Visualizing and Analising Urban Leisure Runs by Using Sports Tracking Data

Özgün Balaban\textsuperscript{1}, Bige Tunçer\textsuperscript{2}
\textsuperscript{1,2}Singapore University of Technology and Design, ETH Zurich, Future Cities Laboratory, Singapore-ETH Center (SEC), Singapore
\textsuperscript{1}balaban_ozgun@mymail.sutd.edu.sg \textsuperscript{2}bige_tuncer@sutd.edu.sg

Recently there has been a significant growth on the usage of personal fitness applications running on smart phones or fitness devices. These applications record millions of GPS points generated from the paths of runners. This data can be analyzed to comprehend behavior of runners within a specific location. In this study, using data generated from several sources such as Endomondo and Strava and other complementary data such as climate data, population data etc., we aim to find out the factors affecting running behavior in urban settings. For this purpose, visualizations of running activities are plotted with different variables by using BIG-DID, a software tool we developed as part of this study. Additionally, an evaluation of the tools used or can be used for data analysis and visualizations discussed. Finally, a linear regression model is introduced, which will be further developed in later stages of this study.

Keywords: Big Data, Urban Visualization, Fitness Applications, Leisure Runs

INTRODUCTION

Urban planners benefit from understanding how public spaces are utilized for certain activities, in order to create more lively public spaces. However, relating the activities undertaken in spaces to certain design actions or intent can be a complex task with dozens, if not hundreds, of different factors affecting behavior of people utilizing these spaces. For a better understanding of this behavior urban planners need data of factors affecting it. Until now data required for this task was coming from traditional sources such as questionnaires or interviews (Reades et al. 2007). Planning to promote healthy active city residents is becoming a priority for most of the cities. As the most available sports activity to public is running/walking, it is becoming crucial that we make our cities more running friendly. Various measures are taken by different organizations to motivate people for running, for example, Singapore HealthHub is organizing national steps challenge where users obtain points as they walk/run more (Healthhub 2017). In addition, the current boom in personal fitness applications that track users’ movement with GPS and record them provides a way for users to track their achievements. This data, originally intended for personal use, can be collected massively, and used to help urban planners understand how streets are used for leisure runs. Although there are certain concerns such as the data having selective bias, its huge availability outweighs this disadvantage.

In this study, personal fitness application data is used along with other data such as climate, street
topology, sociodemographic attributes, etc., to create visualizations and statistical analysis to be used for understanding factors affecting running behavior in urban settings. Additionally, an evaluation of the tools used or can be used for data analysis and visualizations discussed. Finally, a linear regression model is introduced, which will be further developed in later stages of this study.

LITERATURE REVIEW
Running/walking is the focus of this study and there are various reasons for this. First, running as an exercise is easy to start, it does not require a special technique, equipment or field. Also, unlike team sports, it can be done solo, which makes it easier to incorporate in a busy life schedule of a modern urban dweller. There is plenty of research about running being helpful in maintaining a healthy life (Hardman et al. 2009). This fact is generally accepted by governmental agencies and municipalities, but even though these want to increase running activity, it is not easy to achieve this as most of the activity is performed in random locations and the spatial needs of running are not fully known (Scheerder and Breedveld 2015).

The effect of perceived built environment on physical activity has been widely studied mainly by using various data sources such as surveys, interviews, governmental data, and more recently by using crowd-sourced data. In a review study, Harris et al. (2013) collected 318 articles in the field (PABE Physical Activity and Built Environment), out of this 191 were about the relationship of built environment and physical activity, 79 were reviews on the past work in the field, 38 were about methodology, 6 were about interventions to increase physical activity and 4 were about other issues. The low number of papers that are about interventions show that there is still not enough knowledge for the urban planners to plan to increase physical activities happening in the urban space. In one such study, Troped et al. (2011) has demonstrated levels of sprawl in US and women’s involvement in physical activity by using a questionnaire. In the same study, they found out access to recreational facilities increases physical activity levels. On the other hand, there was not any statistically significant change with perceived crime and sidewalk presence. Although there is significant amount of research done in the field, there is still not clear how to interpret the findings of these research.

