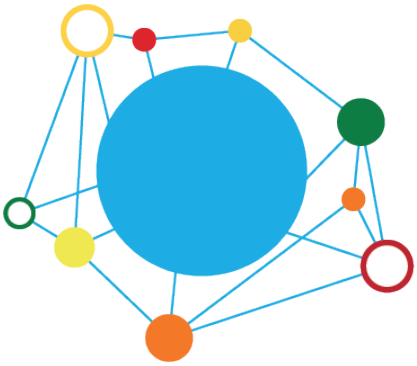


DATA-POP
ALLIANCE

Data-Pop Alliance - UNFPA Maldives Project

CONTRACT No UNFPA/MDV/PSC/17/002

Final Report

July 24, 2018

Gianni Barlacchi, Micol Stock
with supervision from Bruno Lepri, Emmanuel Letouzé

Introduction

This final report follows the first Deliverable, which describes the general scope of the project as well as the proposed approach to develop and evaluate socio-economic indicators exploring Big Data sources and methodologies available in the Maldives. In particular, the aim of this document is to present the findings of the work focused on estimating resident population and population density through Mobile Phone Data.

The report includes:

1. A description of the dataset used
2. An overview of the exploratory analysis and main results
3. An analysis of the population density through mobile data
4. An estimation of the population density using the Voronoi Diagram
5. Conclusions
6. An Annex with suggestions on how to improve the data

1. Data Description

For this project, we obtained a unique mobile usage dataset from a major mobile operator in the Maldives, an island country located in the Indian Ocean.

The dataset used consists of 29,914,732 records of mobile phone logs from January 15 to February 28, 2018. Before sharing the dataset, the telephone company anonymized the data following all local privacy protection laws and regulations.

Each record of mobile usage data contains the information as follows:

- The user ID: a unique identifier attached to a particular user on a given day. It was changed every 24 hours for the same user in such a way that user activities over days cannot be tracked.
- Date and timestamp
- Four types of mobile usage activities: outgoing call, incoming call, outgoing SMS, and incomin SMS.
- Geographic coordinates of cell phone tower (antennas) location: latitude and longitude
- User's gender
- User's age

We generated four datasets that summarize three different aspects of mobile usage behavior that we intended to study: users, temporal pattern, and geography. For each dataset, we summed the total number of mobile usage activities group by variables as follows:

1. User: date, type of activity, user ID, gender, and age.
2. Temporal pattern: date, activity type, and hour of the day.
3. Geography: date, activity type, latitude, and longitude of the antennas. The latitude and longitude of antennas are rounded into a 6 digits accuracy.

As user IDs changed every 24 hours to protect the privacy of users, we calculated all measurements on a daily basis. We explored the dataset using Tableau software, which allowed us to create graphical visualization, drill-down and roll-up the datasets in different ways. In most of the cases, we calculated the median number of activity per day because the distributions were mostly skewed.

We noticed that there were some outlier users with a behavior that exceeds a human being's ability. For example, one user in the dataset sent 8,198 messages in a single day. However, the outlier activities do not impact the result analyzed because a median was calculated, which is more robust to outliers. In order to study the behavior of users, we propose to remove some non-human (bot) users by applying an outlier detection technique such as an inter-quartile rank for one-dimensional data and a local outlier factor for multi-dimensional data.

For our analysis we used the open-source library Pandas, GeoPandas for Python, and the software Tableau for data analysis.

2. Exploratory Data Analysis

2.1 Users

First, we aggregated the data by user to study individual mobile usages behavior. The demographics of users by gender and age group is shown in Table 1. Most of the users' gender is missing in this dataset and therefore cannot be included in our analysis. The median age range of mobile phone users is between 25-34 years old.

Table 1: Average number of users by gender and age group per day

	Outgoing Call	Incoming Call	Outgoing SMS	Incoming SMS
Gender				
Female	9	9	7	9
Male	212	237	126	231
Unknown	26,684	30,920	15,644	39,605
Age group				
12-17 years old	30	35	19	49
18-24 years old	4,047	5,096	3,209	5,881
25-34 years old	10,179	11,567	6,055	14,186
35-44 years old	5,165	5,874	2,768	7,136
45-54 years old	2,843	3,239	1,431	3,600
55-64 years old	1,095	1,199	445	1,294
Others	3,547	4,155	1,849	7,700
Total	26,905	31,166	15,777	39,845

The distribution of the number of mobile usage activities per day is shown in Figure 1. We observe a heavy-tail distribution for every type of activity. Table 2 presents the summary statistics of the distribution.

The median of the outgoing calls, incoming calls, outgoing SMS and incoming SMS are 3, 3, 2, and 3 per day respectively. We find that the maximum numbers of outgoing and incoming SMS are 8,198 and 14,229 which are much higher than the ability of a person to send or receive SMS per day.

Table 2: Summary statistics of the number of mobile usages per day

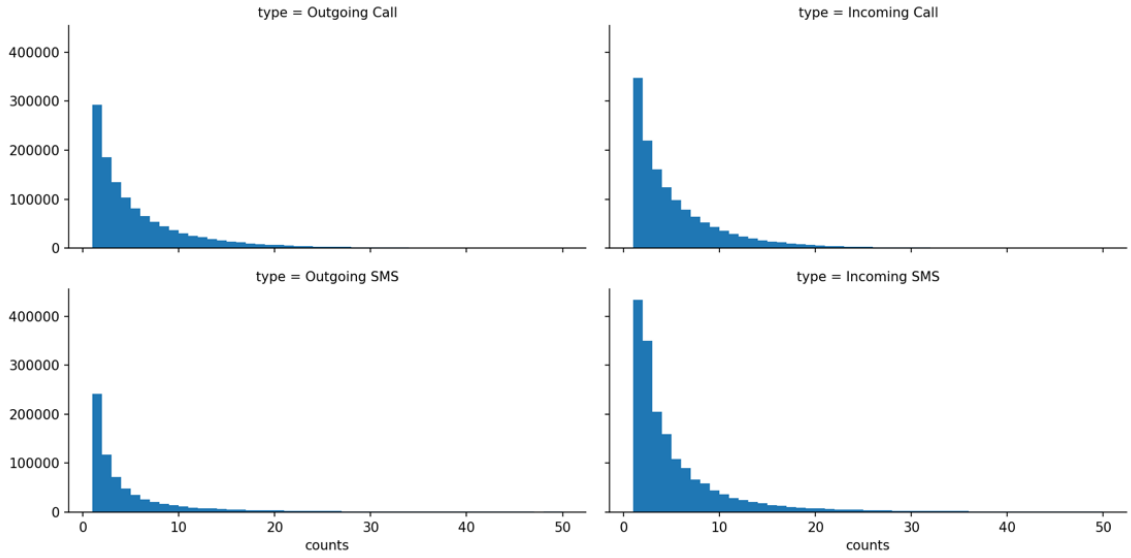


Table 2: Summary statistics of the number of mobile usages per day

Type	mean	std	min	0.25	0.50	0.75	max
Outgoing Call	5.67	6.32	1	2	3	7	150
Incoming Call	5.21	5.60	1	2	3	7	206
Outgoing SMS	6.76	18.93	1	1	2	6	8,198
Incoming SMS	6.11	17.30	1	2	3	6	14,229

The median of mobile usage activities is calculated for different age groups as shown in Table 3. The values are the same for different age groups.

