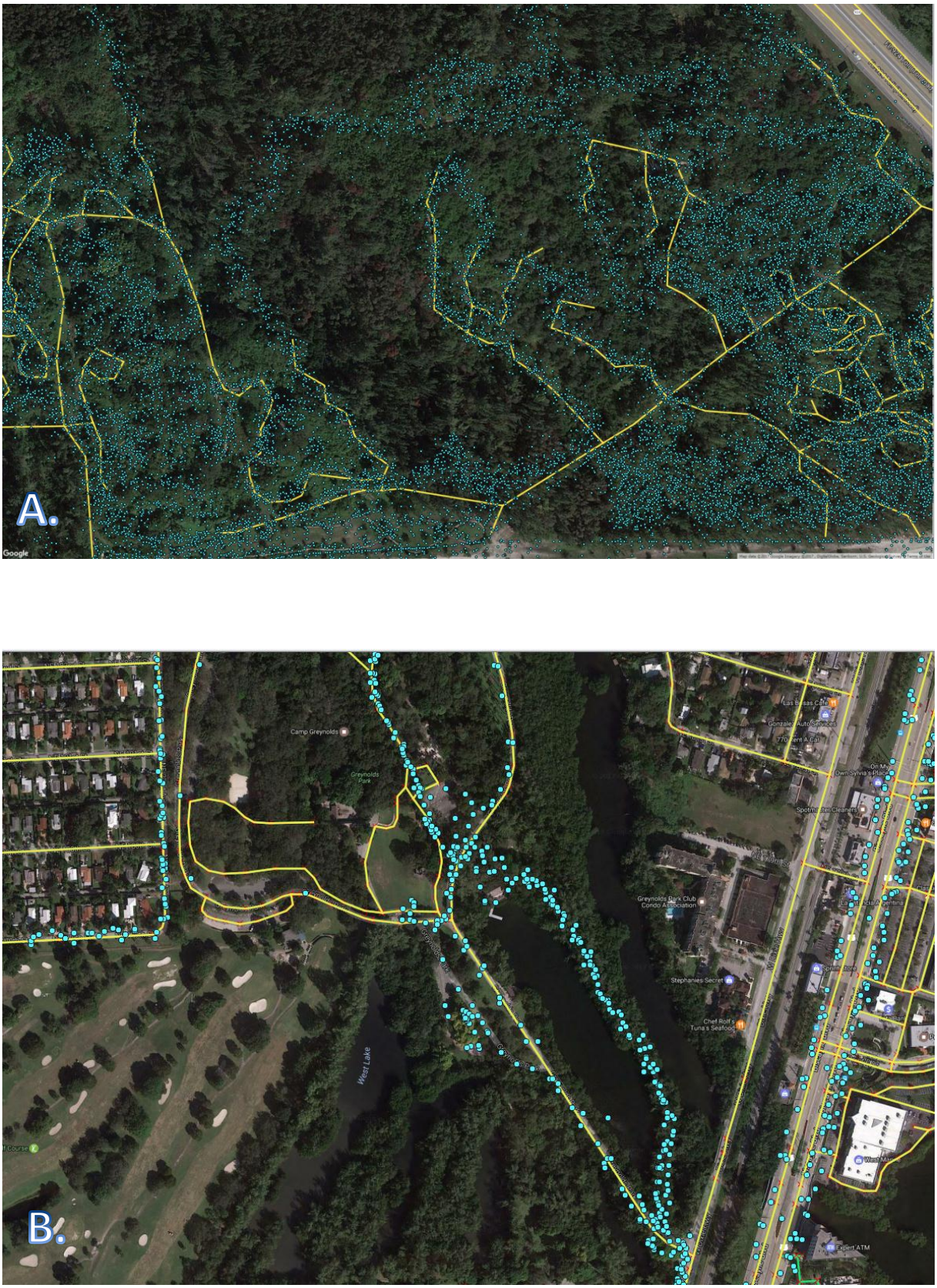


Figure 20. Endomondo Points Outside of Here Navstreets Geometries. Part I.

Figure 20. Endomondo Points Outside of Here Navstreets Geometries. Part II.

Figure 21. Shortcuts Absent in the Here Navstreets Data

Figure 22. Missing Segments Along Existing Roads

We downloaded an OSM dataset from *Geofabrik* and clipped it to the Miami-Dade County area. Then we selected the OSM segments that did not intersect with the Here Navstreets geometries and added them to the mia_dade table with the value of `mapsource` = 'OSM' (all Strava Metro segments have a value of `mapsource` starting with 'NavTeq'). Then with the help of the *Disconnected Islands* QGIS plug-in we determined the groups of OSM segments that were connected to each other but not to the Here Navstreets network. We digitized connections to the OSM 'islands' manually. We also manually digitized the revealed missing shortcuts and to achieve the connectivity

necessary for successful map-matching. Further analysis of the network data has revealed that in rare cases some of street segments were absent in the dataset (Figure 22). We digitized these streets on-screen. After the preparation of the road network was completed, it contained the 186,937 segments (Table 4):

Data source (`mapsource`)	Segment count
NavTeq2016Q1	167,769
OSM	11,863
NavTeq2015	1
Digitized	7,304

Table 4. Network Segment Counts According to the Data Source

There exist an approach to integration of different street networks by buffering one or both datasets, uniting buffers and extracting their centerlines (Sileryte et al., 2016). This approach involves no digitization, but resulting polylines lack the attributes of original road segments. Our approach to the network data integration relied on digitization to a great extent. We have chosen it due to the necessity to retain the original geometry of Strava Metro since athlete and activity counts are tied to these geometries and it is no possible to restore these counts if geometries are changed.

To unify the classification of roads for OSM and Here Navstreets we analyzed both classifications and matched the semantically equal classes. Classification of Here Navstreets roads is given in the reference guide to the dataset. This document comprising over 1,000 pages, is proprietary, confidential and distributed among clients. We have received an older version of this reference manual (NAVTEQ, 2008). The OSM standard of road classification in is available online under http://wiki.openstreetmap.org/. The correspondences determined and applied to classify all road segments according to the Here Navstreets standard are given by Table 5.

Here Navstreets Functional Class	OSM Functional Class	Here Navstreets Functional Class	OSM Functional Class
1	motorway	3	primary_link
2	trunk	3	secondary
2	motorway_link	4	secondary_link
3	trunk_link	5	living_street
4	tertiary	5	track
5	tertiary_link	5	track_grade1
5	footway	5	track_grade4
5	service	5	track_grade5
5	pedestrian	5	residential
5	cycleway	5	unclassified
5	path	5	steps
2	primary		

Table 5. Correspondence between Here Navstreets and OSM Road Functional Classes

Finally, we replaced all *null* values of athlete and activity counts by OSM and digitized road segments with the value *zero* as there were zero Strava cycling activities recorded on these segments. At this point the road network was ready for the map-matching and count comparison.

4.3. Map-Matching: Transformation of GPS Points into Road Segments

A simple approach to map-matching used in this study was offered by Loidl (2016). It consists of the following steps:

1. Create a minimum bounding box of all GPS points
2. Clip the road network accordingly to speed up the processing
3. Create buffers around network segments to reflect the areal nature of roads
4. Create buffers around GPS points to counterbalance possible GPS errors
5. Overlay buffers and select points intersected by only one segment
6. Snap points to this segment
7. Calculate the shortest path using points as stops
8. Match the calculated path to the underlying network

We implemented this algorithm with the help of pgRouting PostGIS extension to map-match 1,385,929 GPS points located in the Miami-Dade County using a buffer of 15 m for segments and 5 m for points. The map-matching results were less accurate than expected: about 70% of the segments per workout. One of the possible reasons could be the high density of road network, including many parallel roads located closely to each other. This resulted in sparse unambiguous GPS points for route calculation. We have made several adjustments to the algorithm, such as:

1. We reduced the network to a buffer of 100 meter around the minimum bounding box to avoid false connections between points that lie close to the border.
2. In case buffer intersection resulted in too few unambiguous points, we repeated it with a network buffer of 7 m to avoid intersection with closely located parallel roads.
3. We employed the known distances between unambiguous points to assess the closeness of the calculated path length to the actual distance between the points. If length of the shortest path calculated between the known points exceeded the actual distance between these points by more than 30%, in most cases this path contained a loop resulting from a non-digitized shortcut (Figure 23). For such cases we broke the track in several parts (Figure 24).
4. At the end of the map-matching we analyzed the presence of dead ends in the resulting segment sequence. We compared the length of the dead end with the distance between the last point captured before the dead end and the first point captured after it. If this distance

was less than twice the length of the dead end segments, we deleted the dead end segments from the resulting track as erroneous.

Figure 23. Error in the Restored Track due to a Not Digitized Shortcut

Figure 24. Restored Track Comprising Two Parts due to a Not Digitized Shortcut

As a result the accuracy of the matched tracks was improved to 90-95% correctly calculated segments. We retrieved segment based tracks of 3,240 workouts that took place in Miami-Dade County from January, 1st to March, 31st 2016. An extract from the resulting map is given by Figure 25.

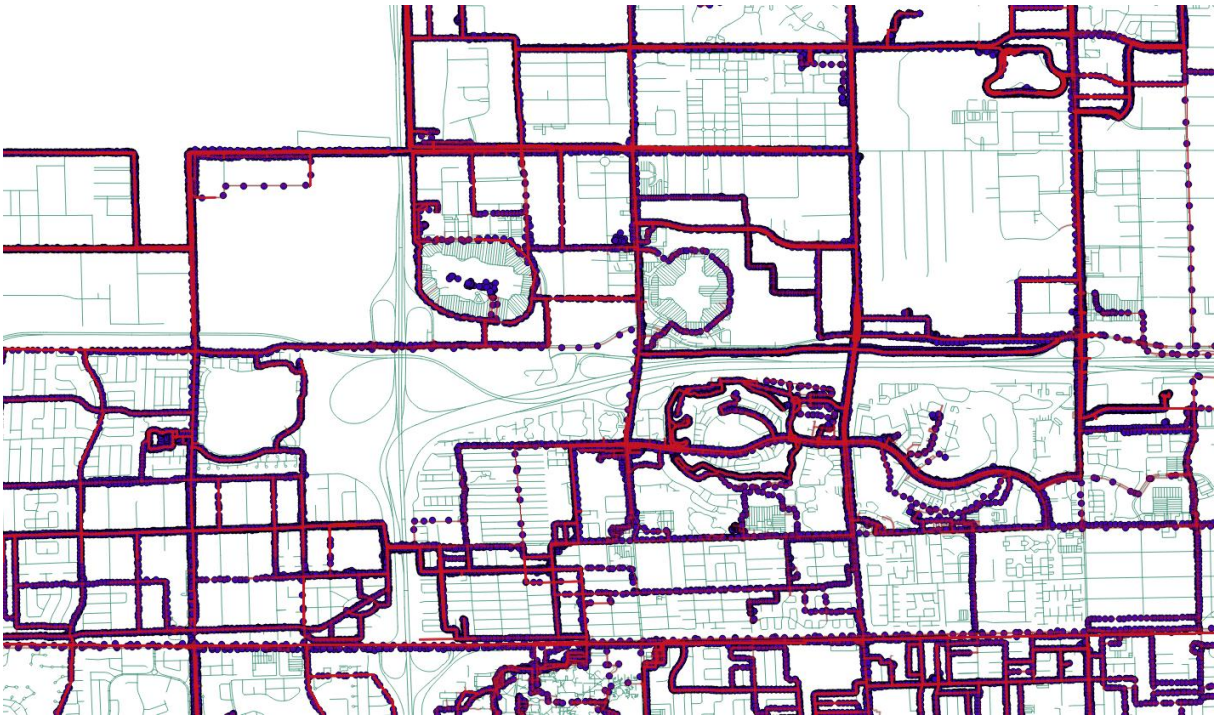


Figure 25. Results of Map-Matching

The last step was the calculation of timestamps for the segments that did not get assigned to an unambiguous point. Endomondo data allows calculating timestamps with precision to one minute. Strava Metro, on the other hand, is classified by five time internals, four of which are five hours long, and the last one is four hours long. Thus the level of temporal detail achievable with Endomondo data is higher than needed for comparison with Strava Metro. For this reason we calculated the missing timestamps not proportionally to the restored path length but as a mean of the surrounding known values.

4.4. Calculation of Segment Based Usage Statistics

In Strava Metro the highest number of unique athletes per road segment counted in both travel directions is 3,394, the highest number of unique activities per segment is 2,954. The total number of unique Endomondo users that logged their workouts in Miami-Dade County in the 1st quarter 2016 is 567, the total number of workouts is 3,240. If the comparison of road usage by Strava and Endomondo cyclists relied on absolute values, it could only show the overall dominance of Strava. To see which roads are preferred by Strava athletes, and which are more popular among Endomondo users we needed to calculate proportions of kilometers travelled along each segment relatively to the total travelled kilometer volume. The concept of bicycle kilometer travelled (BKT) is a common measure of cycle flow assessment (Hochmair et al., 2017; Osama, Sayed, & Bigazzi, 2017). It is usually calculated by multiplication of activity counts per segment by the length of corresponding segments. BKT can be biased in case a certain route is submitted multiple times by the same cyclist, or in case of

mass cycling events. Before the BKT calculation we had to determine the segments that could produce biased BKT values.

First we calculated the number of activities per athlete for each segment and then analyzed whether there were segments with unreasonably big counts. The activity counts per user varied from 1 to 6.8, which are reasonable numbers for a three-month period. For this reason we did not exclude any of the segments from further analysis based on the activity counts per user.

It is not possible to identify from an athlete count on its own whether its high value attributes to a mass cycling event or to a popularity of the route. Also, counts arising from the cycling mass events skew data to the less extent than repeatedly submitted routes. In our dataset approximately 6% of all segments are extreme outliers in terms of athlete counts. As these segments are located mostly along the beach which is a popular tourist and recreational area, or on intersections, we did not exclude them from the further analysis.

To illustrate differences between weekend and weekday we calculated BKT for weekend and weekday data separately. The calculation of BKT followed the following formula:

$$BKT = \left(\text{activity count } \textbf{in} \text{ the digitization direction} + \text{activity count } \textbf{against} \text{ the digitization direction} \right) \times \text{segment length}$$

Endomondo BKT counts were calculated in the following way. All segments corresponding to workouts were saved in the `restored_route` table. For each segment we counted the number of occurrences with a distinct workout ID in digitalization direction plus against the digitalization direction where day of the week in the timestamp had a value from 1 to 5 (Monday to Friday). This resulted in workout counts for each segment. We multiplied the resulting counts by segment lengths, which produced weekday BKT values. The same calculation for the day of the week values equal to 0 (Sunday) or 6 (Saturday) resulted in weekend BKT values. Further we normalized both Strava and Endomondo data by calculating z-scores to enable comparison between Strava and Endomondo. To see which streets were preferred by users of Endomondo, and which by Strava athletes, we subtracted Endomondo z-scores from Strava z-scores.

4.5. Calculation of Area-Based Statistics

Area-based statistics served the following goals:

1. To compare areal preferences of Endomondo users and Strava athletes: in which census tracts the users of each app tend to cycle more eagerly than in other tracts

2. To draw conclusions from attributes present in Endomondo and absent in Strava, such as distance of workouts or age and gender of users.
3. To compare user preferences of roads according to their functional classes between Strava and Endomondo, as well as temporal pattern in cycling.

For the first comparison we have calculated the BKT for Endomondo and Strava for each census tract. Figure 26 shows that absolute count of Endomondo BKT is bigger than Strava BKT in 16 out of 518 census tracts. As Strava has an overall quantitative dominance over Endomondo, in order to analyze the relative preferences of users towards certain areas it was necessary to normalize the BKT counts.

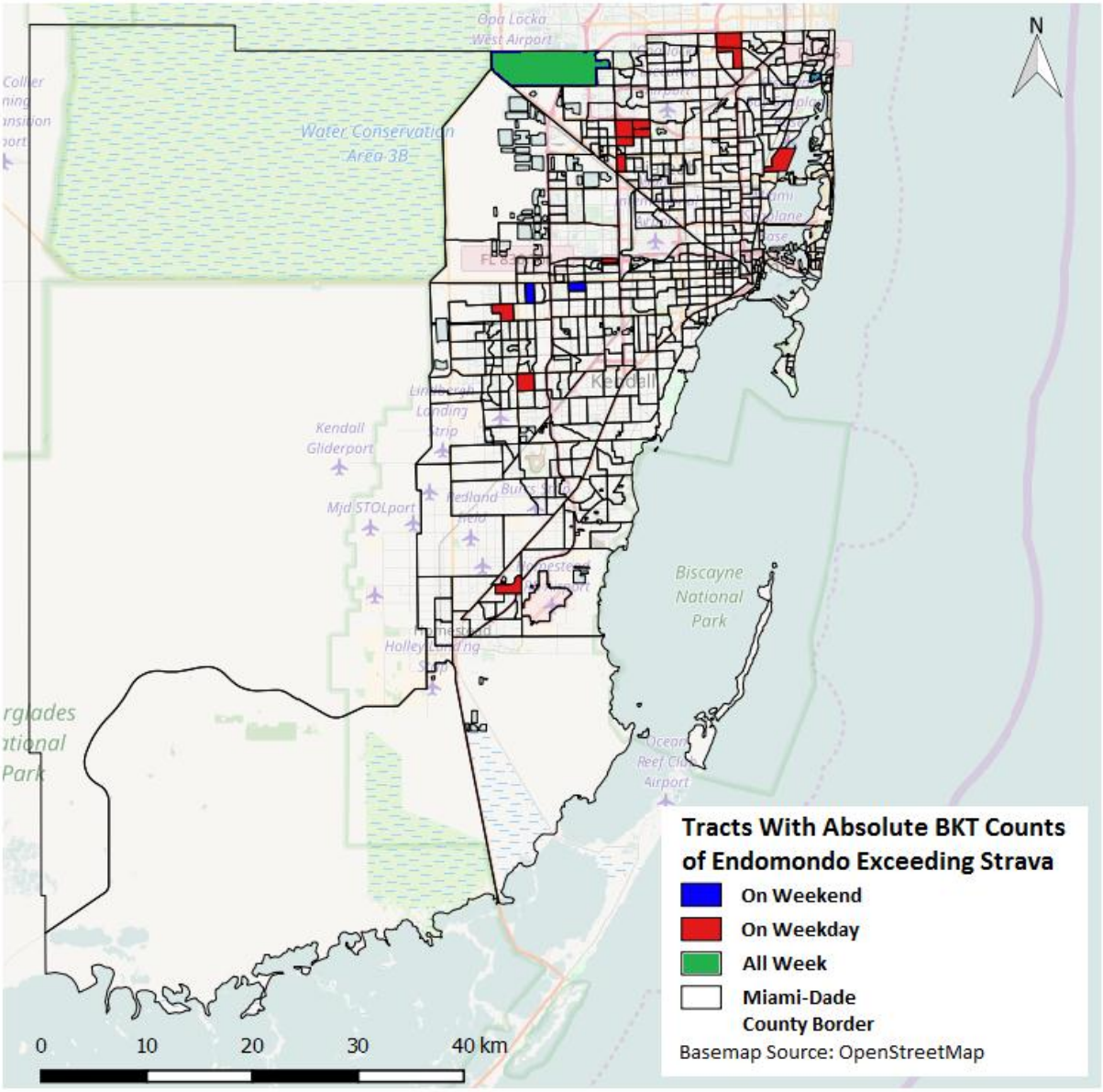


Figure 26. Tracts Where Endomondo Absolute BKT Counts Exceed Strava Absolute BKT Counts

The normalization was done in two steps:

1. We normalized BKT counts in each tract relatively to the length of all street segments in the tract to compensate for differences in tract areas and road network density. Such normalization was done by dividing BKT counts by total road length in the tract.
2. We standardized values resulting from the previous operation to an interval from 0 to 1 for both Strava and Endomondo with the help of the following formula:

$$standardized\ value = \frac{normalized\ value\ in\ the\ current\ tract}{(max_{normalized\ value} - min_{normalized\ value})}$$

Thus for each tract we have received relative preference of the tract compared to other tracts by users of a certain app. To see whether there were differences between preferences of different tracts between Strava athletes and Endomondo users, we subtracted the normalized Endomondo values from the normalized Strava values. Most of the resulting differences were very close to 0. The negative values showed that Endomondo users preferred the corresponding tracts over other areas slightly more than Strava athletes. Positive values indicated a slight preference of Strava athletes to cycle in the corresponding tracts over other areas in comparison to Endomondo users. Values very close to 0 indicated that both Strava athletes and Endomondo users were equally eager to cycle in the corresponding areas. In several areas there were no Endomondo workouts while Strava activities were present. Resulting maps are presented in the next section.