Using personal fitness applications for visualizations and analysis is not new. Almost every personal fitness company that publishes visualizations of certain cities derived from their applications, but these are not meant to be used for analysis as they are not controllable with certain parameters, thus not very useful for urban planners (Strava 2017). Better examples come from transportation planning studies for bicycles. Oksanen et al. (2015) used personal fitness application data for bicyclers for plotting various heatmaps showing the number of cyclers at certain times. Hochmair et al. (2017) used similar data from Strava for estimating the bicycle kilometers travelled per block by using various parameters.

SETUP
Understanding when and where runners are running is crucial to learn about the spatial requirements of running activity. This study takes place in Singapore, and the aim is to understand the behavior of runners in Singapore. Running data is obtained from fitness applications Endomondo and Strava. The technique to retrieve this data by web scraping methods are discussed in a previous study (Balaban and Tunçer 2016). For this study five hundred thousand users and their workouts are scraped from web. From 500000 users, there are 2436 users registered from Singapore and these have completed 44056 workouts of which 38073 are marked as runs by the users.

In addition to the running data, other supplementary data are used in the analysis part, these are: climate data taken from an online weather site that publishes archive of weather data, street topology data retrieved from OpenStreetMaps, census data such as population count, traffic data, and crime data. Every run is divided into certain time intervals (around 5 seconds) and for every time interval a GPS
point is created. From the coordinates of the points, the location of the run is obtained such as the street, district or post code of a nearby location. These are stored in an online MySQL database to be used in the later stage of visualizations and analysis.

After the database is set with the required data, initial step was to decide on the software to make analysis and visualizations to understand spatial requirements of runners. GIS (Geographical Information System) software was a logical choice and QGIS and ARCGIS are tried. ArcGIS comes with an initial challenge as it does not have MySQL connection capability, therefore the data is converted to csv files and imported to ArcGIS, however, this step is time consuming as the data is more than 10M entries and it should be divided into chunks. Both ArcGIS and QGIS have extensive capabilities of displaying geographical data and they provide Python programming options. The second step is statistical analysis, and R and Tableau have been deployed as software options. R is an open source statistical software and it is widely used by the research community whereas Tableau is proprietary software and it is more popular in business analytics. They both provide many options for data analysis, and as their purpose is data analytics, it is easy to connect to different data sources. After experimentation with GIS software and statistical software, the decision was to have a tool that combines both capabilities in one interface which is called BIG-DID see table 1. The reasons for this decision were because he aim is to create a tool for urban designers/planners who will not necessarily have the skills to use either GIS or statistical software. Also, in this way, use of expensive software such as ArcGIS and Tableau is avoided.

BIG-DID runs on a PHP server and it is built by using Laravel framework. It is connected to the MySQL database that holds all the data. When the client starts the system and selects the required parameters, the result of the query is passed to the client. At this stage, certain JavaScript tools are used for creating visualizations, D3 for statistical graphs and Leaflet for map based visualizations. Similarly, all static heatmaps showing runs per kilometer per street are created. This visualization has the same filtering capability so that planners can select certain time frames, climate conditions, etc. Heatmap capability is introduced with the help of Leaflet plugin Leaflet/heatmap.