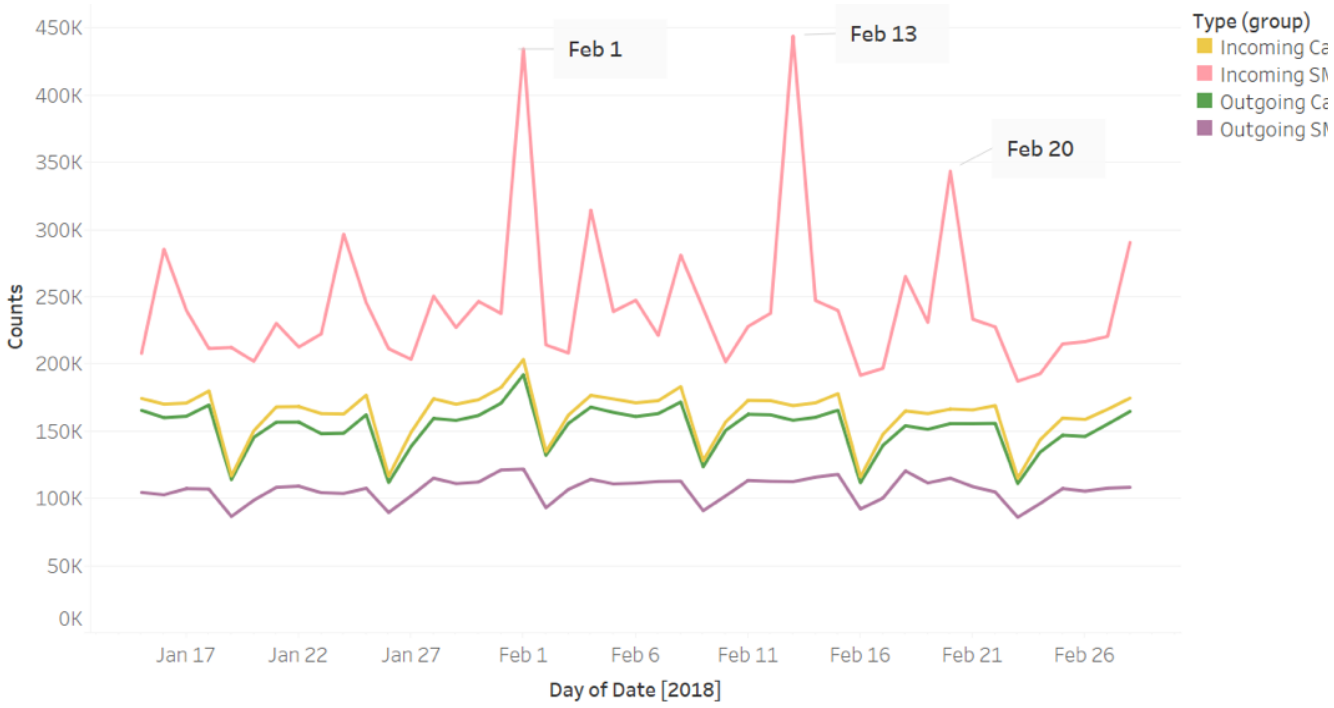
Table 3: Median of mobile usage activities per day by age group

	Outgoing Call	Incoming Call	Outgoing SMS	Incoming SMS
Age group				
12-17 years old	3	3	3	3
18-24 years old	4	4	4	4
25-34 years old	4	4	2	3
35-44 years old	4	3	2	3
45-54 years old	3	3	2	3
55-64 years old	3	3	2	3
Others	2	2	2	2
All users	3	3	2	3

2.2 Temporal Pattern

Next, we analyzed the collective behavior of mobile phone usage over time. The dataset was aggregated by date for different types of mobile usage. Figure 2 shows the time series graph of total mobile usage per day. Three types of activities show the evidence of a weekly seasonality, except the count of incoming SMS which is more fluctuating over time.

Figure 2: Total mobile usage per day



We notice three outlier peaks in the total incoming SMS on February 1st, 13th and 20th, 2018. Moreover, incoming calls, outgoing calls, and outgoing SMS are higher on February 1st than in the weekly seasonality, suggesting some major events were happening in the country during those days.

The evidence for weekly seasonality can be observed from the median mobile usage per weekday, as shown in Figure 3. We find that the volume of activities remained constant from Sunday to Thursday. The activity abruptly decreased on Friday and started to recover to the average volume on Saturday.

