Uber Data Analysis

# Qaseem Raza

*Superior university Gold Campus BSEM-F17-137* bsemf17137@superior.edu.pk

# 

***Abstract*—Urban liveability is a key concept in the New Urban Agenda (NUA) adopted by the United Nations (UN) in 2016. The UN has recognized that effective benchmarks and monitoring mechanisms are essential for the successful implementation of the NUA. However, the timely and cost effective collection of objective international quality of life urban data remains a significant challenge. Urban liveability indexes are often complex, resource intensive and time consuming to collect, and as a result costly. At the same time, competing methodologies and agendas may result in subjective or non-comparable data. Historically, transit has been a central organizing factor around which communities have been built. This paper explores the use of Uber data as a simple real-time indicator of urban liveability. Using data from the Uber Ride Request (URR) API for the Brazilian city of Natal, our preliminary findings suggest that Uber Estimated Time to Arrive (ETA) data is strongly correlated with selected quality of life indicators at a neighborhood and region level. Furthermore, unlike other urban liveability indicators, our findings suggest that Uber ETA data is context-sensitive reflecting daily and seasonal factors thereby providing more granular insights. This preliminary study finds strong evidence that Uber data can provide a simple, comparable, low cost, international urban liveability indicator at both city and neighborhood level for urban policy setting and planning.**

***Index Terms*—Uber, Data science, Urban liveability indicators**

1. **INTRODUCTION**

For nearly five decades, liveability has been referenced as a key attribute for community and urban planning worldwide. More recently, it has been firmly placed in the global policy lexicon by its inclusion in three of the principles and commitments of the New Urban Agenda (NUA) adopted by the UN in 2016. The NUA is notable as it represents a significant international policy commitment in support of the Sustainable Development Goals (SDG), and more specifically SDG11, and what some have referred to as a pro-urban future . SDG11 sets out a goal for the international community to “make cities inclusive, safe, resilient and sustainable”. While it is clear that the authors of the NUA perceived

liveability as playing a role in eradicating poverty (Paragraph 14a), and as an indicator of both social inclusion and cohesion (Paragraph 40) and sustainable urban transport and transit systems (Paragraph 114) no where within the NUA or supporting documents the concept of liveability is defined. This is not entirely surprising. Indeed, authors have commented on the widespread use of the term, despite the ambiguity in meaning in policy documents and scholarly articles.

According to Newton liveability can be defined as a set of attributes of a place, encompassing housing, neighborhood and region aspects that contribute to residents’ quality of life and well-being. A recent review of the literature on relevant indicators of liveability suggests a broad range of contributory indicators across 11 policy domains (the natural environment, crime and safety, education, employment and income, health and social services, housing, leisure and culture, food and other goods, public open space, transport, social cohesion and local democracy), although the relative importance of each is unclear. Ruth and Franklin suggest that a “livable city” requires the needs of the inhabitants of the city to be aligned with “*built infrastructures and ecosystems that provide the goods and services on which lives and livelihoods in the city depend.*” They note that it is difficult to arrive at a gener- ally acceptable definition of liveability because globalization, urbanization, new technologies and environmental constraints are impacting the expectations of the inhabitants. Notwith- standing the definitional ambiguity of the term liveability, this tension between the dynamic subjective perception of the inhabitant versus the objective reality of city infrastructure presents significant measurement challenges.

Policymakers, industry and scholars have all proposed a number of potential indicators to measure liveability for a variety of purposes including discrimination among com- peting hypotheses, structuring the understanding of issues and conceptualizing solutions, tracking performance toward

goals and objectives, discriminating among alternative policies either for specific decisions or general policy directions, and informing general users (public, stakeholders, community) [9]. Indicators can vary widely in terms of geographic granularity (e.g. international, country, city, neighborhood), comprehen- sive vs singular perspectives (e.g. transport, health, population cohorts), measure characteristics (e.g. objective/quantitative vs subjective/qualitative), and stakeholder (e.g. individual, industry, local government, national government).

As a result of this variability, liveability indices are often difficult to compare [1], [10]. Due to the scale and complexity of liveability measurement, it is often resource-intensive, time- consuming, and costly. Unsurprisingly, there has been calls for standardization, disaggregation, accurate and timely infor- mation on both liveability and sustainability.Historically, transit has been a central organising factor around which communities have been built [11]. For over six decades, commentators have consistently noted the role of transit systems and traffic patterns in livable communities [12]. It is unsurprising therefore that transport and transit systems features heavily and widely referenced liveability indicators and is specifically mentioned in the NUA with regards to livability.