We carried out a general analysis of Endomondo data based on gender, date of birth and country of registration provided by Endomondo users, user's date of registration, total workout summary, location of workout GPS points and metadata of individual workouts.

For the analysis of age and gender data we calculated proportions of users belonging to each age/gender group for each reference area, namely Miami-Dade County, Florida, the USA and the world. We also analyzed the mean Endomondo user age for all Florida counties where the number of distinct users in the county that have provided their date of birth was at least 30, since the distribution of Endomondo users according to their age is approximately normal.

We analyzed the proportions of Endomondo users according to their country of registration in Miami-Dade County, Florida and the USA. Further, we calculated proportions of users from different countries of registration for Endomondo users worldwide. Proceeding from the each user's date of registration we analyzed the user count change in four different regions of the world: Miami-Dade County, Florida, the USA and worldwide to see whether the trends are similar or there are regional differences.

We analyzed Endomondo workout metadata and calculated the mean, median and trimmed mean values of workout duration, speed and distance for all counties in Florida where workout count was

at least 30. As we did not uncover any areas where these three measures did not correlate, in this thesis we illustrate only spatial variations in workout distances.

For the last goal, namely comparison of temporal patterns and road class preferences, we calculated the BKT counts for Endomondo users and Strava athletes for each of the five time intervals according to Strava classification, as well as for each road class and normalized the resulting absolute counts to proportions in order to compensate the quantitative predominance of Strava. The following section contains maps and charts that resulted from the preparations described above.

5. RESULTS

5.1. Comparison of Endomondo and Strava Data

Though this was mentioned in the previous sections, we would like to draw special attention to the fact that data analyzed here represents a three-month period from January to March. Thus it is not necessarily representative for spatio-temporal variations and patterns in cycling with the use of Endomondo and Strava apps at other times of the year.

5.1.1. Data Coverage

As Strava is originally US based, we expected to see that Strava coverage is more extensive than the one of Endomondo. As Strava Metro data does not contain information on individual trajectories, the total number of workouts in Miami-Dade County logged by Strava is unknown. However, there are segments in the road network attributed with a count of more than as 4,000 individual activities, which is more than a total count of Endomondo activities in the whole county during the same period of time. But quantitative dominance of Strava does not necessarily mean that the spatial distribution of Strava activities is more extensive that the one of Endomondo workouts.

A map presented by Figure 27 demonstrates the spatial distribution of Strava activities in Miami-Dade County differentiating according to the day of the week. Grey road segments are those with no Strava activities. After having retrieved the Endomondo GPS point data, we have overlapped this map with raw Endomondo trajectories. The resulting map is given by Figure 28. It shows that there are several areas where Endomondo has no coverage while Strava data there is present. At the same time, there are certain regions, mostly of small area, where Endomondo tracks are observed and the Strava data is absent. The segment based Kernel Density Estimation maps that predict BKT for Strava (Figure 29) and Endomondo (Figure 30) on weekday and weekend provide information not only on coverage of both apps but also on spatio-temporal intensity of cycling with the use of these apps in Miami-Dade County. The possible conclusions from the patterns revealed by these maps are discussed in subsection 6.1.

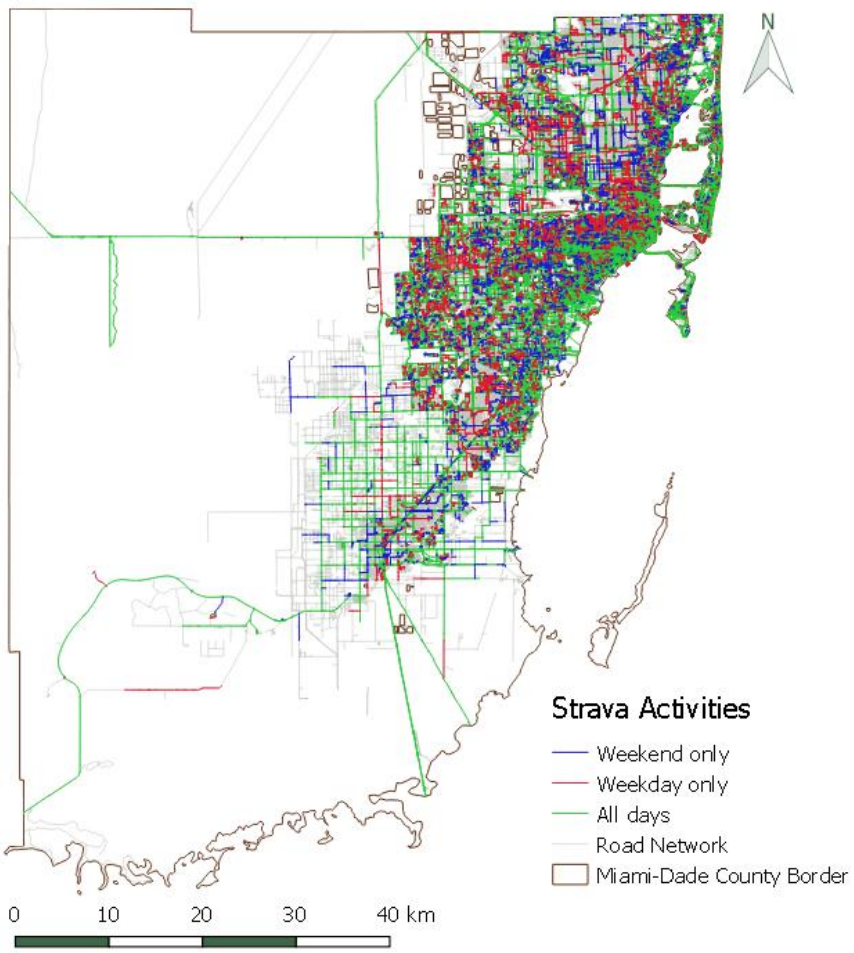


Figure 27. Segment Based Distribution of Strava Activities in Miami-Dade County

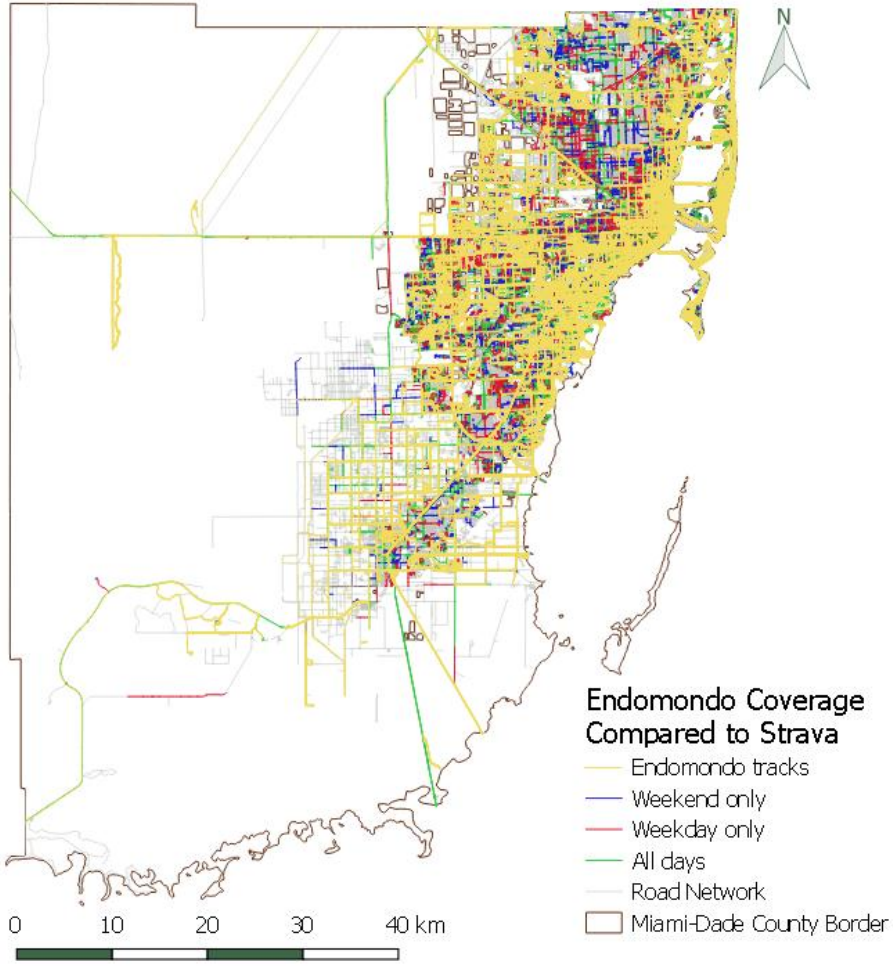


Figure 28. Strava Spatial Distribution Map Overlapped With Raw Endomondo Tracks

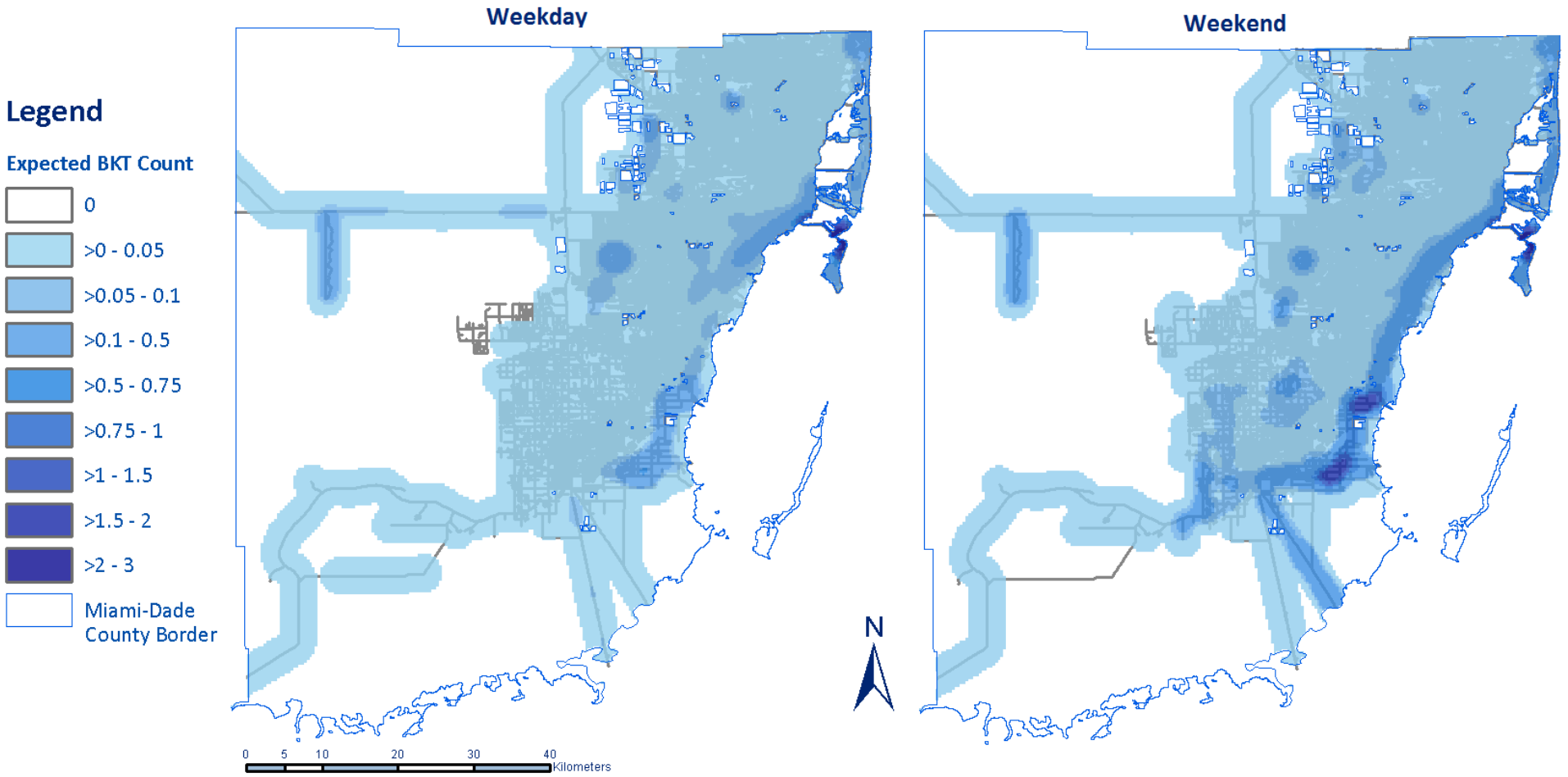


Figure 29. Kernel Density Estimation Maps of Strava Activities Based on Predicted BKT Counts per Cell

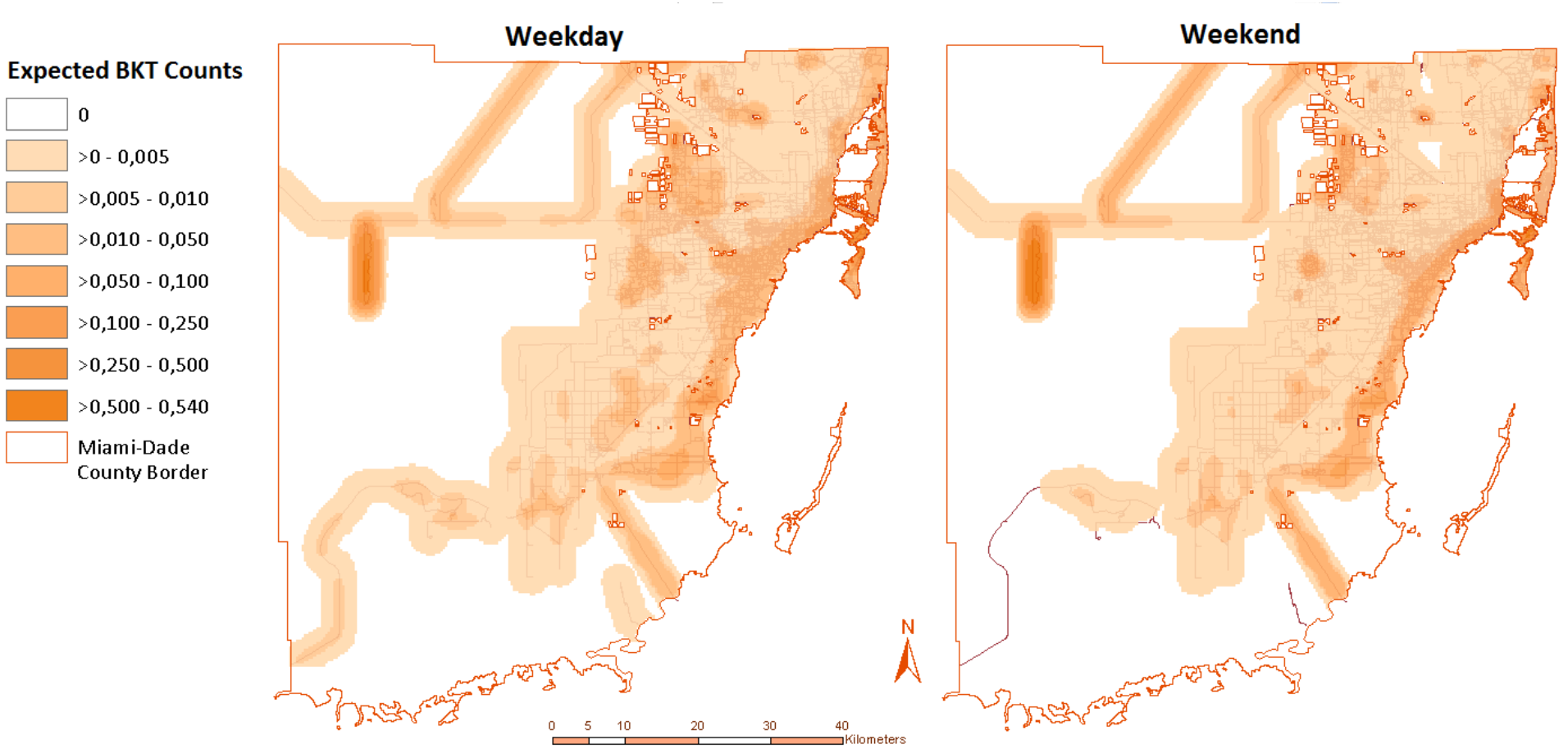


Figure 30. Kernel Density Estimation Maps of Endomondo Activities Based on Predicted BKT Counts per Cell

After comparing the overall activity distribution between Strava and Endomondo we proceeded with the area based statistics of Endomondo and Strava usage in these sub-regions. The goal .of this analysis was to identify travel patterns in different sub-regions of Miami-Dade.