Figure 3 : Median mobile usage per weekday

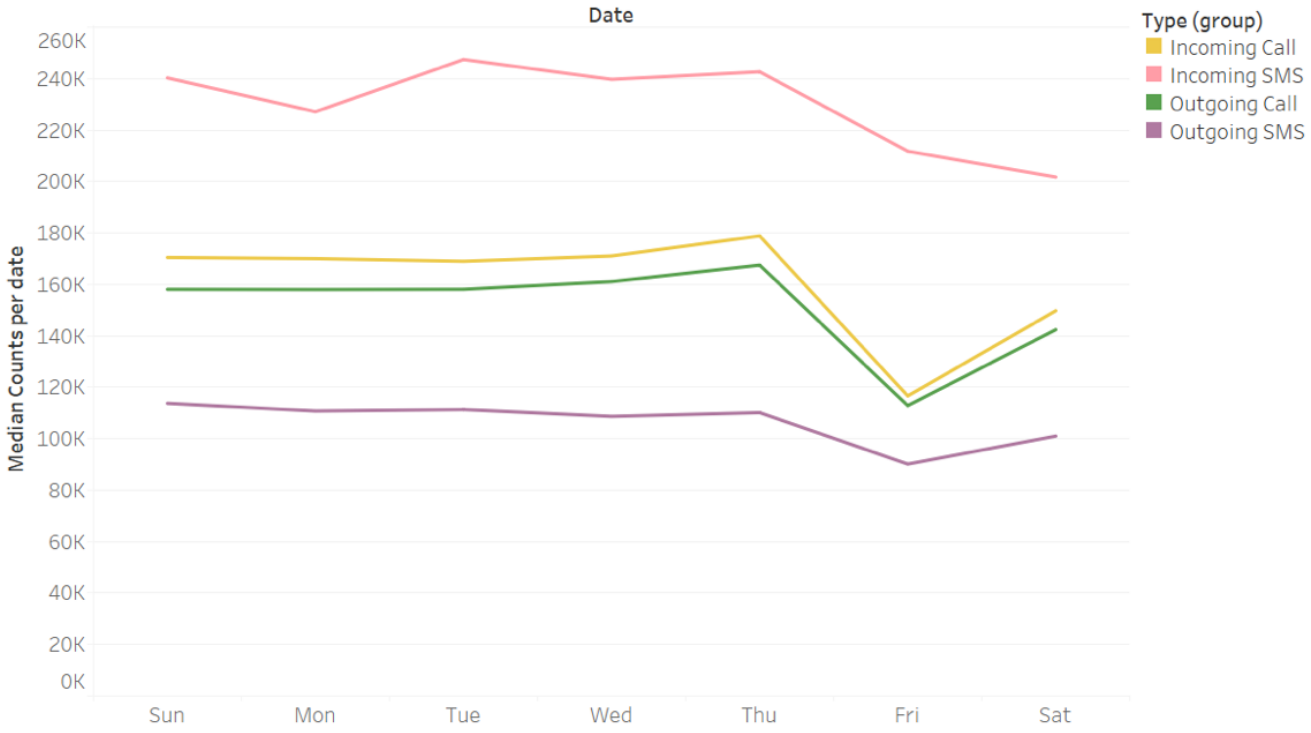


Figure 4 and 5 allow us to investigate the distribution of mobile usage per day in different weekdays. Each dot represents the frequency of mobile usages on a specific day. The box-and-whisker plots are drawn based on interquartile ranges of data on different weekdays.

According to Figure 4, the incoming and outgoing calls show the same weekday pattern. The medians of call activities between Monday and Wednesday are almost the same. The calling activities reached the peak on Thursday, probably because an outlier event happened on February 1st.

The number of calls decreased gradually on Friday and started to increase on Saturday, until it reached the same level on Sunday. We detect some outlier days that significantly deviated from the weekly trend on February 1st.

Figure 4: The box-and-whisker plot of incoming and outgoing call per weekday

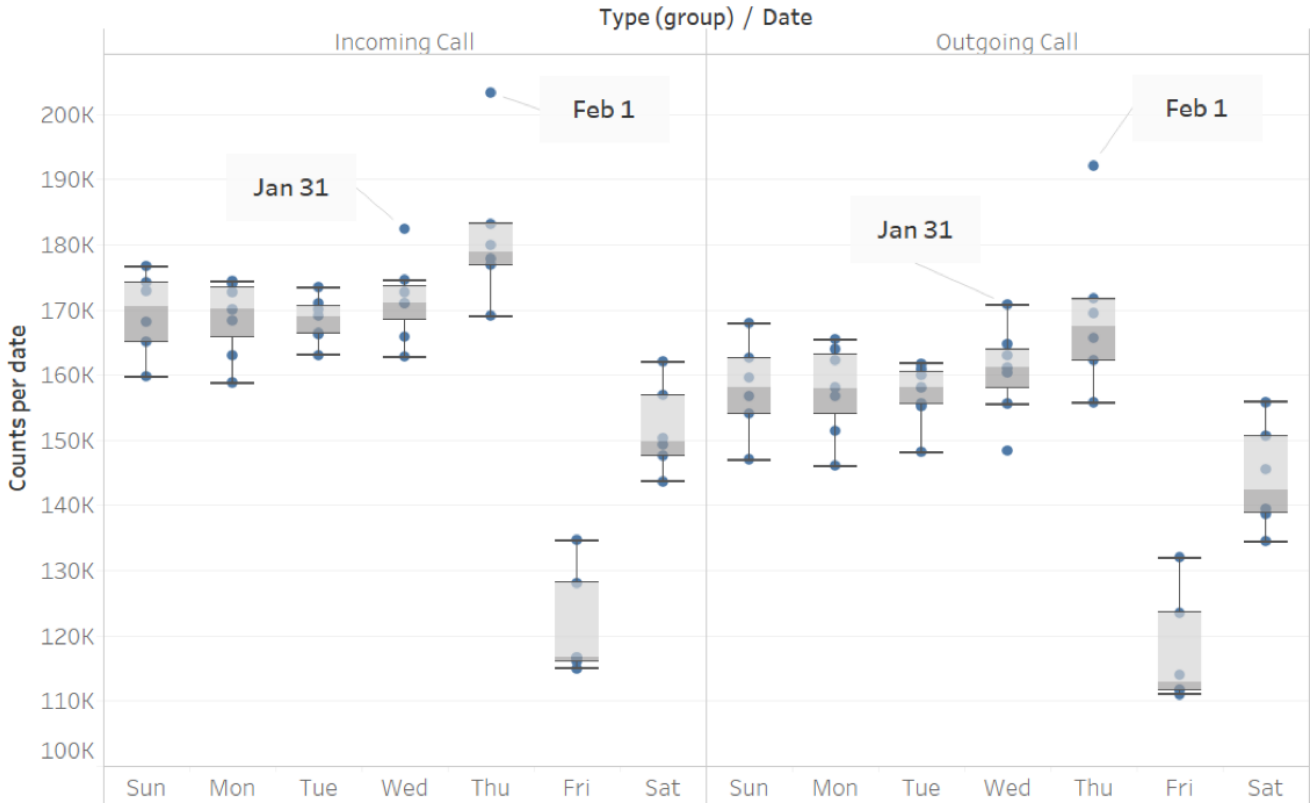
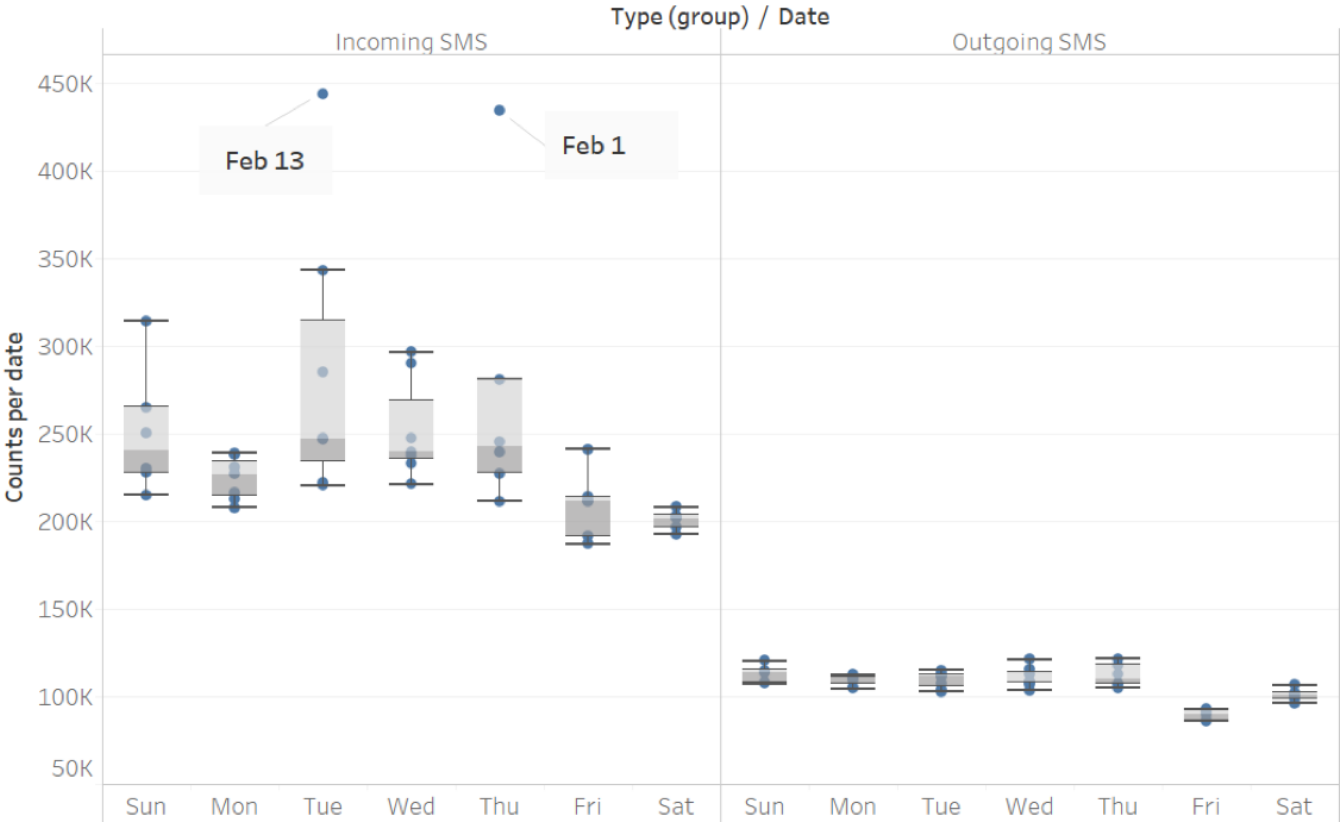