Table 1
Software used in this research

<table>
<thead>
<tr>
<th>Software</th>
<th>Type</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArcGIS</td>
<td>GIS</td>
<td>Easy to create</td>
<td>No MySQL support</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Easy to create geographic visualizations</td>
<td>Hard to include functionality</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Web-based capability</td>
<td>Expensive</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MySQL support</td>
<td></td>
</tr>
<tr>
<td>QGIS</td>
<td>GIS</td>
<td>Free</td>
<td>Hard to include functionality</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Easy to create geographic visualizations</td>
<td>Certain things are buggy/hard to work</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comes up with tools</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>No web-based capability</td>
<td></td>
</tr>
<tr>
<td>BIG-DID</td>
<td>Software</td>
<td>Web-based Interface is shaped by the programmer</td>
<td>Hard to setup initially, needs programming skills to setup</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Web-based</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Statistical</td>
<td>Easy to create statistical graphs</td>
<td>Very slow working with database</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No web-based capability</td>
<td>(10M data points)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Easy to create statistical graphs</td>
<td>Requires programming</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Web based capability (R Shiny)</td>
<td>Can be hard to setup (R Shiny)</td>
</tr>
</tbody>
</table>

Figure 1
Weekday - weekend run statistics
OBSERVATIONS

For urban planners, it is crucial to know when and where do people run, to better understand the spatial requirements of running. Running behavior changes with various conditions; working individuals tend to run either after work hours or during weekends. However, for every city this situation can change, in some parts of the world it might be unsafe when it is dark or it might be too cold in the winter. Therefore, spatial requirements of running also change from city to city. This research focuses on Singapore.

Singapore is a country/city with a tropical climate with a warm temperature throughout the year. There are two periods of rains which occur from mid-November to early March and June to September. During these months, there are heavy rains occurring daily. Also, it is one of the safest cities in the world. Therefore, we expect more runs happening after the sun sets and during weekends.

First analysis is about weekday and weekend runs see figures 1, 2 and 3. As expected, it is observed that on weekends the average running time and distance increases along with the times of runs that are performed. Also, throughout the weekdays after Monday, the number of runs decline until Saturday. Observations of figure 2 and figure 3 reveal the locations of runs that happen weekdays and weekends, respectively. Although these look similar, there are still some differences to observe. In both periods, parks such as East Coast Park (ECP), Bedok Reservoir, etc., are used extensively along with the streets near public housing estates. However, during weekdays, there are more differences in running locations then in weekends, and as the average running time and distance are longer, it is thought that more people go to parks to do longer runs during weekends.

Second analysis is about the time of the day when people run. As Singapore is a tropical city that is warm throughout the year, it is not surprising to see in Figure 4 that people avoid times with strong sunlight, and the most preferred time of the day for runs are before and just after sunset and during sunrise. A significant number of runs performed in Singapore are done without sunlight, hence lighting streets becomes crucial for safety of runners. Figure 5 shows that the situation does not change within the year. Figures 6 and 7 show the runs that are performed during day time and night time. The difference can be observed around public housing estates. There are more sprawling runs whereas night runs are generally concentrated in certain areas, which can be ex-

Figure 2
Weekday runs
plained as people running in the daylight can change their routes whereas people running at night prefer routine.

Singapore, having a tropical climate, has a steady number of runners throughout the year, but this can change in extreme weather conditions. In certain years, forest fires in neighboring Indonesia brings particles to Singapore which is called haze. During those periods, it is difficult to breathe and the government advises people to stay indoors. One example of this occurred in September-October 2015 with maximum PSI level reaching 302 (very unhealthy). During that period, runners avoided days with high particle levels but they run in the days just after haze dissipates. Similarly, during rainy season at certain days Singapore receives more than 100mm of rain. During those periods running patterns change, runners avoid paths that can be slippery or that puddles can form on.

For enabling urban planners to see patterns of running behavior at some parts of city, a dynamic visualization interface is created that shows runners running within the selected part of the city on top of a map in a dynamic timeframe. Users can select the time frame: certain parts of the day, day or night, weekdays and weekends, etc. Users can also select certain demographics such as male, female, or certain age groups, junior, adult, senior, etc. With this capability, planners can see good performing streets which are running friendly and compare these with less running friendly streets.