This research explores the use of Uber Estimated Time to Arrive (ETA) data as a simple real-time indicator of urban liveability. Uber is a ride-hailing service that offers peer-to-peer ride-sharing and other services to 75 million passengers in over 600 cities worldwide [14]. Due to the nature, scale and coverage of Uber’s operations, it provides a unique real-time objective source of data on the interaction between inhabitants of a city and its infrastructure, and in particular its transport infrastructure. It enables comparison of data at multiple levels, for example cities, districts and neighborhoods, but also provides context sensitive data provid- ing insights in to the impact of other factors impacting Uber drivers and trips on a daily basis including traffic incidents, weather and other events. As such, we posit that Uber data may provide a simple, rapid, low cost, time- and context- sensitive indicator of urban liveability. We illustrate this with data sourced from the Uber Ride Request (URR) API for the Brazilian city of Natal. To demonstrate the viability of this premise, Uber data is explored by means of a data-driven approach mainly founded on Exploratory Data Analysis (EDA). In short, the main objective of this study is to show that Uber service inherently reacts to city features and dynamics, hence its data can be used as source for a new liveability indicator.

The remainder of this paper is organized as follows. The following section discusses prominent urban liveability indi- cators and extant scholarly research on Uber and using the Uber Application Programming Interface (API). Section III introduces the empirical context and an outline of the data set used in the study. Section IV presents the methodologies and preliminary findings from a basic EDA on Uber ETA data as an urban liveability indicator. This is illustrated using data from the Brazilian city of Natal and comparing the results with extant liveability research on Natal based on the Urban Life

Quality Indices (ULQI). Section V discusses the contribution of the study, limitations, and avenues for future research. The paper then concludes.

1. **RELATED WORKS**

Two distinct sources of related works are of interest to this study - publications related to urban liveability indicators 1 and those related to using Uber data. The Economist Intelligence Unit (EIU) Global Liveability Index and the Mercer Quality of Living Ranking are two indices referenced widely in policy, media and academic literature. The EIU Global Liveability Index is an annual rating of 140 cities for relative comfort based on 30 qualitative and quantitative factors across five broad weighted categories (stability, healthcare, culture and environment, education, and infrastructure) constructed using a combination of external data points and the judgment of a group of in-house and external analysts [2]. It is primarily used for employee mobility. Similarly, the Mercer Quality of Life Ranking evaluates living conditions in 450+ cities worldwide based on 39 factors, grouped in 10 categories

- political and social environment, economic environment, socio-cultural environment, medical and health considerations, schools and education, public services and transportation, recreation, consumer goods, housing and natural environment. Scores are weighted to reflect their importance to expatriates. Like [2], the primary focus is to support decisions in relation to employee mobility. It should be noted that Mercer do also offer services to municipalities to assess factors that can improve their quality of living ranking [13].

Recognizing the need for standardization of city indicators, the World Bank initiated the Global City Indicators Program (GCIP) in 2006, which is comprised, at the time of writing, of 27 core and 36 supporting indicators [19]. Unlike, the purpose of the GCIP indices is to inform policy making and urban planning. The Global City Institute Facility (GCIF) manages the GCIP dataset and claims to have data from 255 member cities from 82 countries. They go on to note that the GCIP was the framework for ISO 37120, the first international standard on city metrics2. Cities report indicators annually and the benchmark data is provided by the GCIF including verification services.

More recently, the Global Liveable Cities Index (GLCI) has been proposed [20]. The GLCI comprises five categories of in- dicators e.g. economic vibrancy and competitiveness, environ- mental friendliness and sustainability, domestic security and stability, socio-cultural conditions, and political governance. The GLCI places more emphasis on governance than, for example, the GCIP and therefore in their reported analysis of 64 cities, findings are significantly different. Notably, the GLCI seeks to further categorize cities in terms of their attractiveness to different personality types and in this way, it integrates a citizen-centric approach however the scientific rigour behind this classification is lacking in detail.