5.1.2. Bicycle Travel Patterns across Different Sub-Regions and Different Road Classes

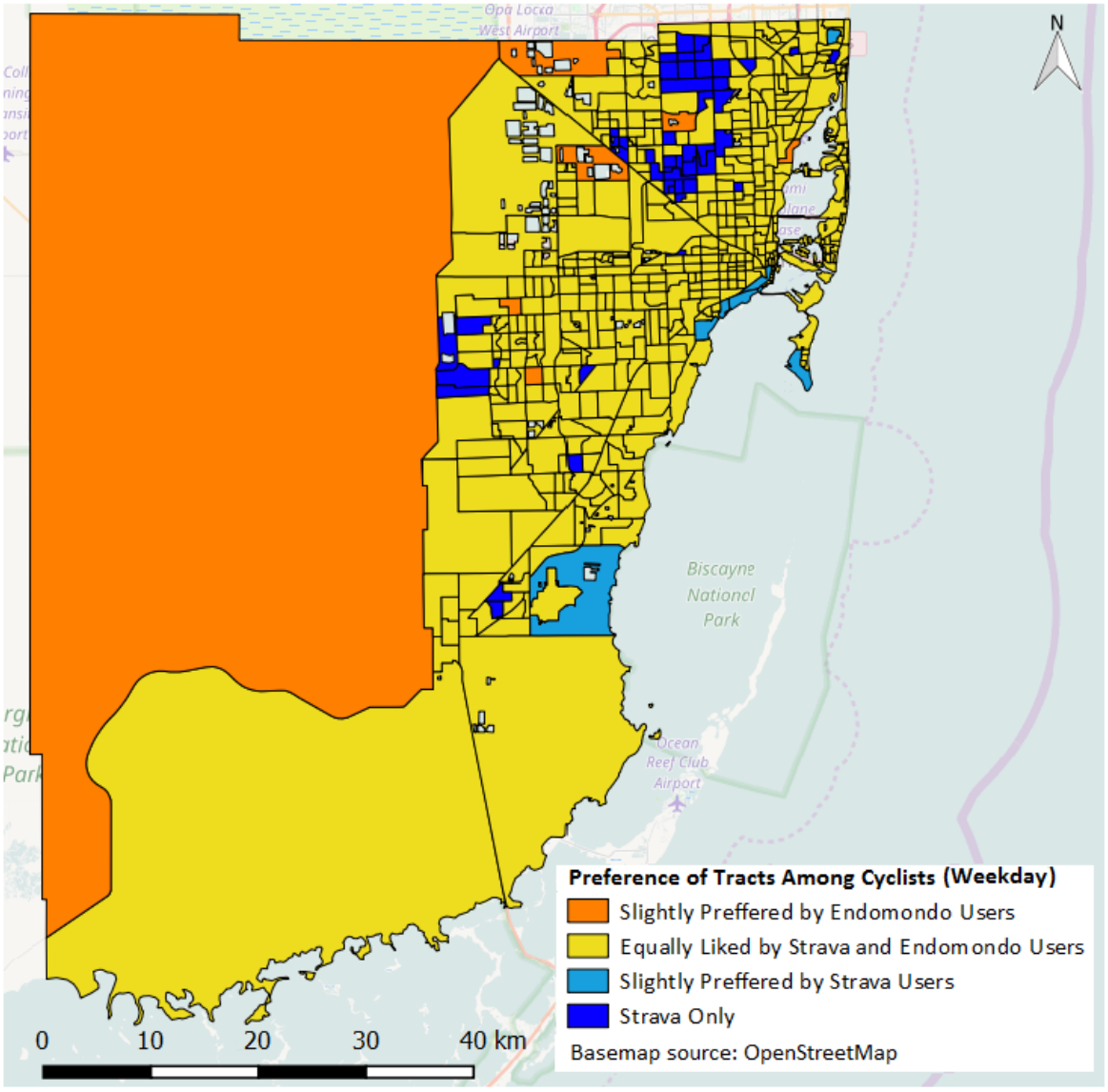


Figure 31. Weekday Areal Preferences of Endomondo and Strava Cyclists

A possible way to analyze the cycling travel patterns is to assess the cyclists' relative preference of certain areas. Miami-Dade County is constituted of 518 census tracts that encompass the population from 2,500 to 8,000 people each and have relatively permanent borders. The data used for census tract representation was retrieved in a form of a shapefile from the official website of the United States Census Bureau (https://www.census.gov/geo/maps-data/data/tiger-line.html). It represents the census tract geometries for Miami-Date County from the year 2010.

Figure 31 demonstrates the weekday preferences of Miami-Dade cyclists. One of the census tracts located to in the Western part of the County has a large area and a very low population density. There are also a few roads in this area. Though the map seems to display that a big portion of the county is more popular among the Endomondo users than among the Strava athletes, in fact the preferences of users of both apps are about the same. There are certain areas where Endomondo data is absent, which is already known from figures considered above.

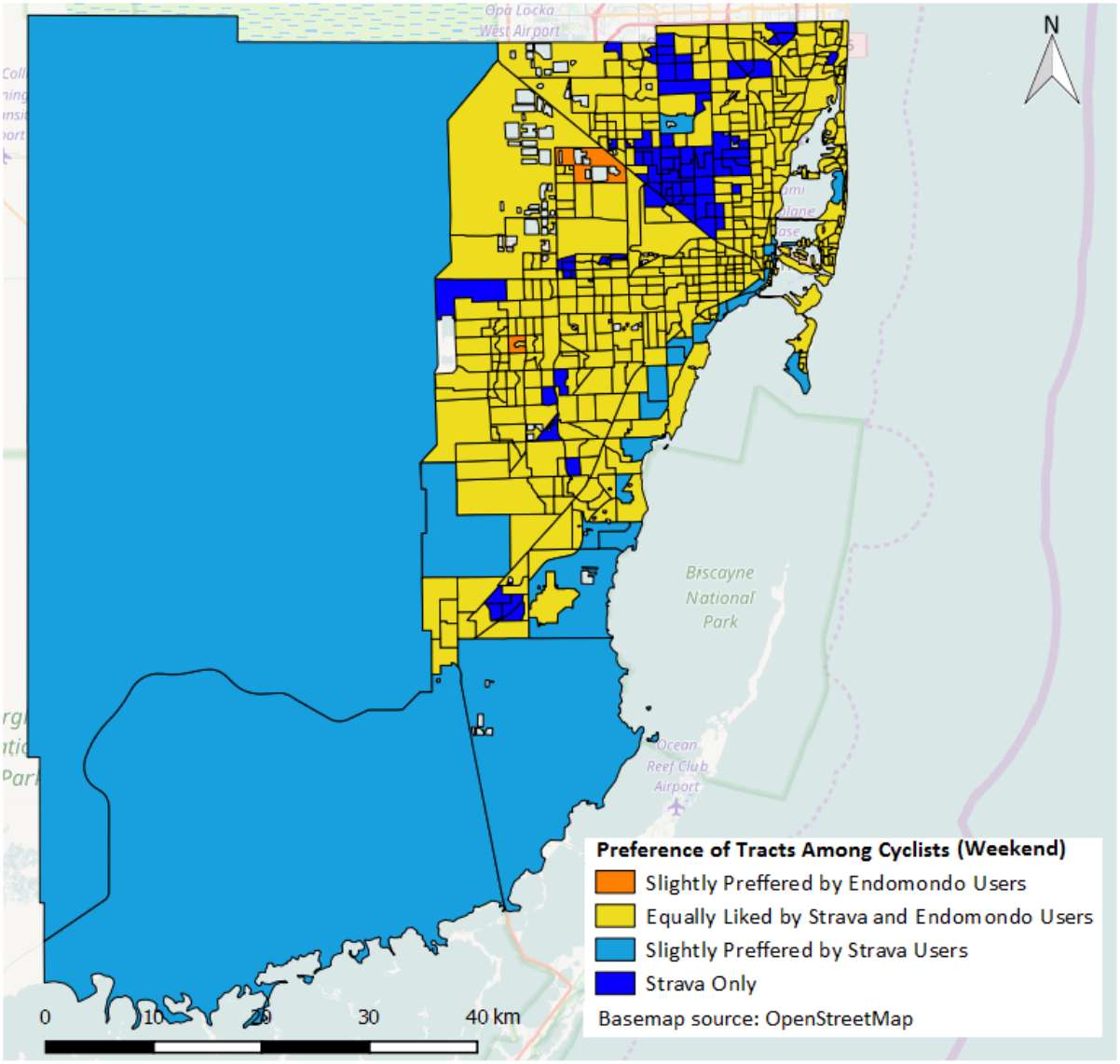


Figure 32. Weekend Areal Preferences of Endomondo and Strava Cyclists

The map presented by Figure 32 is based on weekend activity distributions. It differs from the previous map in a meaningful way implying an overall Strava dominance and showing that regions that are more popular among Endomondo users on weekday can be preferred to a greater extent by Strava athletes on weekend.

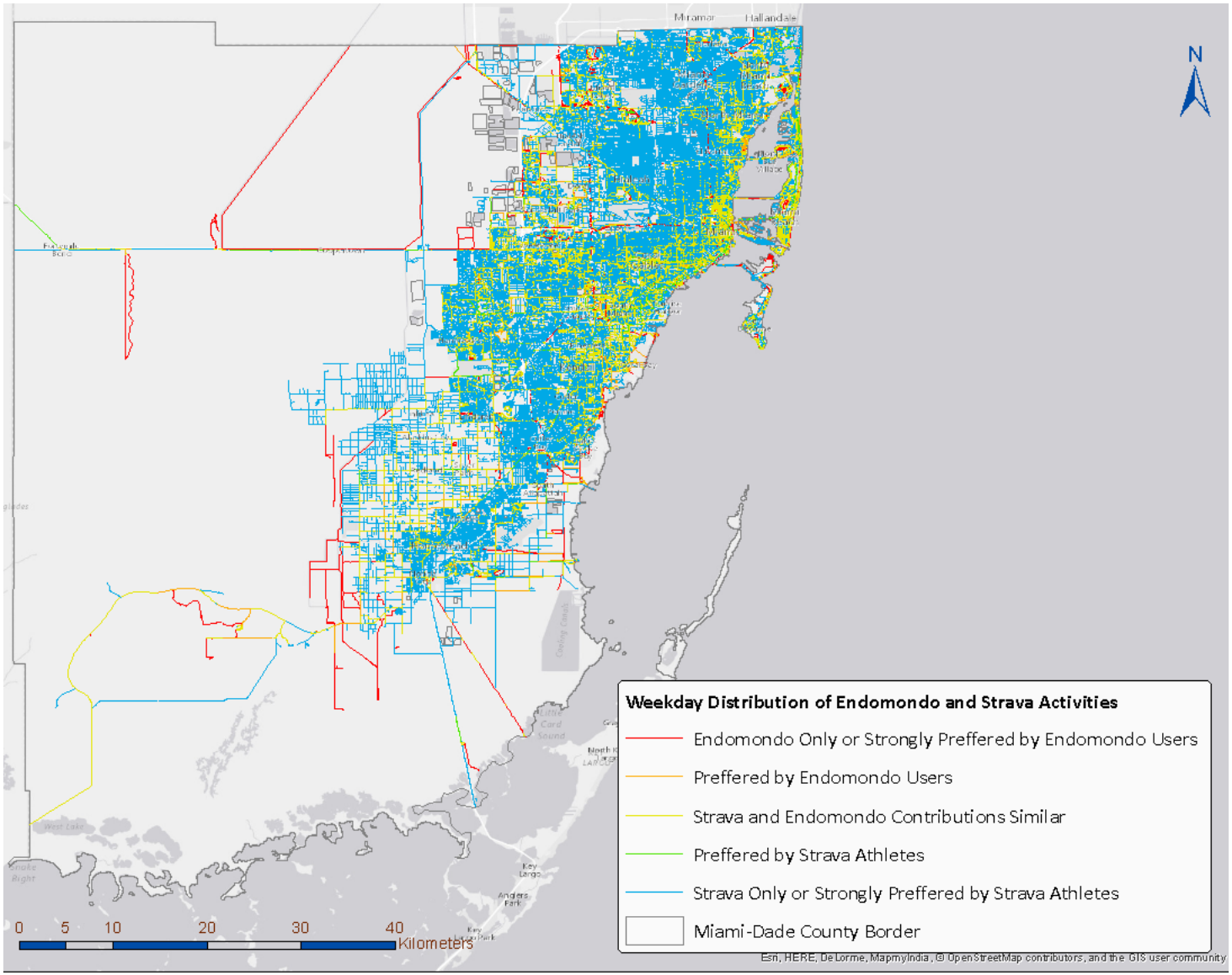


Figure 33. Segment Based Weekday Distribution of Endomondo and Strava Activities

The map to the left (Figure 33) shows the weekday distribution of Endomondo and Strava activities in Miami-Dade at the segment level. Classification of preferences is done based on the differences of z-scores. The z-scores of BKT counts were calculated for each segment. In case the z-score difference for a segment varies between -0.5 and 0.5, a segment belongs to the category of similar contributions. For z-scores between 0.5 and 3 in the negative or in the positive direction, the segment is considered to be preferred by users of Strava (if the value is positive) or Endomondo (for a negative z-score). For the negative z-scores under -3 the segment is called to be strongly preferred by Endomondo users. Segments with Endomondo data only also belong to this category. The segments with high positive z-scores (over +3) are classified as strongly preferred by Strava athletes. The segments with Strava counts only belong to this category, too.

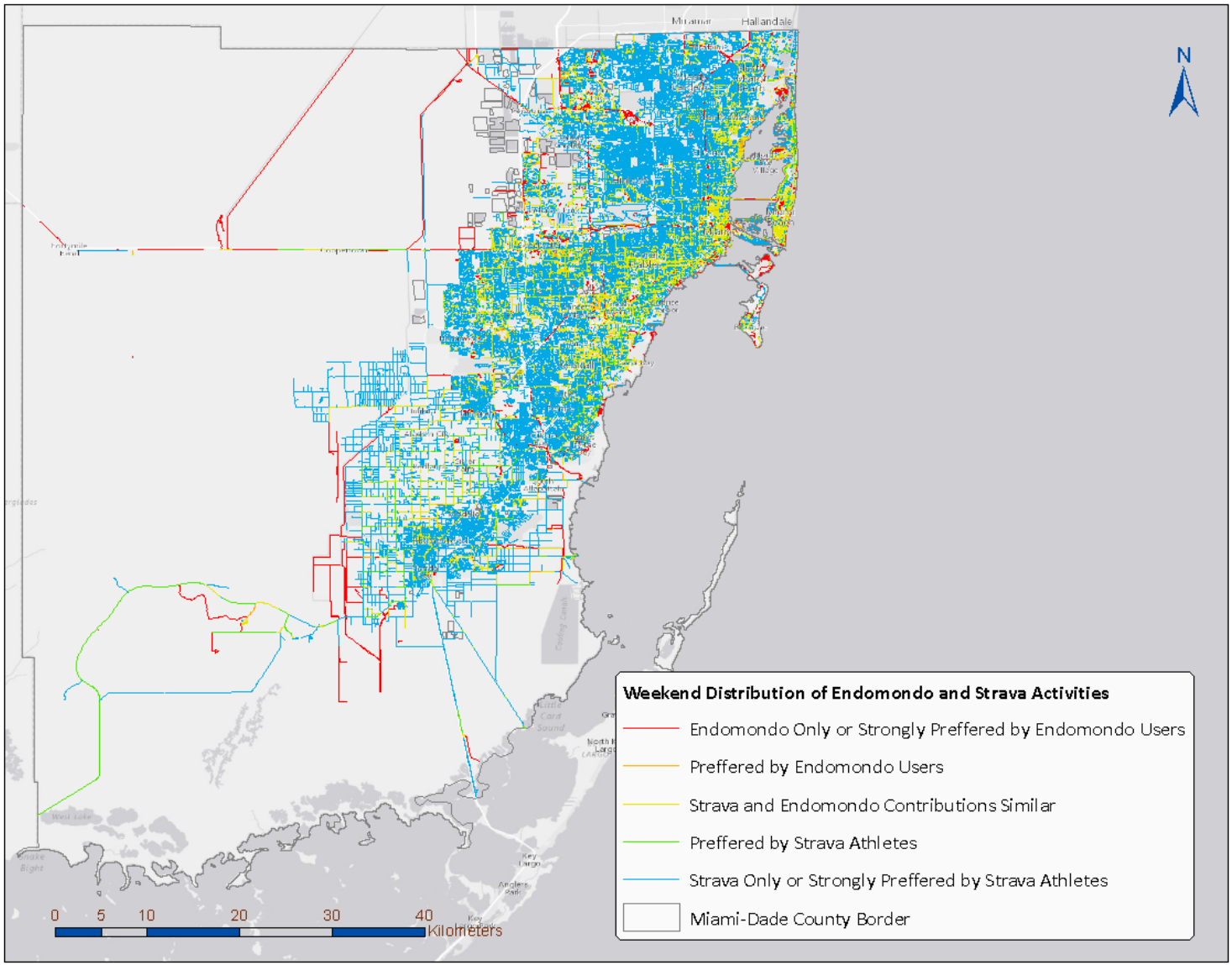


Figure 34. Segment Based Weekend Distribution of Endomondo and Strava Activities

The map displayed by Figure 34 shows the weekend segment based distribution of Endomondo and Strava activities. Though this map provides a higher level of detail than the one representing census tracts, it is also more difficult to read. The most obvious differences from the segment based map of weekday activities can be seen in the Western and South-Western parts of the county. Many streets that were preferably used by Endomondo users on weekday are now marked as preferred or strongly preferred by Strava athletes. This is consistent with the difference between the maps accessing BKT counts in the census tracts. In general, the segment based maps provide good support to the tract based statistics.

The preferences of Strava and Endomondo users towards different road classes (Figure 35) proved to be similar to a certain extent. We also identified some differences that were expected considering the fact that Strava Metro data relies solely on Here Navstreets geometries, while we matched Endomondo GPS points to the road network that was expanded mostly by adding road segments of the fifth functional class.