Last visualization is for street catchment. This shows a catchment visualization for any street in the city that is listed. For this purpose, every point of all runs that use a certain street is drawn on a map, points that have more runs are colored differently denoting the intensity. The aim of this visualization is to show a street’s ability to attract runners, and from
what distance. Also, it shows the connectivity of that location with the rest of the surrounding. In Figure 8, we can observe Bedok Reservoir, which is a lake/park around a residential area, and is used for running frequently. In this plot, the entry points of runners, and where they start and end their runs can be observed.

These visualizations are great for giving an overview of a street’s situation in accommodating runs. While an expert planner might be able to guess why a certain street is not running friendly, these tools do not provide reasons for why they are less running friendly. To be able to understand the reasons, statistical analysis of the streets is described in the next section.

**ANALYTICS**

Visual presentation of data regarding a phenomenon to planners/designers is a powerful way of displaying how that phenomenon happens. However, although visualizations are good for showing what is happening overall, it is difficult to isolate a parameter’s influence on the overall behavior. For this purpose, planners need to include analytics of the data to determine a parameter’s effect on the behavior.

For runs happening in the urban environment, there are four categories of variables, these are: street networks (N), location (L), sociodemographic (S), and climatic (C) variables. The representation can be given as:

\[ R = f(N, L, S, C) \]  

In this equation, R is runs in a street, N is network characteristic of a street (such as typology), L is location characteristics (such as pedestrian walk width), S is sociodemographic characteristics (such as crime rate), C is climate characteristics (such as raining or not).

Network variables include topology, hierarchy, morphology and scale (Hochmair 2017). Topology is the connectivity of the streets to each other, hierarchy is measure of importance of a street such as a main road, morphology shows the “shape and fragmentation”, and scale is the supply.

Location variables include location specific variables such as pedestrian walk width, presence of a covered walkway, presence of a main road, etc.

Climatic variables affect users’ participation in leisure runs. Seasons create peak months for running which are generally summer months in European countries, whereas tropical countries are bounded by rainy seasons such as monsoon. Therefore, the variables affecting the regression model are: temperature, uv, wind speed, rain fall and air pollution.

Social variables include crime rate, gender, age, etc.

With this model at hand, we tried several variables such as weekdays, weather and time. As discussed in the previous section, weekdays affect the count of runs. It was possible to derive a linear regression model for the days of the week with a certainty of around 80 percent. Also, haze and rain have certain effects on the number of runs happening. These variables are added to the linear regression model. However, the model discussed here needs more data and development, and will be described in the later stages of the research. The final model will be important for an urban planner for creating guidelines for running friendly streets.

**CONCLUSION**

This study uses personal fitness data for creating a toolbox for an urban planner in making streets more running friendly. For this purpose, several visualization and analysis tools are created. This toolbox will
Figure 6
Runs with daylight
7am-7pm

Figure 7
Runs in the dark
7pm-7am
help urban planners in making analysis of streets for leisure runs more easy. Although personal fitness data carries a selective bias, it should be noted that it is getting more and more popular, so the bias is getting less erroneous. Also, the number of valid observations these tools give to the urban designer makes it very useful. Therefore, a careful usage of these tools will be very valuable in urban planning process.

ACKNOWLEDGEMENTS
The research was conducted at the Future Cities Laboratory at the Singapore-ETH Centre, which was established collaboratively between ETH Zurich and Singapore's National Research Foundation (FI 370074016) under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. In addition, it is part of PhD study conducted by the author at Architecture and Sustainable Design Pillar at Singapore University of Technology and Design.

REFERENCES


Harris, JK, Lecy, J, Hipp, JA, Brownson, RC and Parra, DC 2013, 'Mapping the development of research on physical activity and the built environment', *Preventive Medicine*, 57(5), pp. 533-540

Hochmair, HH, Bardin, E and Ahmouda, A 2017 'Estimating Bicycle Trip Volume for Miami-Dade County from Strava Tracking Data', *Transportation Research Board 96th Annual Meeting*, pp. 1-17

Oksanen, J, Bergman, C, Sainio, J and Westerholm, J 2015, 'Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data', *Journal of Transport Geography*, 48, pp. 135-144