**Date-Time Day Shift Rusha Neighborhood Region ETAXb ETASc**

18-02-09 09:30 Morning False Pitimbu South 276.0 324.0

18-02-11 14:00 Afternoon True Petro´polis East 222.0 288.0

18-02-01 16:50 Afternoon False Areia Preta East 138.0 234.0

18-02-21 02:50 Late night False Ma˜e Luiza East 684.0 888.0

18-02-03 20:20 Evening False Lagoa Seca East 138.0 156.0

18-02-23 14:50 Afternoon False Capim Macio South 138.0 174.0

18-02-11 08:30 Morning True Lagoa Nova South 198.0 818.0

18-02-03 15:20 Afternoon False Ponta Negra South 138.0 228.0

18-02-07 20:20 Evening False Guarapes West 396.0 630.0

18-02-02 15:40 Afternoon False Ponta Negra South 138.0 276.0

Of specific relevance to this paper, the recent works by Arau´jo and Caˆndido [15], [16] outline the development of ULQI and its application to Natal and its districts. Unlike the indices above, the ULQI does not use any subjective data and is based on four variables (Urban Environmental Infras- tructure, Urban Equipment and Services, Socio-economic, and Safety) comprising 23 quantitative indicators [16].

There are a number of observations to note with regards to the indices above. Firstly, they all comprise indicator sets. As such, they can be complex and difficult to interpret [21]. Secondly, they are all either annual or ad hoc (in the case of GLCI and ULQI) and therefore represent an assessment at a single point in time and are not sensitive to temporal variations. Thirdly, with the exception of ULQI, they com- prise subjective and objective data and weighting. While [20] attempts comparison, it is unclear whether this is methodolog- ically sound. Fourthly, for international comparisons, they all require significant time, effort and funding to collect data.

Regarding research using Uber data, no scientific research on urban liveability using Uber data was identified. This is not surprising as the Uber API has only be available since 2016 and its use for liveability research is relatively novel. The Transportation Sustainability Research Center at the University of California, Berkeley has published a series of papers on the impact of on-demand ride services and its role and impact on urban transportation but not as an indicator for liveability per se [22], [23]. Similarly, there is a large number of articles on the impact of Uber on competition, the workforce and the need for regulation (see, for example, [24]–[26]), however these are not particularly relevant to our research topic. Research studies using the Uber API typically fall in to a small number of topic clusters, namely understanding pricing phenomena [27], opti- mizing itineraries and payoffs [28], consumer research [29], competitiveness research [30], and to build new products and services.

1. **EMPIRICAL CONTEXT**

The empirical context for this study is characterized by

1. the city of Natal in Brazil and (ii) the Uber ride-sharing service. Natal is the capital city of the Brazilian state of Rio Grande do Norte with an estimated population of 885,180 inhabitants [32]. The city has a total area of 170 square kilometres (66 sq miles) and is located on the Atlantic Ocean. The city can be divided in to four administrative regions based on cardinal points (North, South, East, and West) which vary significantly in population and socio-economic terms. The northern and western regions of the city have the largest resident population (and highest population growth rates) but also are characterized by low income levels per capita [32]. The eastern region is the city center with older neighborhoods and is characterized by both high population density and growth while the southern region is a recent addition to the city and contains many of the newer, more modern and up-market restaurants, hotels and associated social infrastructure [32]. Given its location, tourism is a key economic activity for Natal. Natal is a suitable empirical context for this study as it has an



active Uber market resulting from its tourism activities, has neighborhoods with clearly defined socio-economic and pop- ulation differences, and is a relatively small city. Furthermore, as stated earlier, it has been the focus of a pre-existing study on urban livability [15], [16] and therefore has benchmark data with which to compare our proposed indicator.

As discussed earlier, Uber is a ride-hailing service that offers peer-to-peer ride-sharing, taxi cab hailing, food delivery, bicycle-sharing and other services to 3 million drivers and 75 million passengers in over 600 cities worldwide [14]. Over 15 million trips are completed on Uber each day with data on each booking and trip recorded by Uber [14]. In this way, it is a two- sided market in which the Uber platform must intermediate to calibrate supply and demand, while ensuring relatively high satisfaction to both sides [33]. Even though Uber can be used from a tablet or desktop web-interface, it is overwhelmingly used via a Smartphone. Uber collects a vast amount of data, on both riders and drivers, and provides selected access to this data through APIs and Software Development Kits (SDKs) available through its developer website3. One such developer service is the Uber Ride Request (URR) API. The URR provides access to a number of core capabilities of the Uber mobile application, namely service selection (e.g., *Uber X*, *Uber Select*, *Uber Black*, *Uber Pool*, etc.), pick-up and drop-off location specification, estimated time to arrive and estimated price data, and ride requests [34].