Figure 5 shows the incoming and outgoing SMS distribution on different days of the week. The incoming activity is much higher than the outgoing one, probably because incoming messages came from an external source.

The median of box-whisker plots shows that the frequency of sending and receiving messages is of the same amount for the entire week. We detected the outlier events on February 1st and 13th since the frequency of activity was higher than the interquartile range.

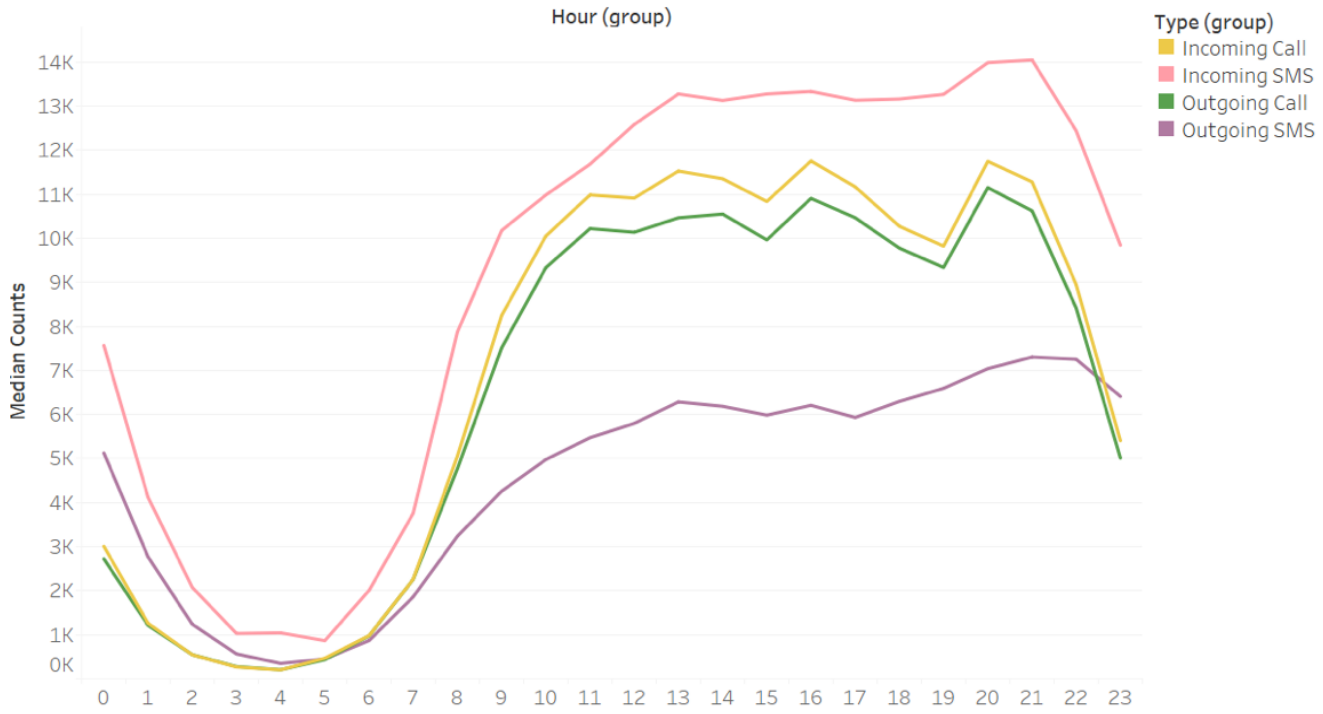
Figure 5: The box-and-whisker plot of incoming and outgoing call per weekday



Next, we analyzed the temporal pattern of mobile phone activities at different hours of the day, as described in Figure 6. The total usage increases rapidly from 6:00 AM to 10:00 AM. After that, the total usage remains constant until 6:00 PM.