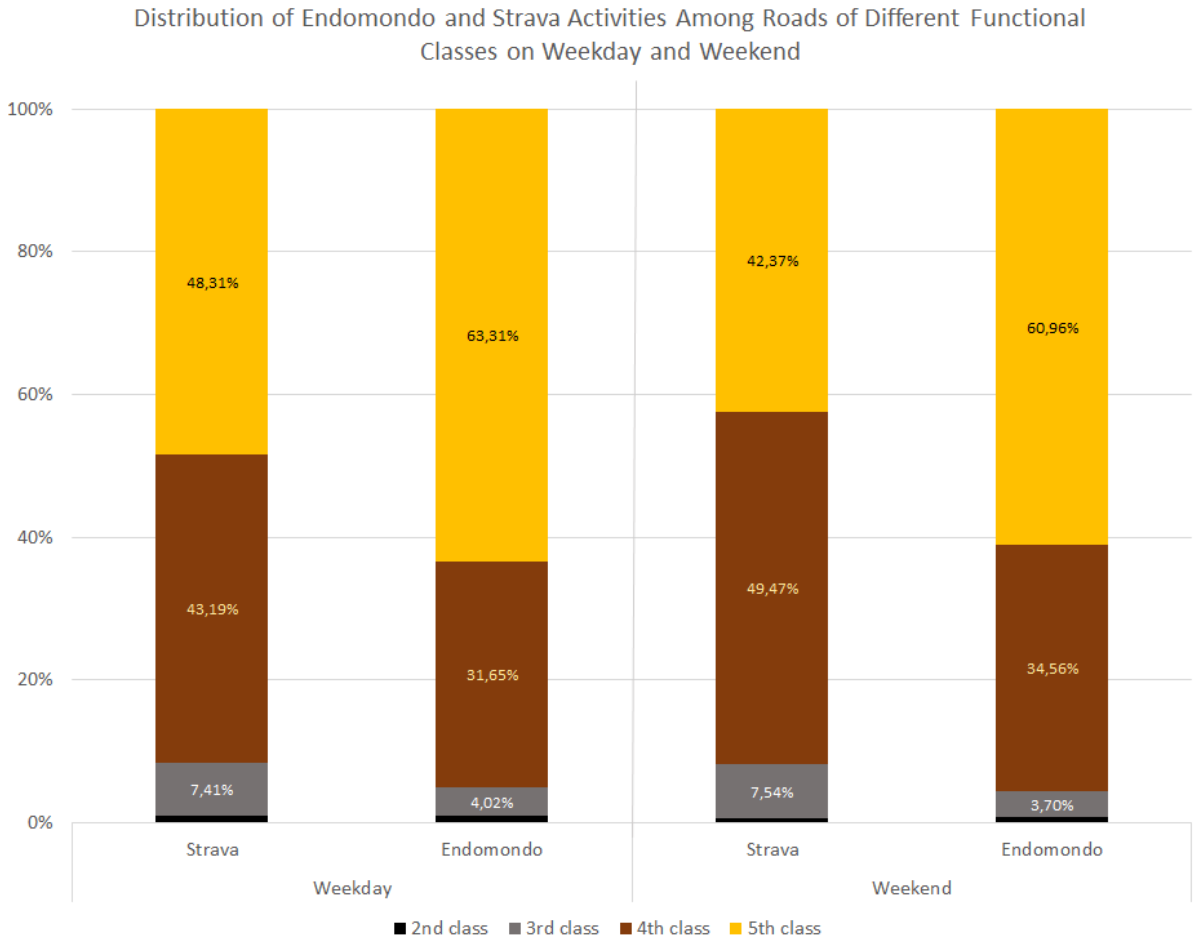


Figure 35. Distribution of Endomondo and Strava Activities Among Roads of Different Functional Classes on Weekday and Weekend

Roads of the second functional class have heavy traffic and are mostly avoided by cyclists using both apps. Strava athletes use roads of the third and fourth functional classes more than Endomondo users both on weekday and on weekend, which can be attributed to the fact that Here Navstreets dataset does not cover some of the recreational areas popular among cyclists and thus the activities of the Strava athletes in these areas are not reflected in the Strava Metro data. Subsequently Endomondo users seem to use the roads of the fifth class more actively than Strava users. The differences in the usage of roads of different functional classes in connection with the time of the day are given in the next subsection.

5.1.3. Temporal Usage Patterns

Figure 36 shows temporal distribution of Strava activities among the roads of different functional classes based on BKT. The weekday chart has two peaks, a higher one in the morning and a lower one in the late afternoon – early evening. The weekend cycling according to Strava data, however, has just one morning peak.

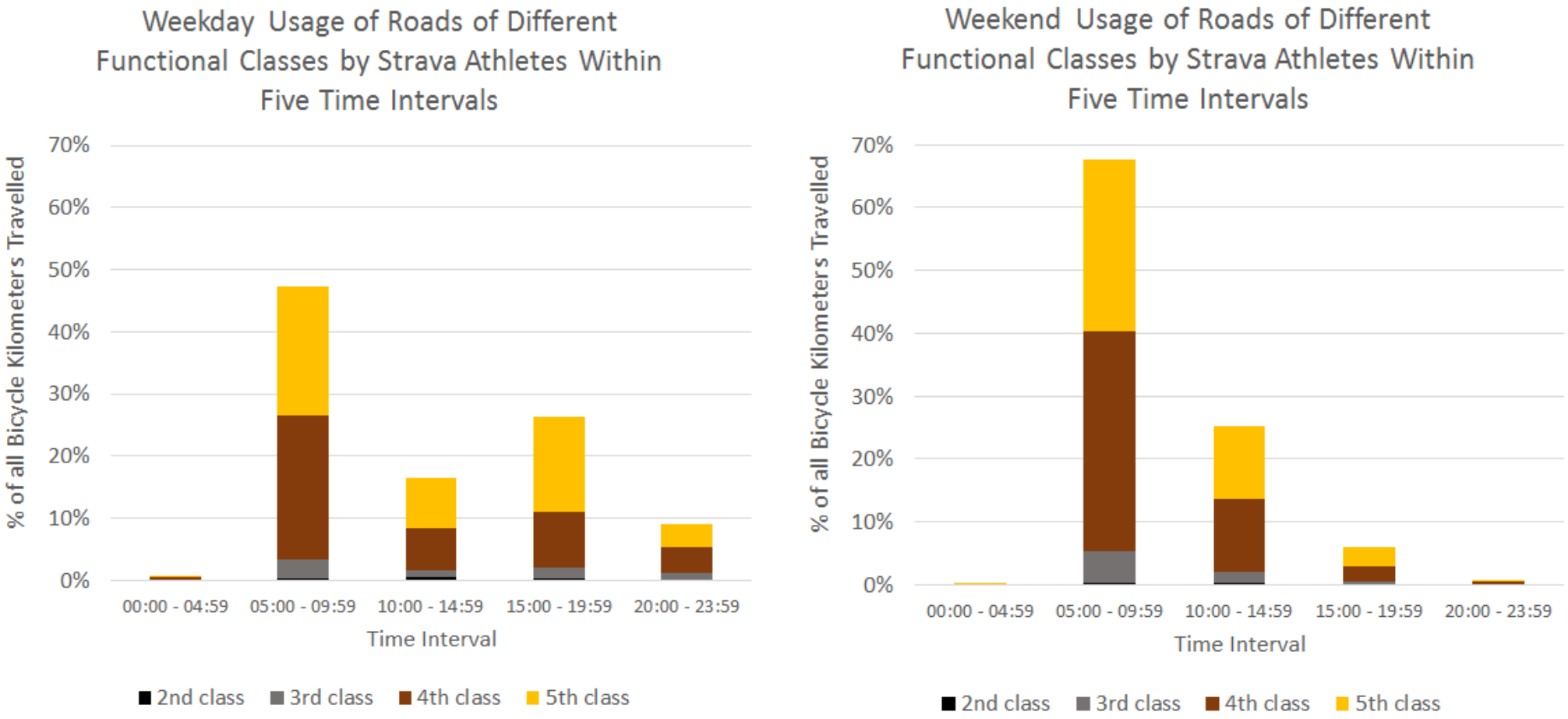


Figure 36. Temporal Distribution Of Weekday and Weekend Strava Cycling Activities Based on BKT

The proportions of BKT travelled along roads of different functional classes do not stay equal. On booth weekend and weekday the streets of the third functional class are used in the morning much more than during the day. The streets of the fourth class are also mostly used in the morning, and the BKT on these streets in the morning hours exceeds BKT on the roads of the fifth class.

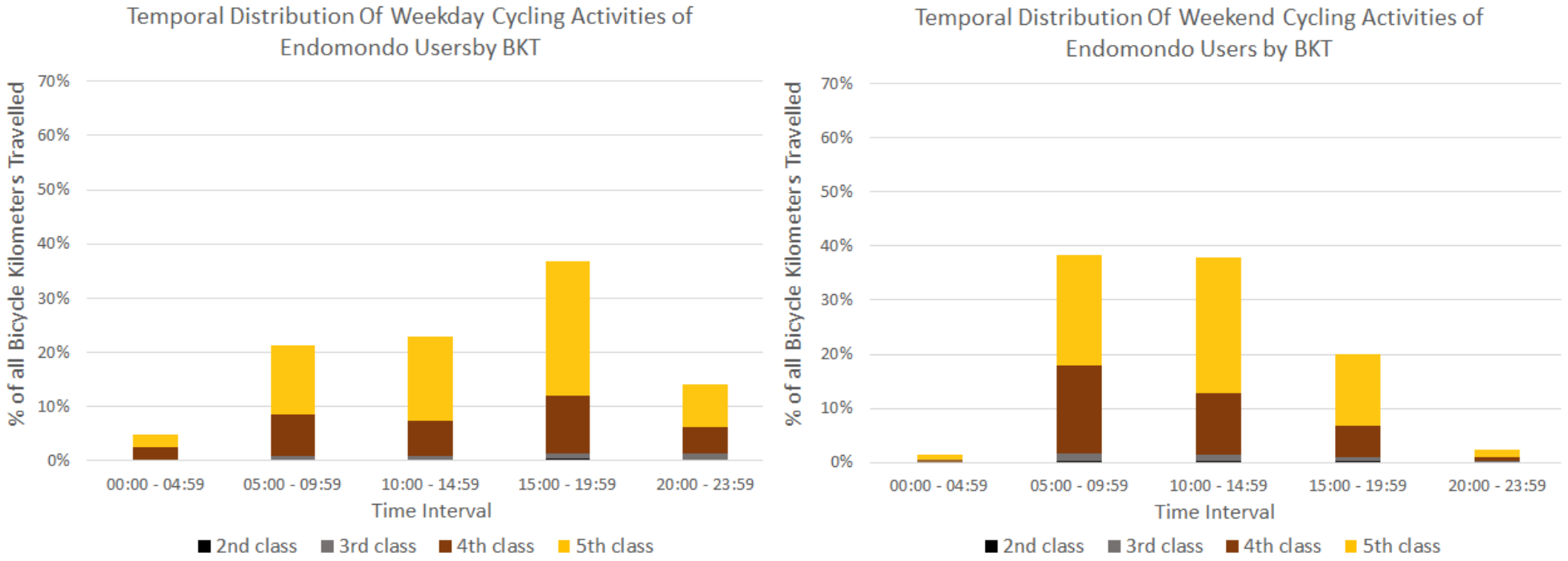


Figure 37. Temporal Distribution Of Weekday and Weekend Endomondo Cycling Activities Based on BKT

Figure 37 provides the temporal distribution of Endomondo workouts among the roads of different functional classes. This distribution is different from Strava distribution since (1) the overall tendency is that the roads of the fifth class are used to some extent more than roads of other classes irrespectively of the day of week or time of the day; (2) there is only one peak in weekday distribution, and the overall change in counts is less abrupt than by Strava. The next figure (Figure 38) illustrates the differences in temporal activity distributions disregarding the road classes in a comparative manner.

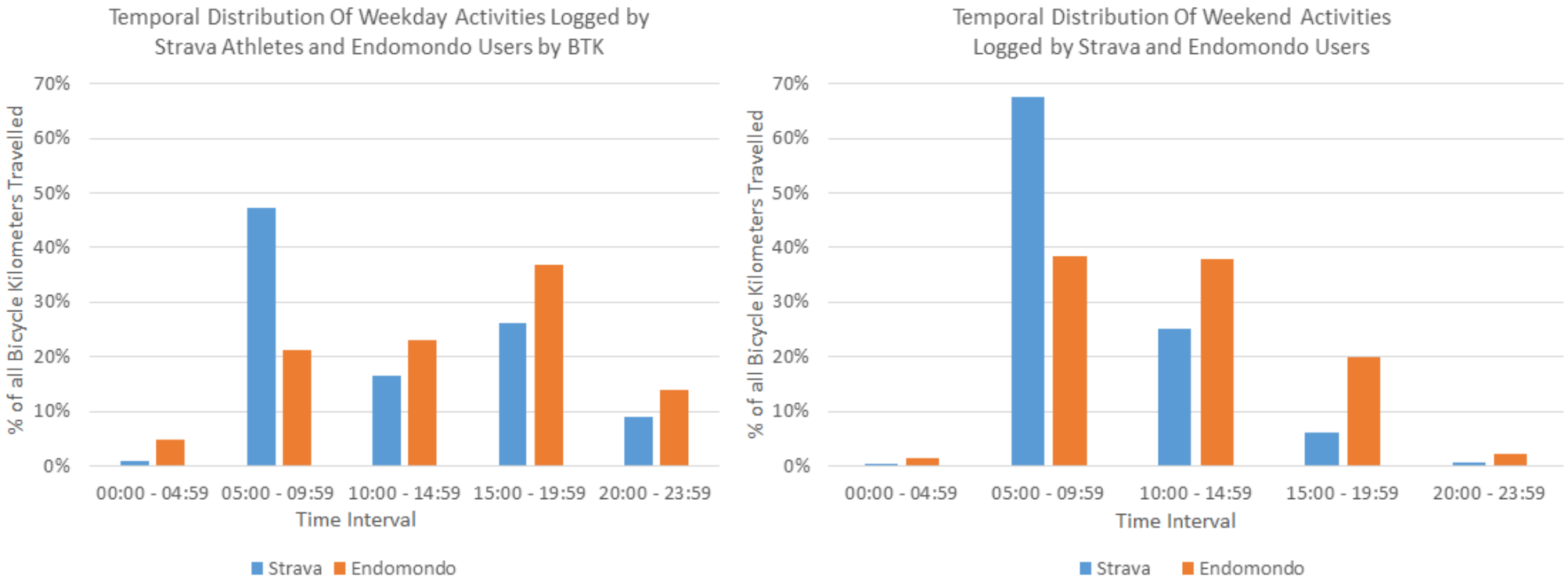


Figure 38. Comparison of Temporal Cycling Activity Distributions between Strava and Endomondo Based on BKT Proportions

Time Interval	Strava		Endomondo	
	Weekday	Weekend	Weekday	Weekend
00:00 - 04:59	0,87%	0,33%	4,92%	1,33%
05:00 - 09:59	47,30%	67,64%	21,19%	38,45%
10:00 - 14:59	16,57%	25,21%	22,99%	37,95%
15:00 - 19:59	26,23%	6,05%	36,86%	19,95%
20:00 - 23:59	9,04%	0,76%	14,05%	2,32%

Table 6. Proportions of Endomondo and Strava Activities according to the Time of the Day

Figure 36 identifies the differences in cycling temporal patterns between Endomondo users and Strava athletes. Additionally to the issued discussed above, Endomondo has a relatively big proportion of nighttime activities. None of the time intervals comprise more than 40% of Endomondo BKT, whereas Strava morning peaks include 47.3% of all BKT on weekday and 67.64% of all BKT on weekend (Table 6).

This completes the analysis based on attributes that Endomondo and Strava have in common. Additional information can be derived from Endomondo data, and this is the content of the next section.

5.2. Endomondo: Selected Usage Patterns

Endomondo data contains age and gender characteristics of users and their country of registration which allows to analyze the demographic characteristics of the Endomondo cycling community. Most of the Endomondo users provide information on their gender. In four analyzed regions the gender information was available for 89-96% of the users: Miami-Dade – 89%, Florida – 91%, the USA – 91.5%, the world – 96%. The gender structure of Endomondo users by region is given by Figure 39. It is common that proportions of male and female cyclists are not equal, which is also the case with Endomondo data.

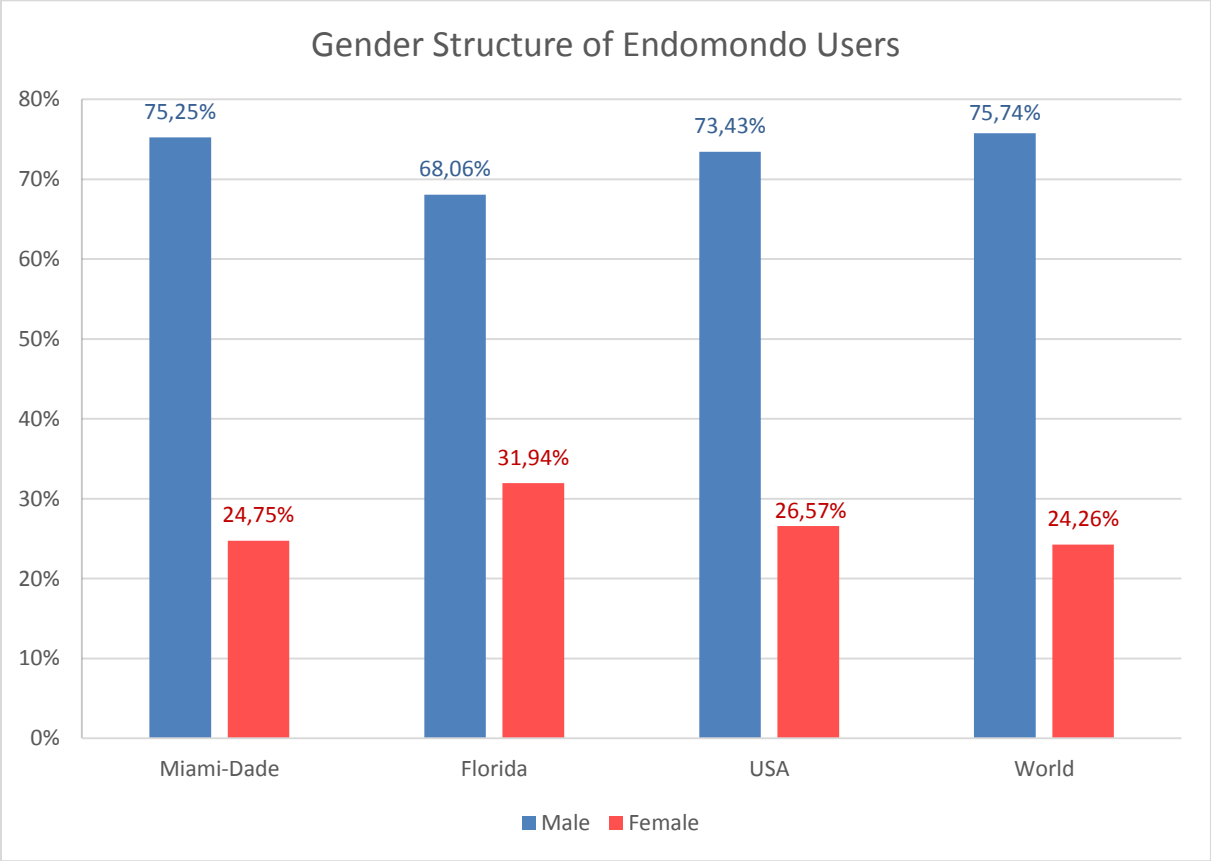


Figure 39. Gender Structure of Endomondo Users

Less than 60% of Endomondo users provide their birth date. We have analyzed the distribution of users of both genders within 5 age groups, namely under 15 years old, from 15 and up to 25 years old, from 25 and up to 40 years old, from 40 and up to 60 years old, from 60 and older. The resulting distribution is given by Figure 40. The distribution of cyclists of both genders among age groups is approximately normal. In the next chapter we discuss the identified trends in detail.