For the purposes of this paper, the URR API was used to retrieve Estimated Time to Arrive (ETA) data (the time in seconds between a request to the Uber system for a driver and the corresponding pick-up time) for a selection of different neighborhoods in Natal. For each selected neighborhood, 10 pick-up locations were defined according to the following criteria:

* + A fixed set of five points of interest (POIs), such as schools, squares, parks, shopping malls, public buildings, etc.;
  + A set of four random land coordinates for which URR API responds (rebuilt for each collection iteration);
  + The geographic centroid of the neighborhood.

The *Time Estimates* endpoint of the URR API returns ETAs for currently available services at a given location. In this case, two Uber services were available in Natal - *Uber X* and *Uber Select*. The ETA for each service is expressed as integers in seconds. Figure 1 illustrates query parameters and response of URR API.

Target data was collected for February 2018. The data collection procedure comprised a coordinate selection stage followed by a multi-threaded data retrieval routine running continuously on a dedicated infrastructure during the data collection period. Every 10 minutes, the mean ETA was calculated from collected data for each neighborhood and results were appended to a time series data set. Figure 2 illustrates the data collection and consolidation process. The resulting data set comprises a 10-minute resolution time series.

TABLE I

SAMPLE UBER ETA DATA FOR NATAL - FEBRUARY 2018

of mean ETA values (for *Uber X* and *Uber Select*) for all neighborhoods in the city totalling 145,152 entries. Table I presents a sample data with temporal augmentations.

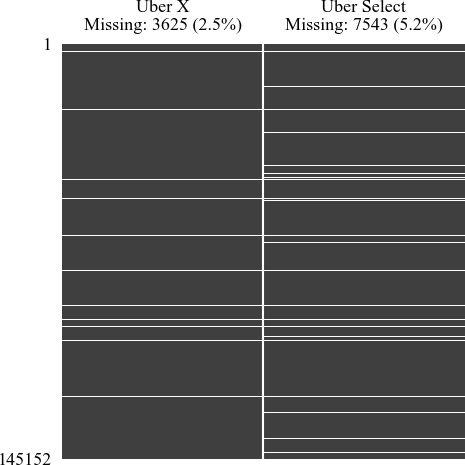
EXPLORATORY DATA ANALYSIS OF UBER ETA DATA Given the preliminary and exploratory nature of this paper,

Exploratory Data Analysis (EDA) was deemed an appropriate initial analytical technique for data driven discovery [35]. EDA has been found to be useful for identifying patterns, trends, correlations or relations among the data to generate insights or hypotheses [36], checking the data sanity (making sure that the data is on the expected scale and format), and identifying missing data, outliers and anomalies [37]. A key principle of EDA is to “*try to let the data speak for themselves*” [38]. Hence, a key component in EDA is the employment of various graphical methods to present data analysis results in potentially intuitive ways in order that different representations of the data will lead to additional indications about ‘models’ of relationships not expected a priori.

In this study, a basic EDA was structured to analyse Uber ETA on a neighborhood basis for Natal. It encompassed data quality, descriptive analytics, time-series and spatial analyses

**Data Quality Analysis**

A critical stage in data analysis is to identify and character- ize missing data in the target data set. By identifying missing data and taking it in to account, awareness of data quality is- sues can guide the data handling strategy for further and deeper analyses. A missing data accounting and visualization toolset



5000

region

East North South West

4000

Count of missing Uber ETA entries

3000

2000

1000

Uber X

0

sub-sampling) for further analysis is not a strong requirement. Missing data can result from a variety of factors such as communication problems, intermittent API response, and ser- vice unavailability. Our analysis suggests that missing entries are related to a lack of response from the URR API for parts of the day for certain pick-up locations. Figure 4 shows that missing entries for both Uber service variants occurred typically late at night and for pick-up locations in the western and northern regions of the city. Such a data absence pattern suggests that there are potential coverage black spots in offered services. Addressing such black spots may improve service

behaviour.

1. *Preliminary Mean ETA Analysis by neighborhood*

To understand high level differences between regions and Uber variants, the mean ETAs by service variant and neigh- borhood was calculated and visualized. Figure 5 shows a grouped bar chart representing mean ETAs for both Uber variants for selected neighborhoods, for both normal and rush hour periods. For these neighborhoods, the chart highlights an overall consistent mean ETA difference between Uber variants, while only a small deviation in mean ETAs between normal and rush hour periods. This suggests that choice of service variant may be more impacting in ETA than the choice of taking service at normal or in rush hours.

comprising the Python language libraries *missingno* [41] and *matplotlib* [40] was used to provide a quick visual summary of the completeness (or lack thereof) of the target data set. Figure 3 shows a nullity matrix manifesting missing entry count and dispersion over ETA data.