The calling volume increases again during the early-night and decreases at a rapid pace until it reaches the lowest point at 4 o'clock, mainly because it is the time at which most people are asleep. In contrast to the hourly pattern, the total number of outgoing SMS increases slowly from 5:00 PM to 10:00 PM.

Figure 6 : Median mobile usage per weekday



Next, Figure 7 and 8 show the box-and-whisker plots of total mobile usage activities per day in different hours. Humans are diurnal animals, naturally active during the daytime, and our circadian rhythm reflects this. The median in Figure 6 shows the difference between wake hours and sleep hours along the human circadian rhythm of the population of the Maldives, in line with results of experiments made in other countries. In Figure 7, the outgoing calls hourly pattern is similar to the one for the incoming calls. The outliers in both call activities are on the same day and at the same hour.

In Figure 8, we see that the number of outgoing and incoming SMS during the wake hours are higher than during sleep hours. Many outliers in incoming SMS are detected during daytime: we can infer that those outliers came from external sources since the outlier did not exist in the pattern of outgoing SMS.

Figure 7: The box-and-whisker plot of incoming and outgoing call per hour

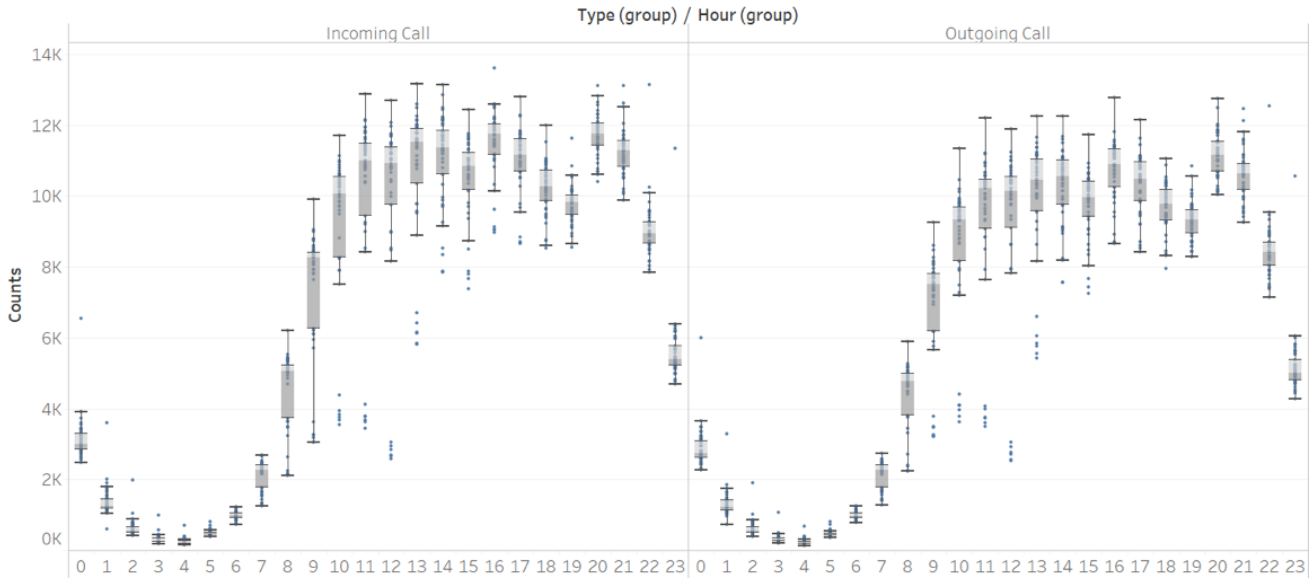


Figure 8: The box-and-whisker plot of incoming and outgoing SMS per weekday

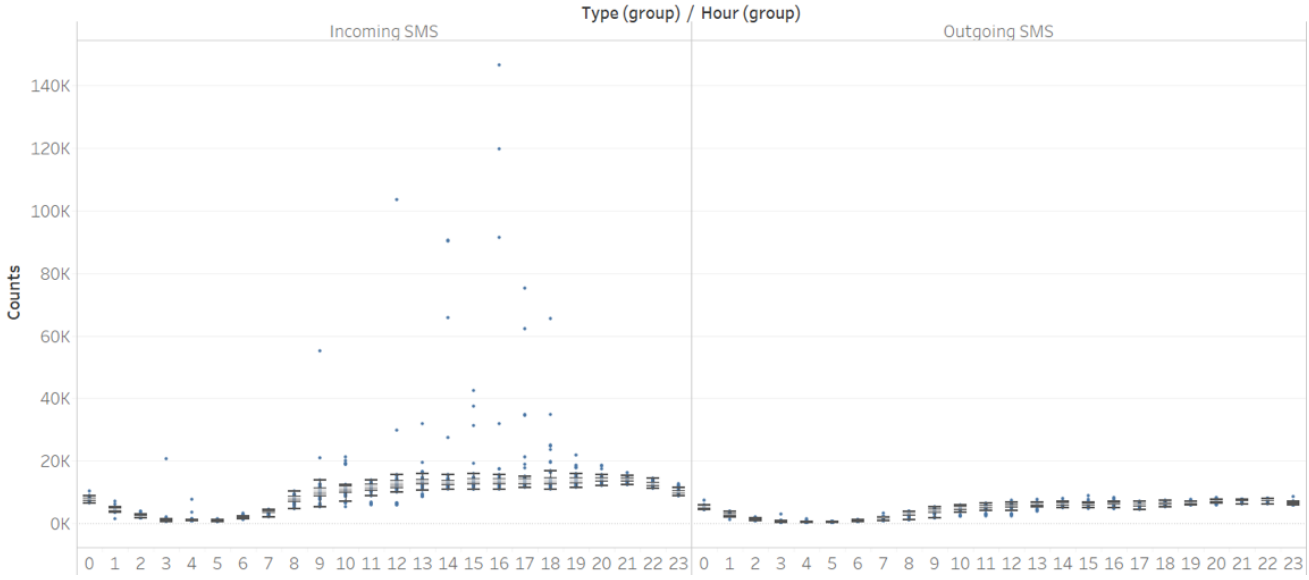
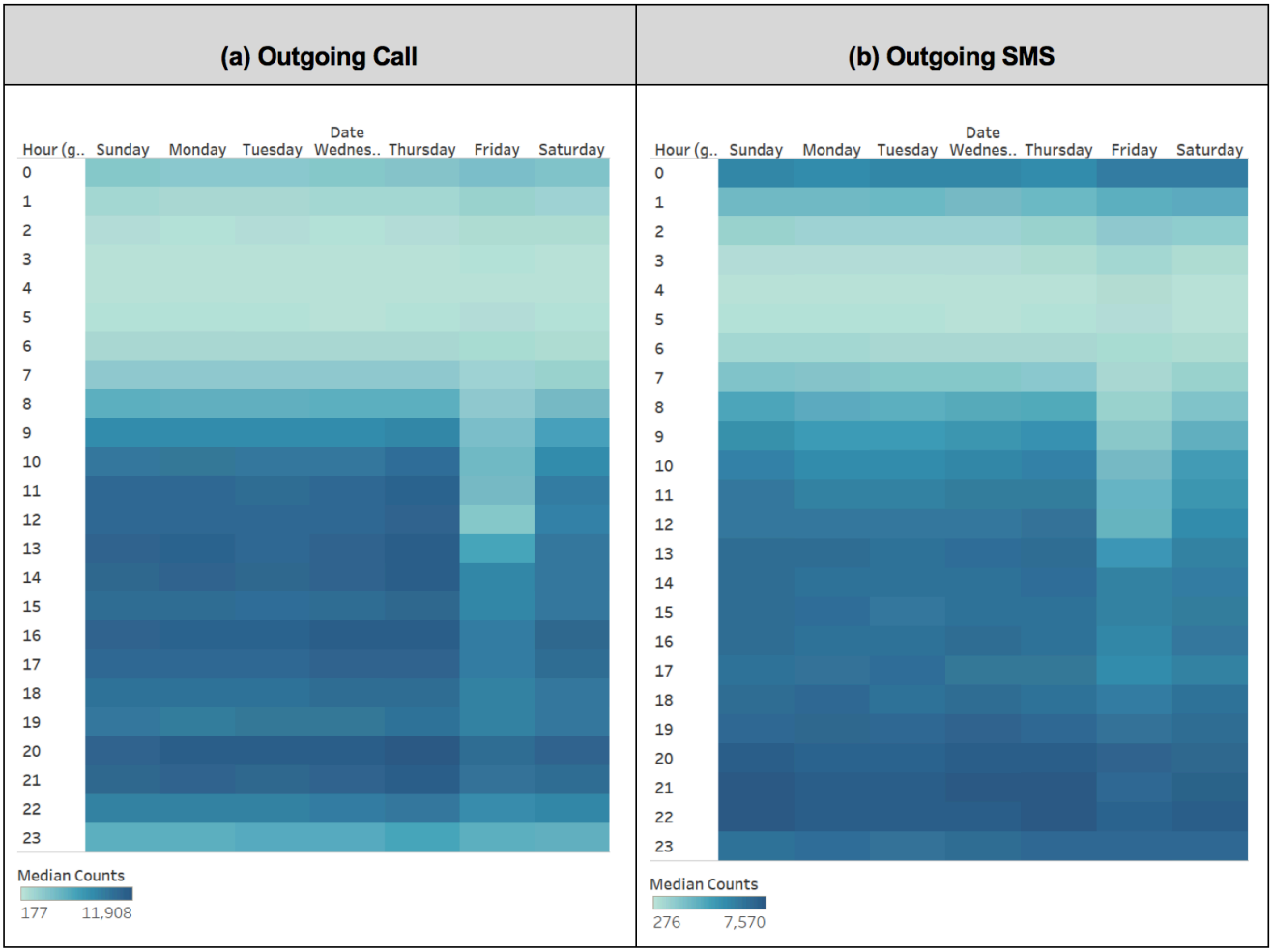


Figure 9 shows the heatmap of the outgoing calls and outgoing SMS median calculated by weekday and hour of the day.

The hourly patterns of outgoing calls and SMS from Sunday to Thursday look quite similar. On Friday, the outgoing call and SMS volumes are lower than usual, especially during the daytime, and start to increase to the average level on Saturday.

Figure 9: Median of outgoing call and SMS by weekday and hour



3. Population Density From Mobile Phone Data

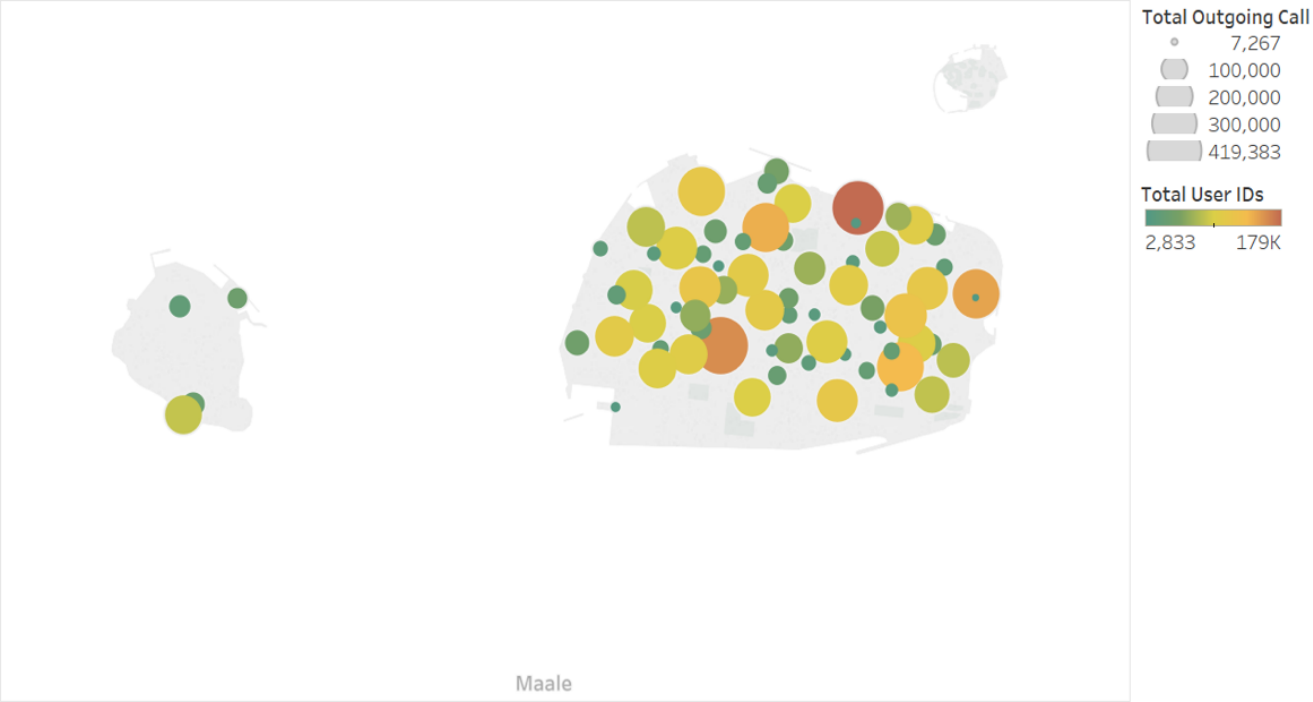
We found 73 cell phone towers in the dataset, of which 72 are located in Malé, the capital and the most populous city in the Maldives. Another cell phone tower is located in the south of Villingili island, specifically at the coordinate (0.749, 73.43746). In this paper, we focused on the outgoing call activities in Malé and Villingili island.