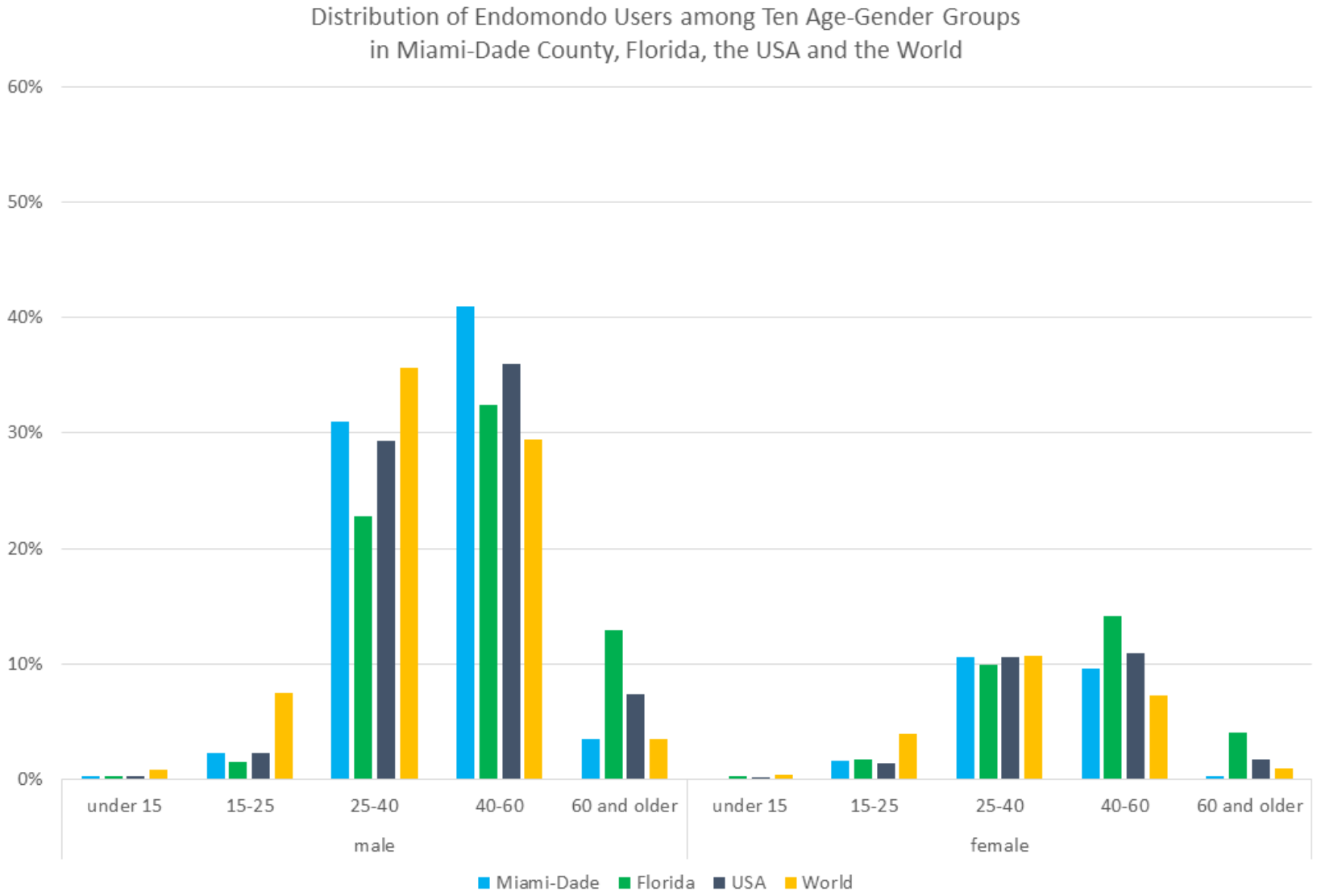


Figure 40. Distribution of Endomondo Users among Ten Age-Gender Groups

The next map (Figure 41) shows the distribution of Endomondo cyclists according to their mean age in the counties of Florida. Many counties have insufficient sample sizes (under 30). The map shows that the mean age of cyclists overall is 40 and higher, and several counties average in age of over 50.

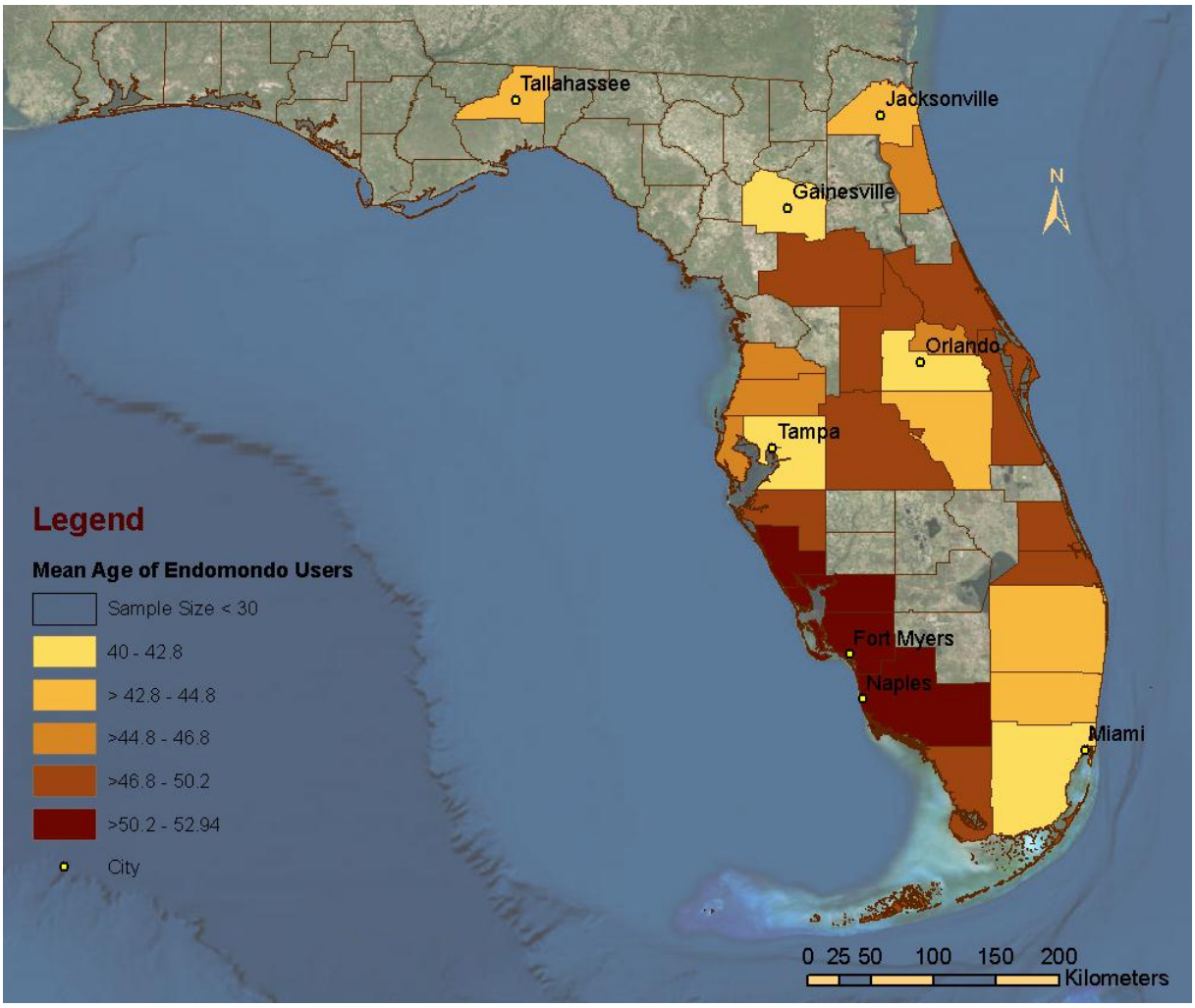


Figure 41. Mean Age of Endomondo Users

We compared the country of registration of Endomondo users cycling in Miami-Dade with those cycling in Florida and the USA to reveal whether there were certain regional patterns not common for the country as a whole. Figure 42 shows that the proportion of the users with the country of registration "USA" is bigger for Florida than for Miami-Dade (94.2% vs. 84.2%), and even bigger for the USA as a whole (96.4%). In Miami-Dade County the proportion of Endomondo users from Latin America is relatively high (7.1% vs. 1.75% for Florida and 0.65% for the USA). Proportions of users registered in Denmark, Great Britain and Poland are also relatively high compared to the state of Florida and the USA. Figure 43 shows that Poland and Denmark are the top two countries with the highest Endomondo user counts worldwide. The USA occupies the fourth place in this comparison.

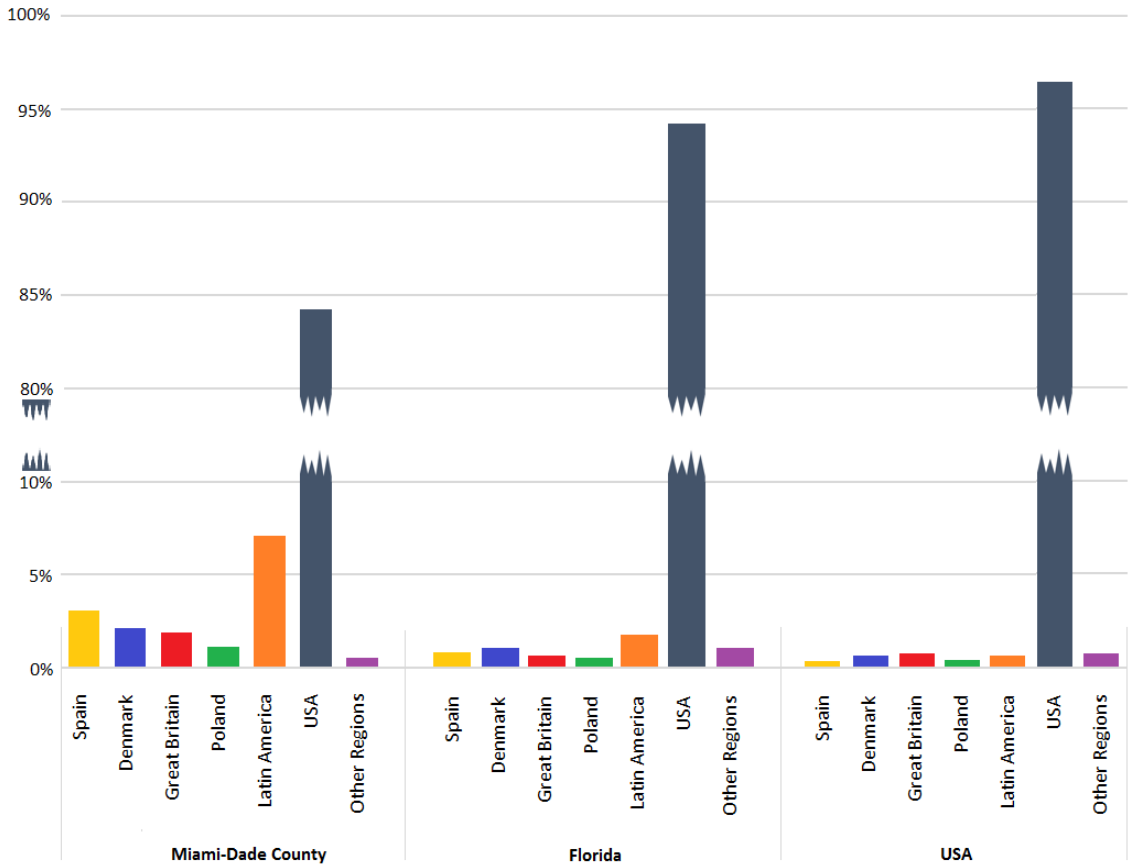


Figure 42. Distribution of Endomondo Cyclists according to the Country of Origin

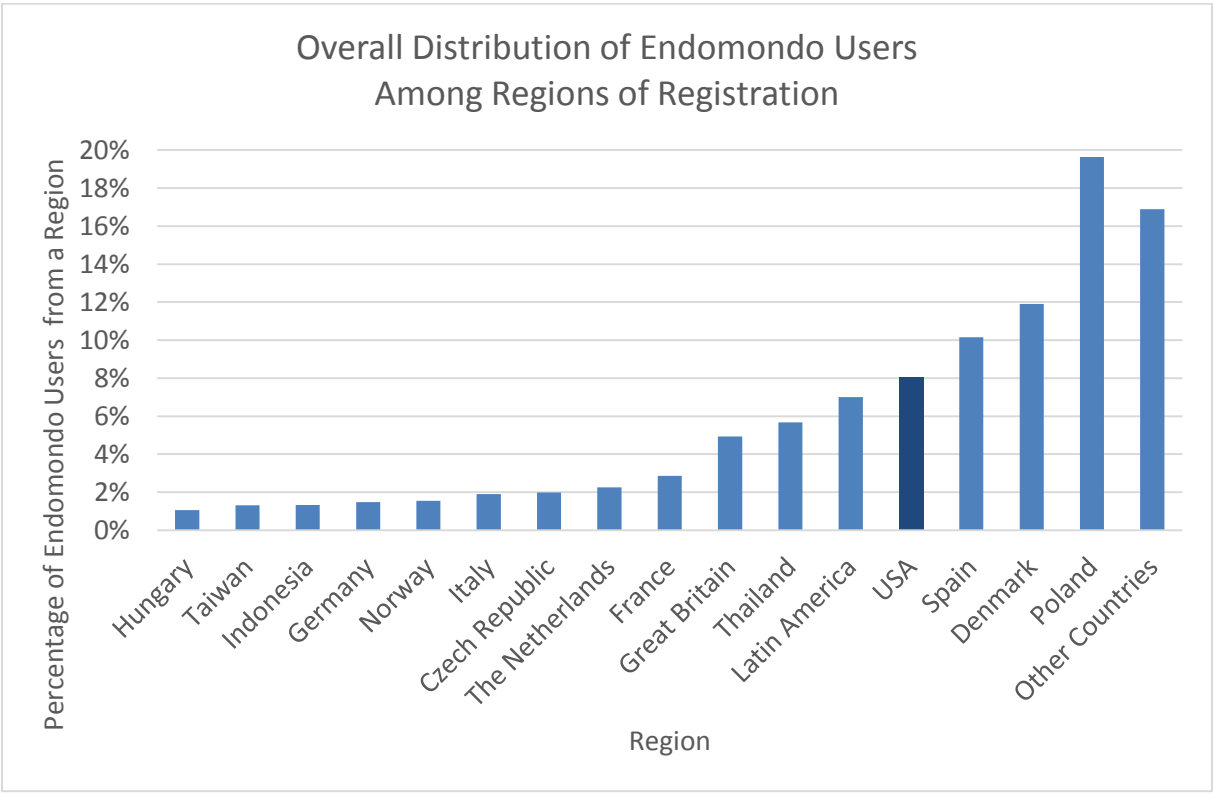


Figure 43. Overall Distribution of Endomondo Users

The next analysis involved workout distances. As Figure 44 shows, the Endomondo dataset for Florida contained some extreme outliers. The same was characteristic for the workout duration data. We excluded these outliers from the further analyses.

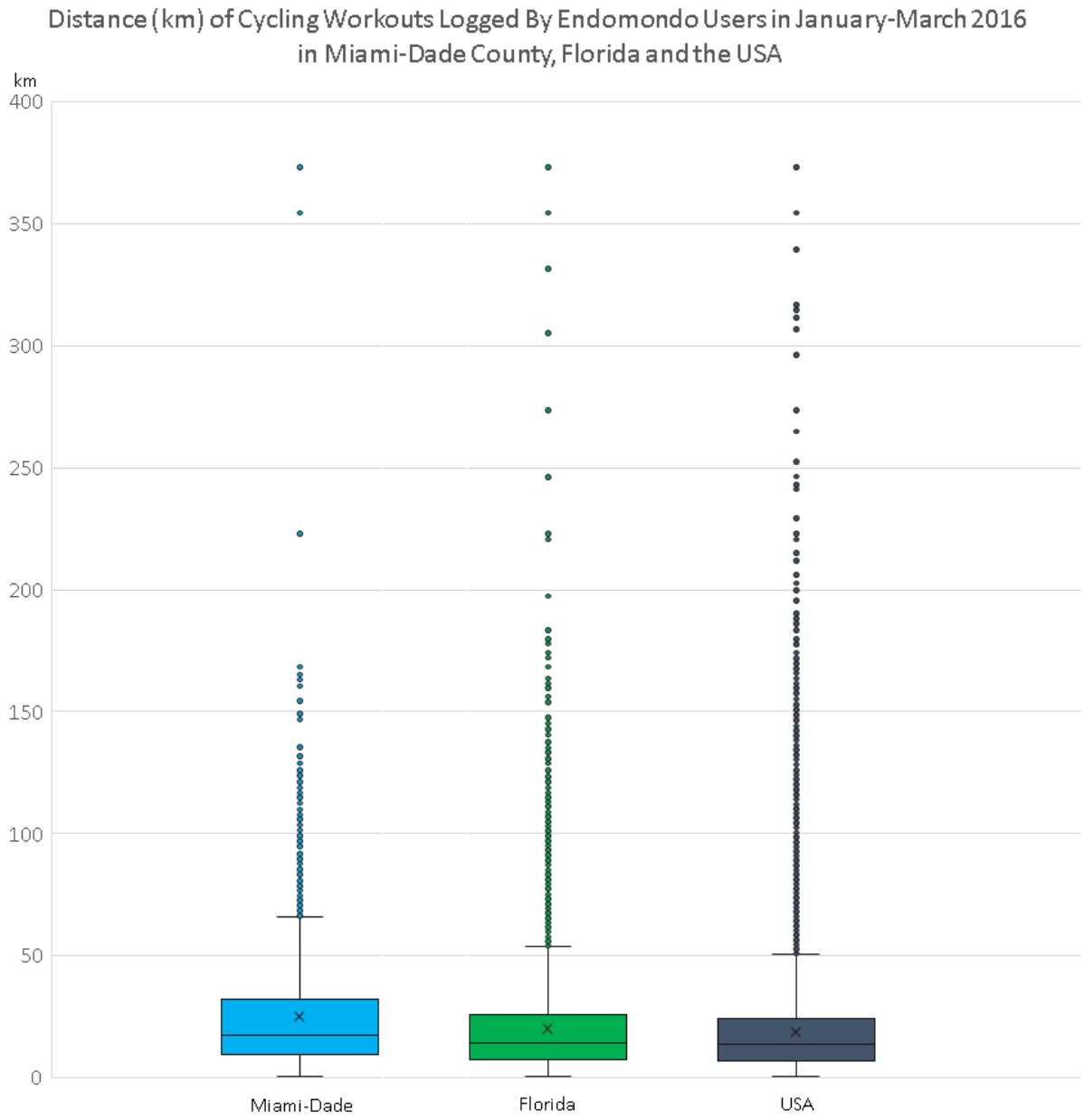


Figure 44. Boxplots of Endomondo Cycling Workout Distances in Miami-Dade, Florida and the USA

Figure 45 shows the mean distance values for cycling activities in Florida counties calculated after the outlier removal from the Endomondo dataset. For this analysis we disregarded the trip purpose and the age differences which will be considered in detail in the next chapter.