Mean ETA data was plotted (i) hourly by Uber service variant (see Figure 6, and (ii) daily by both service variant and region for the month of February 2018 (see Figure 7). The former illustrates, as expected, an increase in mean ETAs during rush hour periods that explained by increased traffic, traffic jams and demand for the Uber Service. Also as expected, a great increase in ETA is observed late in the night, probably due to a lower demand or low availability of drivers. The latter month-wide analysis is more insightful. The clustered and distinct curves in Figure 7 reveals that different city regions have a different Uber ETA profile. Visual peak detection analysis clearly identifies numerous peaks that cur- sory desk research can reveal to be context-sensitive. They are related to (i) city-wide events, in this case, Carnival, and

1. weather events e.g. Natal experienced a very rainy period in February 2018. Increased ETAs during Carnival may a result of demand surges from tourists or lower driver supply during the holiday period. Increased ETAs during rainy days may indicate infrastructural or public transit problems in the city. This analysis in particular highlights both the potential but limits of Uber data for urban livability research; it has
2. **CONCLUSIONS**

Global policymakers are emphasizing the need to measure, monitor and respond to factors impacting liveability in urban settings. Extant urban livability indicators are used for a variety of purposes not least community and urban planning, stakeholder engagement, and labor mobility and workforce planning. Existing measurements of liveability rely on method- ologies require time-consuming, resource-intensive, and costly data collection and are often complex for users to understand and use.

This paper presented a first study in using Uber ETA data as a simple, low cost, and international urban liveability indicator. We illustrate its use in the case of Natal and compared it to prior urban livability research based on the ULQI [16]. Our findings presented compelling evidence that Uber data can act as a simple real-world measurement of urban liveability. Furthermore, our analysis suggests that Uber provides more granular data providing context-sensitive insights on how events, weather and other activities impact a city and its communities.

**ACKNOWLEDGMENT**

This work was partly funded by the World Technology Universities Network (WTUN), the Irish Centre for Cloud Computing and Commerce, an Enterprise Ireland/IDA tech- nology centre, and the Post-graduate Program in Electrical and Computer Engineering (PPgEEC) at the Federal University of Rio Grande do Norte (UFRN) at Natal/Brazil.

REFERENCES

[1]M. Lowe, C. Whitzman, H. Badland, M. Davern, L. Aye, D. Hes,

* 1. Butterworth, and B. Giles-Corti, “Planning healthy, liveable and sustainable cities: how can indicators inform policy?” *Urban Policy and Research*, vol. 33, no. 2, pp. 131–144, 2015.

[2]The Economist Intelligence Unit, “Global liveability index 2018,” 2018. [3]United Nations, “New urban agenda,” 2016.

[4]C. Barnett and S. Parnell, “Ideas, implementation and indicators: epis- temologies of the post-2015 urban agenda,” *Environment and Urbaniza- tion*, vol. 28, no. 1, pp. 87–98, 2016.

[5]F. Caprotti, R. Cowley, A. Datta, V. C. Broto, E. Gao, L. Georgeson,

C. Herrick, N. Odendaal, and S. Joss, “The new urban agenda: key opportunities and challenges for policy and practice,” *Urban research & practice*, vol. 10, no. 3, pp. 367–378, 2017.

[6]United Nations, “Glossary of the habitat iii,” 2016.

[7]P. W. Newton, “Liveable and sustainable? socio-technical challenges for twenty-first-century cities,” *Journal of Urban Technology*, vol. 19, no. 1, pp. 81–102, 2012.

[8]M. Ruth and R. S. Franklin, “Livability for all? conceptual limits and practical implications,” *Applied Geography*, vol. 49, pp. 18–23, 2014. [9]H. J. Miller, F. Witlox, and C. P. Tribby, “Developing context-sensitive

livability indicators for transportation planning: a measurement frame- work,” *Journal of Transport Geography*, vol. 26, pp. 51–64, 2013.

[10]A. Ley and P. Newton, “Creating and sustaining liveable cities,” in *Developing living cities: From analysis to action*. World Scientific, 2010, pp. 191–229.

[11]R. Cervero, *Transit-oriented development in the United States: Experi- ences, challenges, and prospects*. Transportation Research Board, 2004, vol. 102.

[12]D. Sauter and M. Huettenmoser, “Liveable streets and social inclusion,”

*Urban Design International*, vol. 13, no. 2, pp. 67–79, 2008. [13]Mercer, “Quality of life rankings 2018,” 2018.