Figure 10 shows the subdivision of the city. In our dataset, there are 68 cell phone towers on Malé island and four cell phone towers on Villingili island.

Figure 11 displays the map of the islands of Malé and Villingili with the color scale indicating the total number of users and the size indicating the total number of outgoing calls in different cell phone tower locations. It shows that most of the calling activities are in the Malé island. We detected the dense calling activities and number of users in the northern Henveiru and the middle of Machchangolhi. In Villingili island, the higher calling activities are in the southern part of the island while there are some activities in the northern part at two separate towers.

Figure 10: Subdivisions of Malé, Maldives (Source: https://en.wikipedia.org/wiki/Malé)

Figure 11: Total outgoing calls and total number of users per cell phone towers in Malé, Maldives



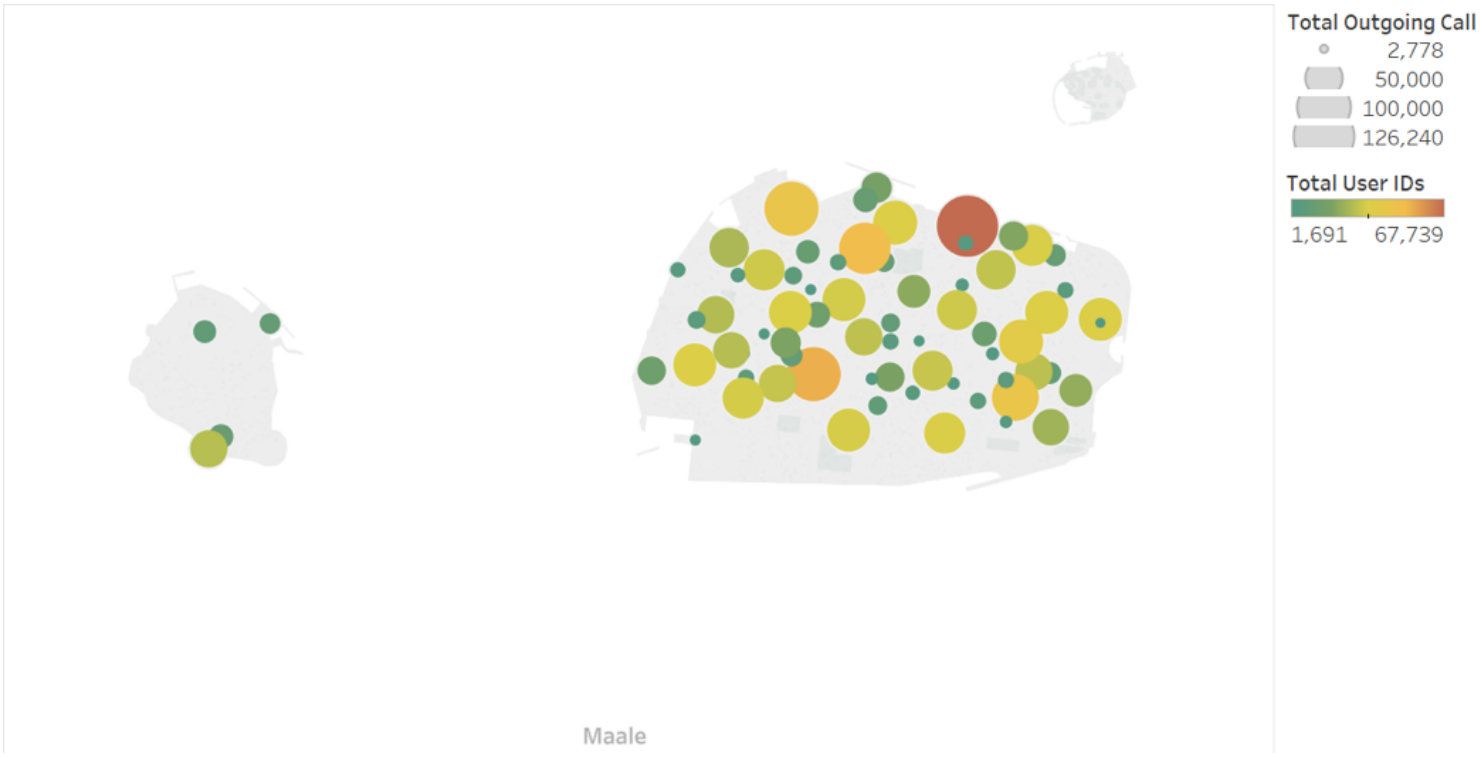
The map of calling activities at different hours of the day is shown in Figure 13. From 1:00 AM - 6:00 AM activity is very low as these are sleep hours for most people. User calls increase from the morning and reach a peak in the afternoon. We applied the different color and circle size scales for each hour bin to observe the relative changes in densely calling locations.

The maps show that the pattern of outgoing calls during 1:00 AM - 6:00 AM is similar to 7:00 PM - 00:00 AM probably. We assume that the dense location during these hours defines residential neighborhoods since people stay at home during the night. The outgoing call location maps during 7:00 AM - 12:00 PM and 1:00 PM - 6:00 PM are similar probably because people are at workplaces during the day.