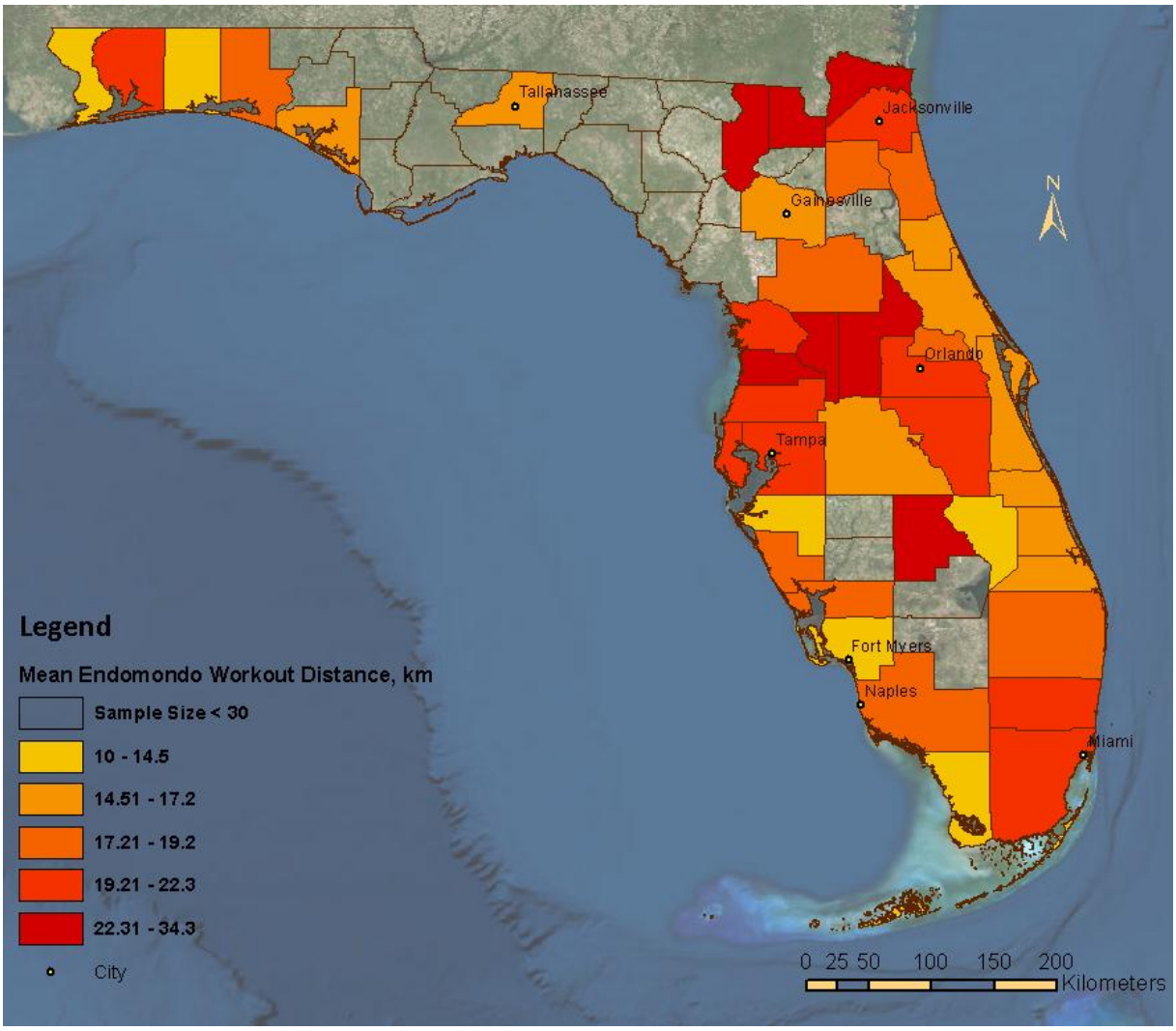


Figure 45. Mean Endomondo Cycling Distance in Florida

To summarize the content of chapter 5, we have compared the two datasets: Strava Metro and Endomondo GPS points with relevant metadata and user data for Miami-Dade County. Though Strava data had more extensive coverage, the Endomondo data allowed to see the bicycle travel patterns from a different viewpoint. With these datasets it is possible to both make assessment of BKT thus revealing routes and areas preferred by cyclists, and also to get a general understanding of demographic characteristics of the local cycling community.

6. DISCUSSION

6.1. Summary of Identified Similarities and Differences in Travel Patterns

All differences in travel patterns can be divided in two big groups: platform dependent differences that show how Strava and Endomondo datasets are different from each other, and differences

identified between certain regions, time intervals or groups of users that are revealed by analysis of data originating from one platform.

Figures 27 and 28 that presented the segment based distribution of Strava activities in Miami-Dade County and the overlapping Endomondo tracks have shown that the overall coverage of Strava data is more extensive than the one of Endomondo. The Kernel Density Estimation maps (Figure 29, 30) showed that Strava has not only the bigger spread, but a substantial quantitative dominance. We have analyzed the trends in Endomondo user counts in four regions of the world. The goal was to see whether there was an overall tendency of growth in user counts and whether this tendency was characteristic for Miami-Dade and Florida. Figure 46 shows the resulting charts.

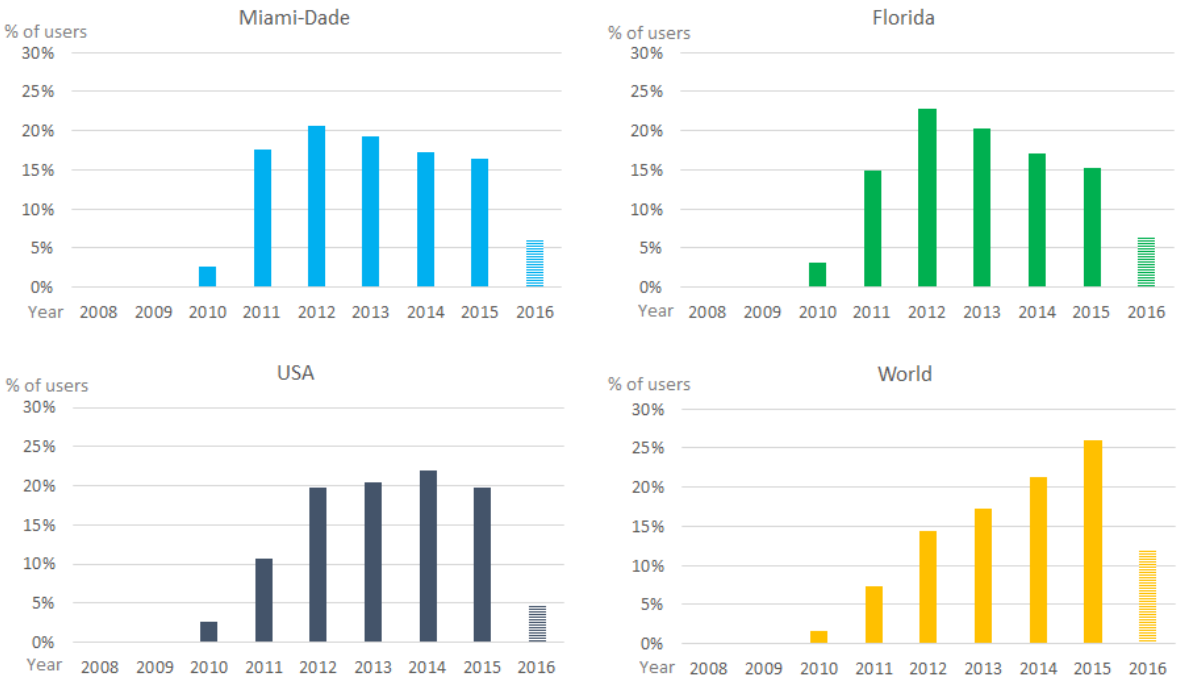


Figure 46. Endomondo User Count Distribution according to the Year of Registration

While there is an overall increase in Endomondo user counts worldwide, this is not the case for the USA and its subregions where Endomondo user counts tend to decline. More than 25% of all Endomondo users in our sample worldwide were registered in 2015. However, this number is less for the USA (under 20%), and even less for Florida (15%). In Miami-Dade 16.4% of users were registered in 2015. Endomondo counts in the USA are already relatively low compared to Strava. Therefore if this declining tendency persists, Endomondo coverage might become insufficient for the travel pattern analysis use.

Kernel Density Estimation maps for Strava and Endomondo on weekday and weekend had some common patterns. On weekend the cycling activities are more focused in certain recreational locations. There are more activities logged along the coastline, on the nearby islands (Virginia Key,

Key Biscayne), in the Everglades National Park, including the Shark Valley (large oval area in the West of the county), and on the roads towards Florida Keys in the South East. The differences in these maps are the activity counts, which are much higher for Strava, as well as some local differences. Endomondo data looks more localized. It contains more weekday activities in rural and recreational areas than Strava does. On the other hand, Strava data is present in the wide areas where Endomondo data is absent (Figure 31, 32). At the level of census tracts, the majority of the areas are equally preferred by users of both fitness apps independent of the day of the week. There is a tendency for Strava athletes to ride more on weekend in the areas that are preferred by Endomondo users on weekday. This implies that the apps might be used for different purposes.

The segment based comparison of route preferences by Strava and Endomondo users is difficult to visually analyze for the scale of a county as a whole. It can be, however, very useful in order to see how presence of certain cycling facilities influences the popularity of the roads. This kind of analysis appeared to be not possible within our study due to the time intensity of Endomondo data retrieval. But the developed procedures and tools simplify this analysis in the future.

We have identified the differences in road usage patterns (Figures 35, 36) according to their functional classes. The dominance of roads of the fifth class in Endomondo tracks can be a result of differences in the underlying road network between Endomondo and Strava. As the majority of the road segments that were added to the Here Navstreets dataset belong to the fifth class, this could shift the proportions of the fifth road class usage in Endomondo. It is possible that Strava athletes log more activities in the recreational areas but Strava Metro data lacks this information since the corresponding geometries are absent in Here Navstreets. In this case Endomondo data can contribute to Strava in a meaningful way.

The temporal differences in the usage of Strava and Endomondo can be partially explained by the fact that Strava promotes logging commute activities. Two weekday peaks of activities in Strava in the morning and in the evening could be attributed to the logged commute trips. Endomondo on the other hand has just one weekday peak in the evening. This, again, suggest that more sport workouts than other types of trips are logged with Endomondo. Considering that the first quarter of the year is characterized by good weather with no extreme heat during the day, we could expect that the cycling activities would be more or less proportionally distributed between 5 AM and 20 PM on weekend. Though the weekend distribution of the activities in Endomondo and Strava datasets are similar, Strava has a more obvious morning peak. A deeper analysis is needed in order to understand the reasons.

Figure 40 demonstrated the distribution of male and female cyclists among 5 age groups. In general the majority of cyclists belong to the category from 25 to 60 years old. The counts of male cyclists between 40 and 60 are high in Miami-Dade compared to other regions (over 40%), and in Florida is a

relatively high count of female cyclists in this age category (about 14%). Florida is also characterized by relatively high counts of cyclist in the category 60+ compared of both genders compared to other regions. There are few cyclists under 25 in the USA including its subregions comparing to worldwide counts.

The distribution of cyclists according to their mean age in Florida counties reveals some interesting patterns. The most counties with big cities (Miami, Orlando, Tampa, and Gainesville) are characterized by the youngest cyclists, slightly over 40. The West of Florida (counties around Naples, Fort Myers) is characterized by a higher mean age of cyclists (over 50). This can be attributed to the fact that these counties are a popular retirement area.

The analysis of mean Endomondo cycling workout distances in Florida also results in an interesting spatial pattern. The areas of the most distant workouts are forests and national parks (Okefenokee National Wildlife Refuge in the Nord-East; Osceola National Forest, John M. Bethea State Forest, Cypress Creek Wildlife Management Area in the North; Apalachicola National Forest in the North-West; Seminole State Forest to the North of Orlando; Lake Apopka to the North-West from Orlando; Chassahowitzka Wildlife Refuge, and Citrus Wildlife Management Area on the West coast of Florida not far from Orlando; Kissimmee Prairie Preserve State Park in the center of Florida).

6.2. Differences in Application of Strava and Endomondo Data for Bicycle Travel Pattern Analysis

We have determined that Endomondo data provides useful information for the assessment of demographic characteristics of the cyclist community. As retrieval of Strava user data is possible to a limited extent, Endomondo data, though less extensive in coverage and less dense than Strava, provides a good additional source of the demographic cyclist data.

Proceeding from the results of our analysis we can conclude that Endomondo data tends to have more recreational character than Strava, and thus its application can be slightly different. Strava Metro data is generally used for the route choice analysis in order to improve the overall cycling infrastructure. Endomondo data in Florida is not extensive enough for this kind of analysis. On the other hand, this data is good for analysis of relative popularity of cycling recreational facilities such as cycling parks. It can also allow to improve the cycling maps my adding shortcuts commonly used by regular local cyclists.

Endomondo data can be applied to explore the fitness and health-related part of cycling workouts as many of the contain sensor data. Additionally there is weather data that can be used to analyze region related cyclists' preferences and limiting weather conditions.

The user data and workout metadata that are present in the Endomondo dataset can be used for various kinds of analysis. Some of the examples are given below.

1. Analysis of correlation between distance and duration of workouts among different age groups cycling with different trip purposes. Goal: to find out whether there is a strong relationship between workout distance and duration, or there are some groups of cyclists with unexpectedly low or high performance.

For this analysis we used the data of worldwide users in each age/trip purpose category. To operate the sample sizes over 30 we had to exclude some of gender/age the groups from analysis.

Trip Purpose and Age	Sample size	Mean Distance, km	Mean Duration, min
Transport, Under 15	33	2,94	13,58
Transport, 15 to 24	310	2,54	9,65
Transport, 25 to 39	1081	3,45	15,13
Transport, 40 to 59	912	3,89	16,34
Transport, 60 and older	166	4,03	15,67
Sport, Under 15	19	3,83	19,41
Sport, 15 to 24	133	3,90	18,80
Sport, 25 to 39	382	4,58	21,96
Sport, 40 to 59	264	5,30	27,69
Sport, 60 and older	56	5,07	23,69
Mountain Biking, 15 to 24	83	4,63	17,44
Mountain Biking, 25 to 39	427	5,56	21,63
Mountain Biking, 40 to 59	2083	5,90	23,73
Mountain Biking, 60 and older	1973	5,61	22,33

Table 7. Mean Distance and Duration of Workouts of Male Endomondo Users

Trip Purpose and Age	Sample size	Mean Distance, km	Mean Duration, min
Transport, 15 to 24	133	1,91	6,49
Transport, 25 to 39	382	2,84	9,77
Transport, 40 to 59	264	3,36	11,18
Sport, Under 15	56	2,94	9,06
Sport, 15 to 24	83	3,52	16,04
Sport, 25 to 39	427	3,17	11,67
Sport, 40 to 59	2083	3,79	15,29
Sport, 60 and older	1973	4,20	16,79
Mountain Biking, 15 to 24	267	4,30	16,09
Mountain Biking, 25 to 39	34	5,05	18,28
Mountain Biking, 40 to 59	104	4,78	17,08

Table 8. Mean Distance and Duration of Workouts of Female Endomondo Users

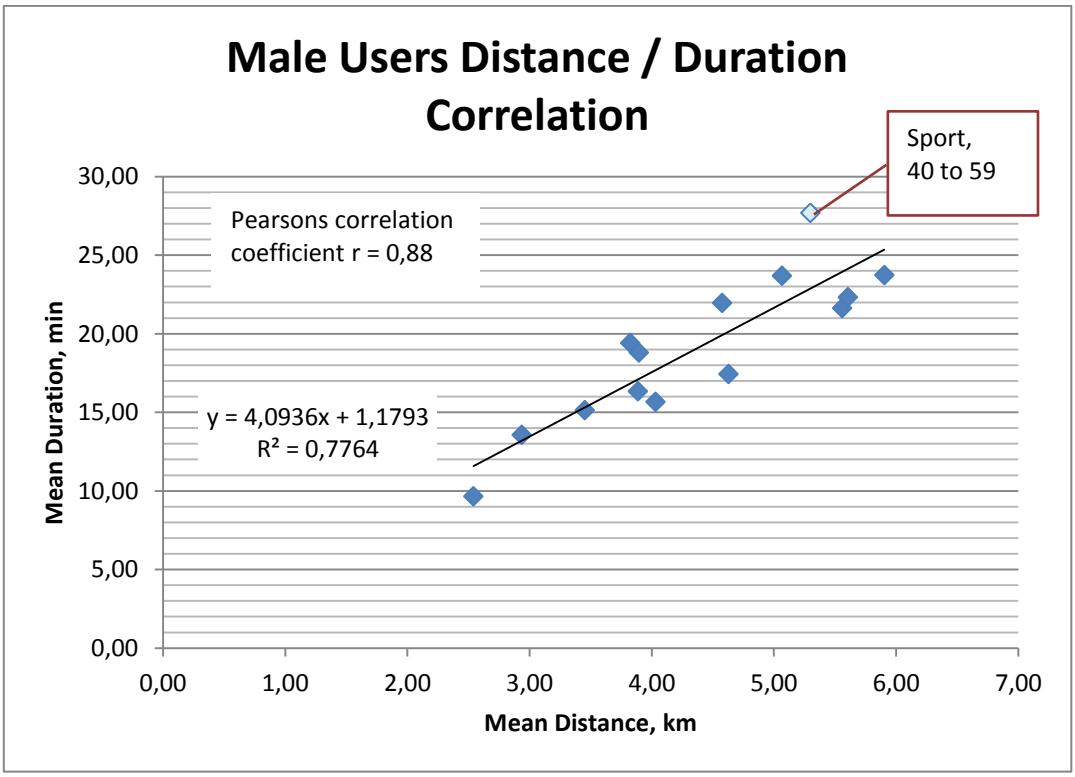


Figure 47. Correlation of Mean Distance and Duration of Workouts of Male Endomondo Users

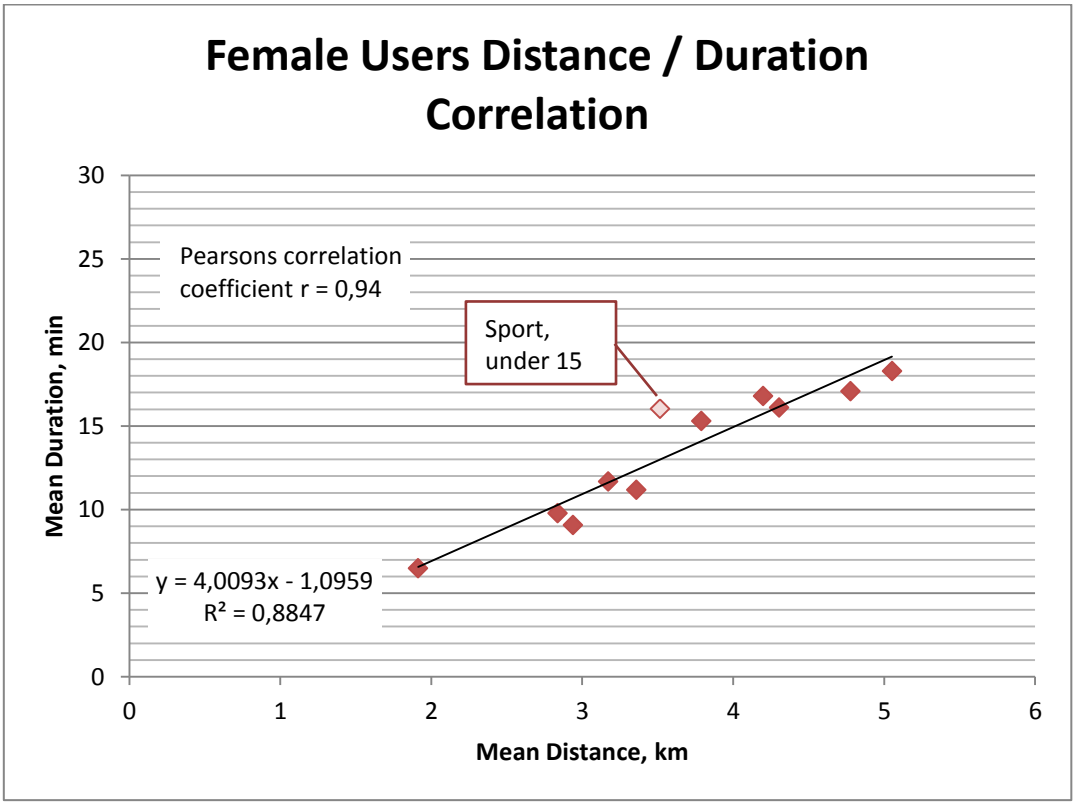


Figure 48. Correlation of Mean Distance and Duration of Workouts of Female Endomondo Users

With the use of the Pearson's correlation coefficient (r = 0,88) and the linear regression (coefficient of determination equal to 0,7764) we determined that there is a strong positive correlation between the workout distance and workout duration of male cyclists (Figure 47). The group of cyclists with relatively low performance in comparison with others is the group of cyclists between 40 and 59 years of age exercising cycling as sport. For the female cyclists (Figure 48) this correlation is even stronger (correlation coefficient r = 0,94, coefficient of determination is 0,8847). The only group of females with relatively low performance are sport female sport cyclists under 15. Sport cycling is more competitive than transport biking, and attracts more people than mountain biking. And young boys tend to be stronger than young girls. These factors can explain the young girls' relatively low performance. The relatively low performance of men between 40 and 59 requires deeper analysis.

2. Analysis of influence of age and gender on the mean workout distance

There are several possibilities of how age and gender factors can influence the workout distance, namely (1) they may have no influence or (2) one of them may influence the distance, and another not, or (3) they can influence the workout distance simultaneously but independently; or (4) there might be an interaction between these two factors. Such an analysis can be done by means of the two-way ANOVA.

Gender		Mean Distance, km	Standard Deviation, km	Sample Size
Female	<15	3,33	2,07	57
	15-25	2,73	2,15	262
	25-40	3,55	2,57	996
	40-60	4,00	2,71	780
	60+	3,66	2,37	111
	Total	3,61	2,58	2206
Male	<15	3,71	2,24	131
	15-25	3,57	2,73	907
	25-40	4,47	3,40	3986
	40-60	5,13	8,62	3916
	60+	4,87	2,90	549
	Total	4,67	6,09	9489

Table 9. Mean Distances of Cycling Activities among Endomondo Users of Different Gender/Age Groups

Table 9 shows the input data that we used to analyze whether age and gender have influence on the mean workout distance. Figure 49 gives a graphical representation of this table. The significance values resulting from the ANOVA test are 0.0001 for gender, 0 for age and 0,867 for their combination. Thus we can conclude that there is a statistically significant effect of gender on travel distance The effect of age is also statistically significant. The combination of gender and age does not

explain the variance on travel distance (p-value = 0.867 > 0.05), which means there is no significant interaction between gender and age factors.

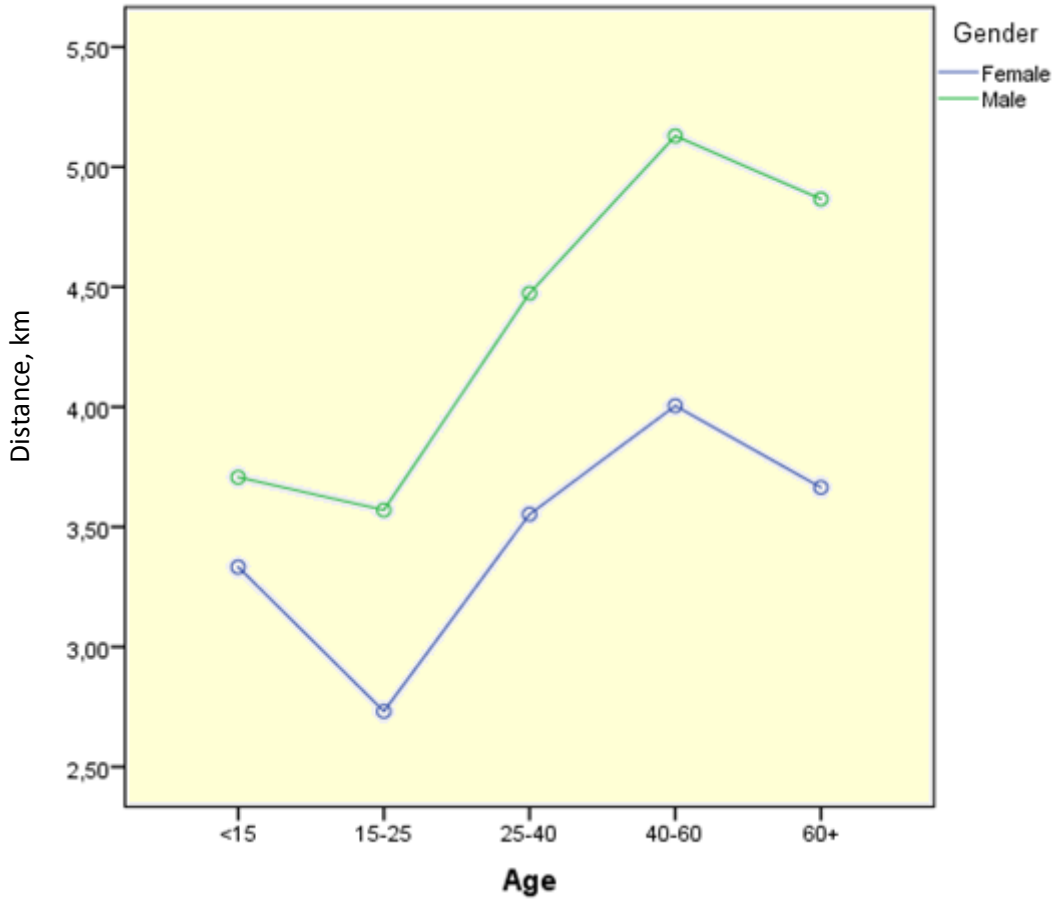


Figure 49. Mean Distances of Workouts According to Gender and Age

The ANOVA test has shown that the effects of age and gender on travel distance are significant. For the gender we have only two groups to compare, so no post hoc tests are needed. To find out which of the five age groups are significantly different from each other we ran the Scheffe's post hoc test. The pairwise comparison of the age groups showed that the following groups differed from each other: (1) 15-25 years and 25-40 years (P-value 0.001), 15-25 years and 40-60 years (P-value 0), 15-25 years and 60+ years (P-value 0.015), 25-40 years and 40-60 years (P-value 0.002).

3. Analysis of Endomondo users' gender distribution according to trip purposes. Goal: to determine whether there is the same percentage of male and female users in all cycling groups (transport, sport, mountain biking), and if there are differences, whether these are statistically significant.

To perform this analysis we used the worldwide Endomondo user counts in each cycling category and performed the Chi-Square test. The distribution of users according to their gender and cycling category e is given by Figure 50. We can visually estimate that the proportion of female cyclists in the

Mountain Biking Category is less than in the Transport category. But is this difference statistically significant? The p-Value associated with the Chi-Square test is equal to 0 < 0.001. There is strong evidence that proportions of cyclists is not even among the cycling categories.

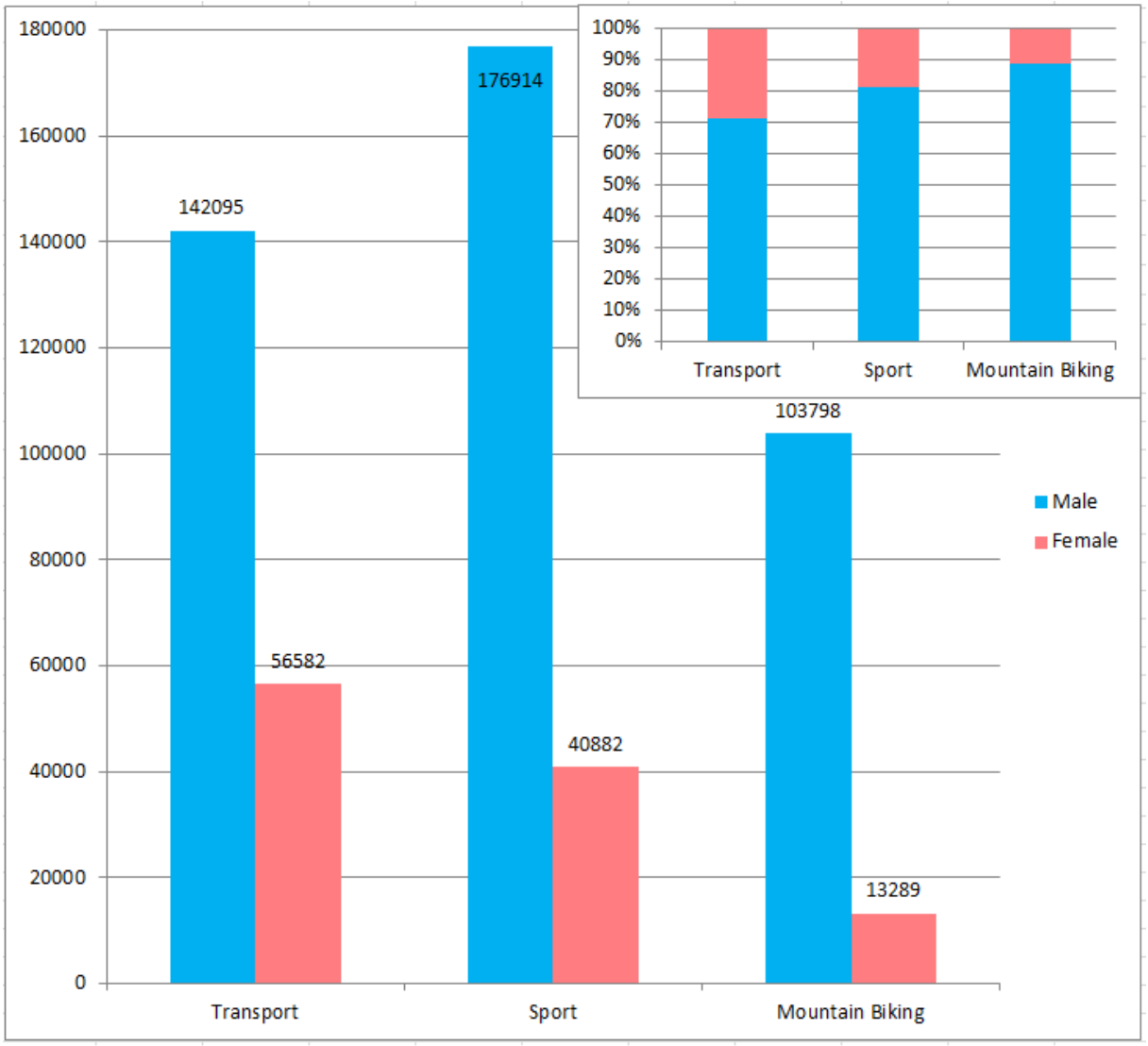


Figure 50. Distribution of Endomondo Users According to Their Gender and the Logged Trip Purpose

As we can see, Endomondo data has a wide range of applications. In regions that have sufficient Endomondo coverage it can be used to analyze the preferred cycling routes and temporal travel patterns. It can be used as a data source complementing other datasets with better coverage in the area if, as in case with Strava Metro data in Florida, it covers small recreational areas and shortcuts absent in Strava. It is a rich additional data source in cases when there exists an interest in the demographic cyclists' data. The next chapter discusses the limitations of the performed study and its results.

7. DATA QUALITY AND LIMITATIONS OF STUDY RESULTS

As with all VGI, GPS tracking data faces certain quality issues. As Strava Metro dataset is delivered to customers in a pre-processed format it is impossible to assess the quality of the underlying point dataset. The Endomondo points on the other hand were available as raw data, and we could detect the some issues that influence the overall quality of the data and conclusions resulting from its analysis. In general we have determined the following limitations:

1. Some of the Endomondo tracks logged as cycling activities definitely do not belong to this category. We have determined that Endomondo dataset contains some tracks located on water in the sea, and some of the tracks represent flight trajectories. Several tracks had the distance of more than 2,000 km and the speed of over 500 kmh.
2. On the other hand, there are Endomondo workouts with unrealistically low speed (less than 2 kmh) and short distances (less than 500 m).
3. In the analysis of Endomondo user demographic data we rely on the assumption that the provided age/gender data is true. However, it is possible that not all users provide truthful demographic data. Thus the results of the analysis based on this data may be faulty.
4. Some of the trajectories do not get logged with the use of Endomondo app but are synchronized with the app by means of Tapiriik (https://tapiriik.com/) or by other means. It can result in doubled counts of synchronized activities if Endomondo is used as an additional data source together with data from another app, such as Strava. The same limitation is valid for Strava as Tapiriik can synchronize data both ways (and also with many other fitness tracking apps).
5. The map-matching algorithm we used matched 90-95% of Endomondo points correctly, and thus there are certain road segments that were falsely assigned to Endomondo tracks, and there are some that were left out by mistake.
6. The quality of GPS signal can attribute to creation of falsely matched segments in areas with natural or man-made obstacles such as high buildings or trees. This complicates the correct map-matching in parks where activities happen under the trees. Some of the Endomondo GPS points in rural areas were located 70 m apart from the closest road. In this case it is possible that data is incorrect, or the activity took place on an off-road terrain. These kinds of activities do not get map-matched.
7. There is a self-selection bias in cycling data that coming from a certain app (Shearmur, 2015), no matter whether it is Endomondo, Strava or another fitness tracker.
8. Within the time frame of this project we have retrieved sufficient Endomondo data for the first quarter 2016. This data is not representative for other times of the year, and thus analysis result cannot be generally extrapolated on cycling travel patterns as a whole.

8. CONCLUSIONS AND FUTURE WORK

Within the scope of this project we compared the suitability of Endomondo and Strava bike tracking data for cycling travel analysis. To achieve this we have fulfilled the following task:

- Development of tools for GPS data retrieval and management
- Endomondo data retrieval
- Design of an entity relationship model and data management in an spatial database system
- Integration of heterogeneous road network data, including HERE NAVSTREETS and OpenStreetMap data, at the geometry and attribute level
- On-screen digitizing of road network data
- Map-matching between GPS tracking points and road segments
- Descriptive methods (e.g. histograms, Kernel Density Estimation maps) to compare coverage and spatio-temporal usage patterns between Strava and Endomondo at the road segment level
- Aggregation of trip characteristics at different levels of U.S. census reference polygons, e.g. cumulative bicycle kilometers traveled per census tracts
- Statistical tests (e.g. two-way ANOVA, chi-square test) to identify significant differences in Endomondo and Strava usage patterns

Some of the research findings are that Endomondo is less used in South Florida than Strava for the analyzed time frame of three months, and hence provides less detail about travel patterns in South Florida. However, Endomondo tracks are observed on many more small roads and off-road tracks (e.g. in parks) than Strava, complementing Strava data in a meaningful way. There were also some differences in the temporal distribution of trips over the day between both platforms, indicating that the platforms are to some extent used for different trip purposes. This research is unique in the sense that it is the first work that compares Strava and Endomondo data at this level of detail.

The analysis results presented in this paper rely on a small part of the retrieved Endomondo data. We have more than a billion GPS points of Endomondo workouts worldwide that can be analyzed in order to compare regional differences in travel patterns and Endomondo app usage. An area of special interest is Thailand. Located not in Europe where high count of activities are expected, Thailand being roughly 3 times the area of Florida about 18 times more workouts (more than 454,000 vs. 27,000 in Florida).

We had no possibility to analyze the influence of changes in cycling infrastructure on bike activity counts due to insufficient data. A possible direction for the future development of this study is to retrieve Endomondo data for at least one year interval and to carry out such an analysis.

Endomondo and Strava GPS tracking data can complement each other for bicycle travel analysis in some regions, as we have seen on the example of Florida. They can be used together as an additional information source providing new insights into bike travel patterns and the peculiarities of the local cycling community.

9. REFERENCES

Barsukov, N. (2014a). Maps of running routes in European cities. Retrieved from http://barsukov.net/visualisation/2014/07/25/endomondo/

Barsukov, N. (2014b). Generating running route maps. Retrieved from http://barsukov.net/programming/2014/07/26/endomondo-code/

Beebom Media. (2016). What is GLONASS And How It Is Different From GPS. Retrieved from https://beebom.com/what-is-glonass-and-how-it-is-different-from-gps/

Benkler, Y., & Nissenbaum, H. (2006). Commons-Based Peer Production and Virtue. *Journal of Political Philosophy*, *14*(4), 394–419.

Bergman, C., & Oksanen, J. (2016). Estimating the Biasing Effect of Behavioural Patterns on Mobile Fitness App Data by Density-Based Clustering. In T. Sarjakoski, M. Y. Santos, & L. T. Sarjakoski (Eds.), *Lecture Notes in Geoinformation and Cartography. Geospatial data in a changing world* (pp. 199–218). New York NY: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-319-33783-8_12

Bierlaire, M., Chen, J., & Newman, J. (2013). A probabilistic map matching method for smartphone GPS data. *Transportation Research Part C: Emerging Technologies*, *26*, 78–98. https://doi.org/10.1016/j.trc.2012.08.001

Bucher, B., Falquet, G., & Metral, C. (2016). Enhancing the management of quality of VGI: contributions from context and task modelling. In C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, & R. Purves (Eds.), *European Handbook of Crowdsourced Geographic Information* (pp. 132–142). London: Ubiquity Press.

Bureau of Economic and Research. (2016). *Florida Estimates of Population: 2016*. University of Florida. Retrieved from College of Liberal Arts and Sciences - Bureau of Economic and Research website: https://www.bebr.ufl.edu/population/data

Capineri, C. (2016). The Nature of Volunteered Geographic Information. In C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, & R. Purves (Eds.), *European Handbook of Crowdsourced Geographic Information* (pp. 15–33). London: Ubiquity Press.

Cintia, P., Pappalardo, L., & Pedreschi, D. (2013). "Engine Matters": A First Large Scale Data Driven Study on Cyclists' Performance. In *13th International Conference on Data Mining Workshops (ICDMW)* (pp. 147–153). https://doi.org/10.1109/ICDMW.2013.41

Cortes, R., Bonnaire, X., Marin, O., & Sens, P. (2015). Stream Processing of Healthcare Sensor Data: Studying User Traces to Identify Challenges from a Big Data Perspective. *Procedia Computer Science*, *52*, 1004–1009. https://doi.org/10.1016/j.procs.2015.05.093

Criscuolo, L., Carrara, P., Bordogna, G., Pepe, M., Zucca, F., Seppi, R.,. . . Rampini, A. (2016). Handling quality in crowdsourced geographic information. In C. Capineri, M. Haklay, H. Huang, V. Antoniou,

J. Kettunen, F. Ostermann, & R. Purves (Eds.), *European Handbook of Crowdsourced Geographic Information* (pp. 57–74). London: Ubiquity Press.

Endomondo. (n.d.). File Export. Retrieved from https://support.endomondo.com/hc/en-us/articles/213219528-File-Export

Endomondo. (2013). The World is Going Green. Retrieved from https://blog.endomondo.com/endomondo-reaches-20000000-users-the-world-is-going-green/

Endomondo. (2015). Making a difference by motivating people to lead more active lives: What We Do. Retrieved from https://www.endomondo.com/about

Fang, Q. (2014). Least Square Analysis on Exercise Duration of Recorded Data from the Endomondo Fitness App. Retrieved from https://pdfs.semanticscholar.org/3622/5c90b27349cd27de3528acabd6b72e7c749a.pdf

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Google. (n.d.). Encoded Polyline Algorithm Format: Google Maps API - Google Developers. Retrieved from https://developers.google.com/maps/documentation/utilities/polylinealgorithm

Griffin, G. P., & Jiao, J. (2015). Where does bicycling for health happen?: Analysing volunteered geographic information through place and plexus. *Journal of Transport & Health*, *2*(2), 238–247. https://doi.org/10.1016/j.jth.2014.12.001

Haworth, J. (2016). Investigating the Potential of Activity Tracking App Data to Estimate Cycle Flows in Urban Areas. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLI-B2*, 515–519. https://doi.org/10.5194/isprsarchives-XLI-B2-515-2016

Hochmair, H. H., Bardin, E., & Ahmouda, A. (2017). Estimating Bicycle Trip Volume From Miami-Dade County From Strava Tracking Data. In *Transportation Research Board 96th Annual Meeting.* Washington DC, United States.

Hood, J., Sall, E., & Charlton, B. (2013). A GPS-based bicycle route choice model for San Francisco, California. *Transportation Letters*, *3*(1), 63–75. https://doi.org/10.3328/TL.2011.03.01.63-75

Hudson, J. G., Duthie, J. C., Rathod, Y. K., Larsen, K. A., & Meyer, J. L. (2012). *Using Smartphones to Collect Bicycle Travel Data in Texas: Final Report*. Texas Transportation Institute.

Jagadeesh, G. R., & Srikanthan, T. (2015). Probabilistic Map Matching of Sparse and Noisy Smartphone Location Data. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC). 15-18 Sept. 2015, Las Palmas, Gran Canaria, Spain* (pp. 812–817). Piscataway, NJ: IEEE. https://doi.org/10.1109/ITSC.2015.137

Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, *52*, 90–97. https://doi.org/10.1016/j.jtrangeo.2016.03.006

Kerr, A. W., Hall, H. K., & Kozub, S. (2002). *Doing statistics with SPSS*. London: SAGE.

Koller, H., Widhalm, P., Dragaschnig, M., & Graser, A. (2015). Fast Hidden Markov Model Map-Matching for Sparse and Noisy Trajectories. In *18th International Conference on Intelligent Transportation Systems* (pp. 2557–2561). https://doi.org/10.1109/ITSC.2015.411

Loidl, M. (2016). A simple map matching approach for bicycle GPS tracks. Retrieved from https://gicycle.wordpress.com/2016/05/13/a-simple-map-matching-approach-for-bicycle-gps-tracks/

Moore, D. S., & McCabe, G. P. (2003). *Introduction to the practice of statistics* (4th ed.). New York, N.Y., Great Britain: W.H. Freeman and Co.

NAVTEQ. (2008). *NAVTEQ's NAVSTREETS Street Data: Reference Manual v3.0*. Chicago, Illinois: NAVTEQ.

Newson, P., & Krumm, J. (2009). Hidden Markov Map Matching Through Noise and Sparseness. In *17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS 2009)* (pp. 336–343). Seattle, WA.

OpenStreetMap. (2017). Downloading Data. Retrieved from http://wiki.openstreetmap.org/wiki/Downloading_data

Osama, A., Sayed, T., & Bigazzi, A. Y. (2017). Models for estimating zone-level bike kilometers traveled using bike network, land use, and road facility variables. *Transportation Research Part A: Policy and Practice*, *96*, 14–28. https://doi.org/10.1016/j.tra.2016.11.016

Peck, R., & Devore, J. L. (2012). *Statistics: The exploration and analysis of data* (7th ed.). Australia, United States: Brooks/Cole Cengage Learning.

Quddus, M. A., Ochieng, W. Y., & Noland, R. B. (2007). Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, *15*(5), 312–328. https://doi.org/10.1016/j.trc.2007.05.002

Rayer, S., & Wang, Y. (2014). Measuring Population Density For Counties In Florida. Retrieved from https://www.bebr.ufl.edu/population/website-article/measuring-population-density-counties-florida

Romanillos, G., Zaltz Austwick, M., Ettema, D., & Kruijf, J. de. (2016). Big Data and Cycling. *Transport Reviews*, *36*(1), 114–133. https://doi.org/10.1080/01441647.2015.1084067

Schweizer, J., & Rupi, F. (2014, May). Map matching and cycling infrastructure analyses with SUMO and python. In *SUMO2014. Modeling Mobility with Open Data* (pp. 195–205). Berlin - Adlershof: Deutsches Zentrum für Luft- und Raumfahrt e.V. - Institut für Verkehrssystemtechnik.

Shearmur, R. (2015). Dazzled by data: Big Data, the census and urban geography. *Urban Geography*, *36*(7), 965–968. https://doi.org/10.1080/02723638.2015.1050922

Sileryte, R., Nourian, P., & Spek, S. v. d. (2016). Modelling Spatial Patterns of Outdoor Physical Activities Using Mobile Sports Tracking Application Data. In T. Sarjakoski & Santos M. Y. (Eds.), *Lecture Notes in Geoinformation and Cartography* (pp. 179–197). Berlin: Springer.

Strava. (n.d.). Strava's V3 API. Retrieved from https://strava.github.io/api/

Strava. (2014a). Strava Metro: Better Data for Better Cities. Retrieved from http://metro.strava.com/, http://metro.strava.com/faq/

Strava. (2014b). Strava Metro: Frequently Asked Questions. Retrieved from http://metro.strava.com/faq/

Strava. (2016a). Strava Labs: Welcome Developers. Retrieved from http://labs.strava.com/developers/

Strava. (2016b). The Year in Stats: Incredible activities and achievements from the Strava community. Retrieved from http://blog.strava.com/2016-stats/

Under Armour. (2015). Under Armour Acquires Endomondo and MyFitnessPal to Establish the World's Largest Digital Health and Fitness Community. Retrieved from http://investor.underarmour.com/releasedetail.cfm?ReleaseID=894685

Velonews. (2012). MapMyRide aims for Strava's KOM. Retrieved from http://www.velonews.com/2012/06/news/mapmyride-aims-for-stravas-kom_223318

Watkins, K., Ammanamanchi, R., LaMondia, J., and Dantec, C. A. L. (2016). *Comparison of Smartphone-based Cyclist GPS Data Sources*. Washington, D.C.

Whitfield, G. P., Ussery, E. N., Riordan, B., & Wendel, A. M. (2016). Association Between User-Generated Commuting Data and Population-Representative Active Commuting Surveillance Data - Four Cities, 2014-2015. *MMWR. Morbidity and mortality weekly report*, *65*(36), 959–962. https://doi.org/10.15585/mmwr.mm6536a4

TABLE OF FIGURES

LIST OF TABLES

Appendix 1. Attributes of Strava Metro Data

Data source: Strava Metro for Florida – Bicycling and Running Data – Edges
Originator: Florida Department of Transportation
Publication Date: November 6, 2015
Contact:
Florida Department of Transportation
Transportation Statistics Office
Chris Francis
850-414-4848

Field Name	Data Source	Field Name	Data Source
EDGE_ID	Strava, LLC	ATHCNT_3	Strava, LLC
ATHCNT_0	Strava, LLC	RATHCNT_3	Strava, LLC
RATHCNT_0	Strava, LLC	ACTCNT_3	Strava, LLC
ACTCNT_0	Strava, LLC	RACTCNT_3	Strava, LLC
RACTCNT_0	Strava, LLC	TATHCNT_3	Strava, LLC
TATHCNT_0	Strava, LLC	TACTCNT_3	Strava, LLC
TACTCNT_0	Strava, LLC	ACTTIME_3	Strava, LLC
ACTTIME_0	Strava, LLC	RACTTIME_3	Strava, LLC
RACTTIME_0	Strava, LLC	CMTCNT_3	Strava, LLC
CMTCNT_0	Strava, LLC	ATHCNT_4	Strava, LLC
ATHCNT_1	Strava, LLC	RATHCNT_4	Strava, LLC
RATHCNT_1	Strava, LLC	ACTCNT_4	Strava, LLC
ACTCNT_1	Strava, LLC	RACTCNT_4	Strava, LLC
RACTCNT_1	Strava, LLC	TATHCNT_4	Strava, LLC
TATHCNT_1	Strava, LLC	TACTCNT_4	Strava, LLC
TACTCNT_1	Strava, LLC	ACTTIME_4	Strava, LLC
ACTTIME_1	Strava, LLC	RACTTIME_4	Strava, LLC
RACTTIME_1	Strava, LLC	CMTCNT_4	Strava, LLC
CMTCNT_1	Strava, LLC	ATHCNT	Strava, LLC
ATHCNT_2	Strava, LLC	RATHCNT	Strava, LLC
RATHCNT_2	Strava, LLC	ACTCNT	Strava, LLC
ACTCNT_2	Strava, LLC	RACTCNT	Strava, LLC
RACTCNT_2	Strava, LLC	TATHCNT	Strava, LLC
TATHCNT_2	Strava, LLC	TACTCNT	Strava, LLC
TACTCNT_2	Strava, LLC	ACTTIME	Strava, LLC
ACTTIME_2	Strava, LLC	RACTTIME	Strava, LLC
RACTTIME_2	Strava, LLC	CMTCNT	Strava, LLC
CMTCNT_2	Strava, LLC	MAPSOURCE	Florida Department of Transportation

ATHCNT_0 (1 / 2 / 3 / 4): Count of unique athletes on the street segment from 12 AM to 4:59 AM (5 AM to 9:59 AM / 10 AM to 2:59 PM / 3 PM to 7:59 PM / 8 PM to 11:59 PM) for the rollup period. This

number represents the number of athletes going with the direction the street segment was digitized. Digitized direction does not correspond with the travel direction of the roadway.

RATHCNT_0 (1 / 2 / 3 / 4): Count of unique athletes on the street segment from 12 AM to 4:59 AM (5 AM to 9:59 AM / 10 AM to 2:59 PM / 3 PM to 7:59 PM / 8 PM to 11:59 PM) for the rollup period. This number represents the number of athletes going against (reverse) the direction the street segment was digitized. Digitized direction does not correspond with the travel direction of the roadway.

ACTCNT_0 (1 / 2 / 3 / 4): Count of trips (regardless of unique athletes) on the street segment from 12 AM to 4:59 AM (5 AM to 9:59 AM / 10 AM to 2:59 PM / 3 PM to 7:59 PM / 8 PM to 11:59 PM)for the rollup period. This number represents the number of trips going with the direction the street segment was digitized. Digitized direction does not correspond with the travel direction of the roadway.

RACTCNT_0 (1 / 2 / 3 / 4): Count of trips (regardless of unique athletes) on the street segment from 12 AM to 4:59 AM (5 AM to 9:59 AM / 10 AM to 2:59 PM / 3 PM to 7:59 PM / 8 PM to 11:59 PM)for the rollup period. This number represents the number of trips going against (reverse) the direction the street segment was digitized. Digitized direction does not correspond with the travel direction of the roadway.

TATHCNT_0 (1 / 2 / 3 / 4): Total number of unique athletes on the street segment regardless of direction of travel from 12 AM to 4:59 AM (5 AM to 9:59 AM / 10 AM to 2:59 PM / 3 PM to 7:59 PM / 8 PM to 11:59 PM)for the rollup period.

TACTCNT_0 (1 / 2 / 3 / 4): Count of trips (regardless of unique athletes) on the street segment regardless of the direction of travel from 12 AM to 4:59 AM (5 AM to 9:59 AM / 10 AM to 2:59 PM / 3 PM to 7:59 PM / 8 PM to 11:59 PM)for the rollup period.

ACTTIME_0 (1 / 2 / 3 / 4): Median time in seconds for all trips on the street segment during the date and from 12 AM to 4:59 AM (5 AM to 9:59 AM / 10 AM to 2:59 PM / 3 PM to 7:59 PM / 8 PM to 11:59 PM)for the rollup period. This number represents the time of cyclists going with the direction the street segment was digitized. Digitized direction does not correspond with the travel direction of the roadway.

RACTTIME_0 (1 / 2 / 3 / 4): Median time in seconds for all trips on the street segment during the date and from 12 AM to 4:59 AM (5 AM to 9:59 AM / 10 AM to 2:59 PM / 3 PM to 7:59 PM / 8 PM to 11:59 PM)for the rollup period. This number represents the time of cyclists going against (reverse) the direction the street segment was digitized. Digitized direction does not correspond with the travel direction of the roadway.

CMTCNT_0 (1 / 2 / 3 / 4): Sum of the commute activities for the street segment summarized from 12 AM to 4:59 AM (5 AM to 9:59 AM / 10 AM to 2:59 PM / 3 PM to 7:59 PM / 8 PM to 11:59 PM)for the rollup period. Commute definition from Strava Metro User Guide “Commuter data is derived by three methods: 1. Commute flag that is native to the Strava experience. 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”

ATHCNT: Count of unique athletes on the street segment for the date specified in the file name. This number represents the number of athletes going with the direction the street segment was digitized. Digitized direction does not correspond with the travel direction of the roadway.

RATHCNT: Count of unique athletes on the street segment for the date specified in the file name. This number represents the number of athletes going against (reverse) the direction the street segment was digitized. Digitized direction does not correspond with the travel direction of the roadway.

ACTCNT: Count of trips (regardless of unique athletes) on the street segment for the date specified in the file name. This number represents the number of trips going with the direction the street segment was digitized. Digitized direction does not correspond with the travel direction of the roadway.

RACTCNT: Count of trips (regardless of unique athletes) on the street segment for the date specified in the file name. This number represents the number of trips going against (reverse) the direction the street segment was digitized. Digitized direction does not correspond with the travel direction of the roadway.

TATHCNT: Total number of unique athletes on the street segment regardless of direction of travel for the date specified in the file name.

TACTCNT: Count of trips (regardless of unique athletes) on the street segment regardless of the direction of travel for the date specified in the file name.

ACTTIME: Median time in seconds for all trips on the street segment for the date specified in the file name. This number represents the time of cyclists going with the direction the street segment was digitized. Digitized direction does not correspond with the travel direction of the roadway.

RACTTIME: Median time in seconds for all trips on the street segment during for the date specified in the file name. This number represents the time of cyclists going against (reverse) the direction the street segment was digitized. Digitized direction does not correspond with the travel direction of the roadway.

CMTCNT: Sum of the commute activities for the street segment summarized for the date specified in the file name. Commute definition from Strava Metro User Guide “Commuter data is derived by three methods: 1. Commute flag that is native to the Strava experience. 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”

MAPSOURCE: This stores the source/provider of the base map information for the line segment on which the data is aggregated.