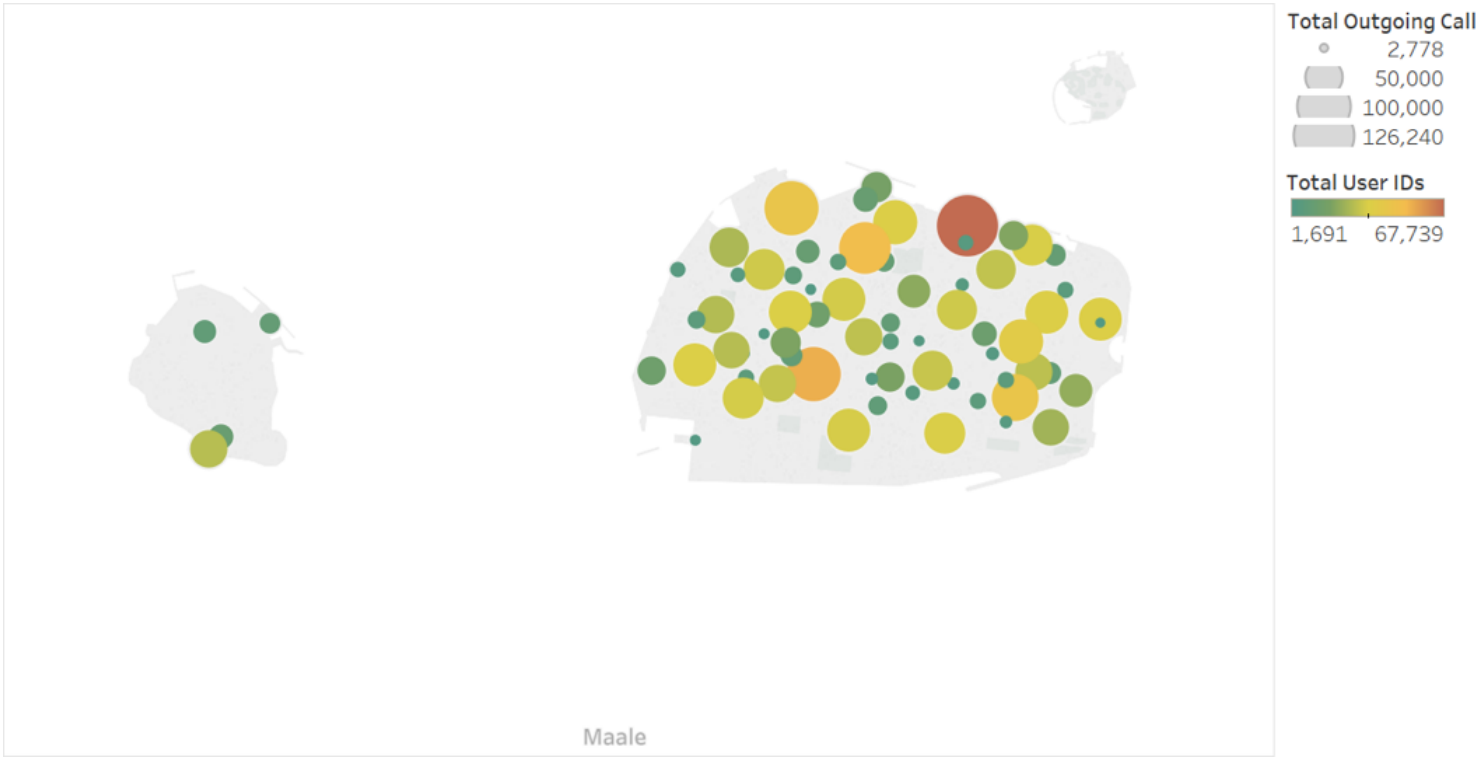
Figure 12: The total outgoing call and total number of users per cell phone towers divided by hours of the day in Malé, Maldives

12.1) 01:00 AM - 6:00 AM

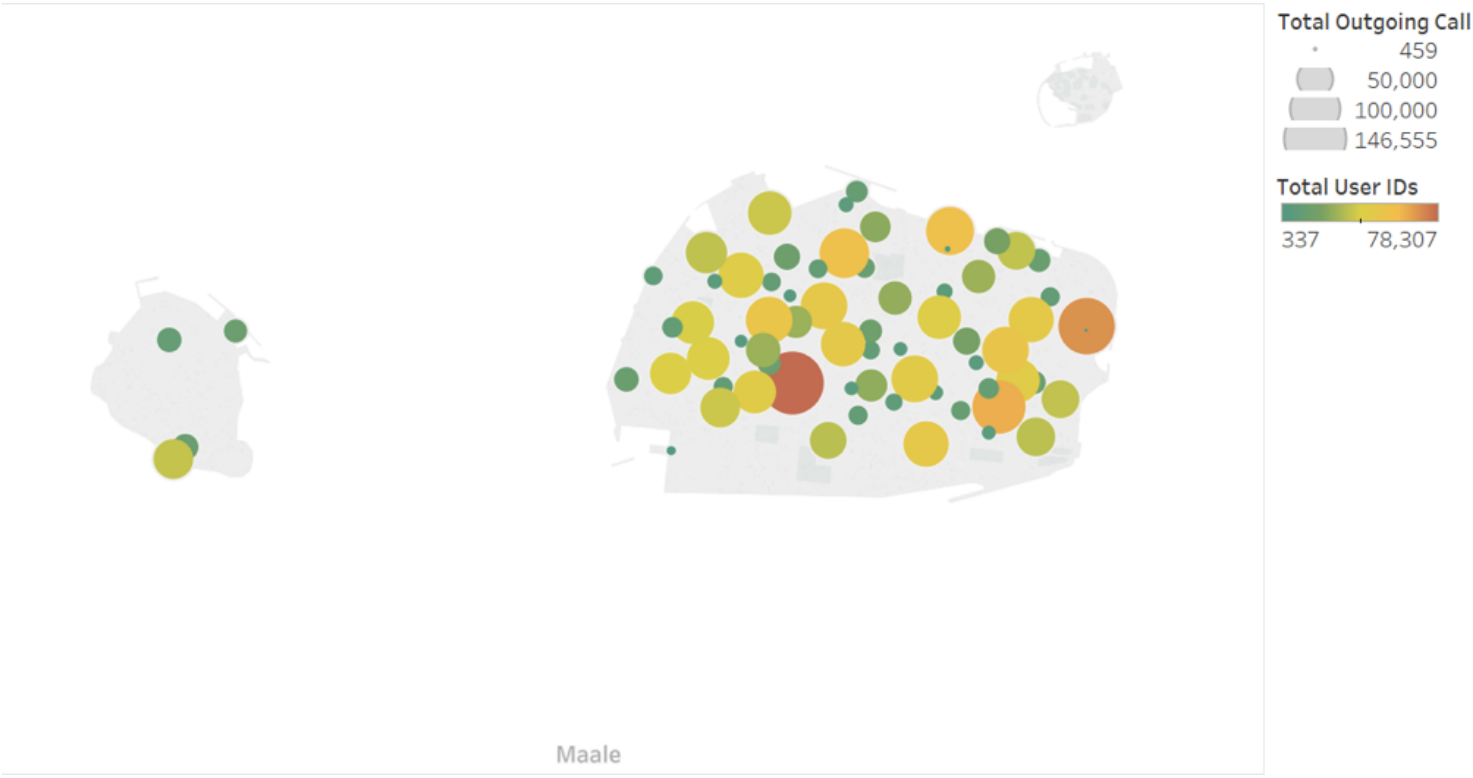
12.2) 7:00 AM - 12:00 PM



12.3) 1:00 PM - 6:00 PM



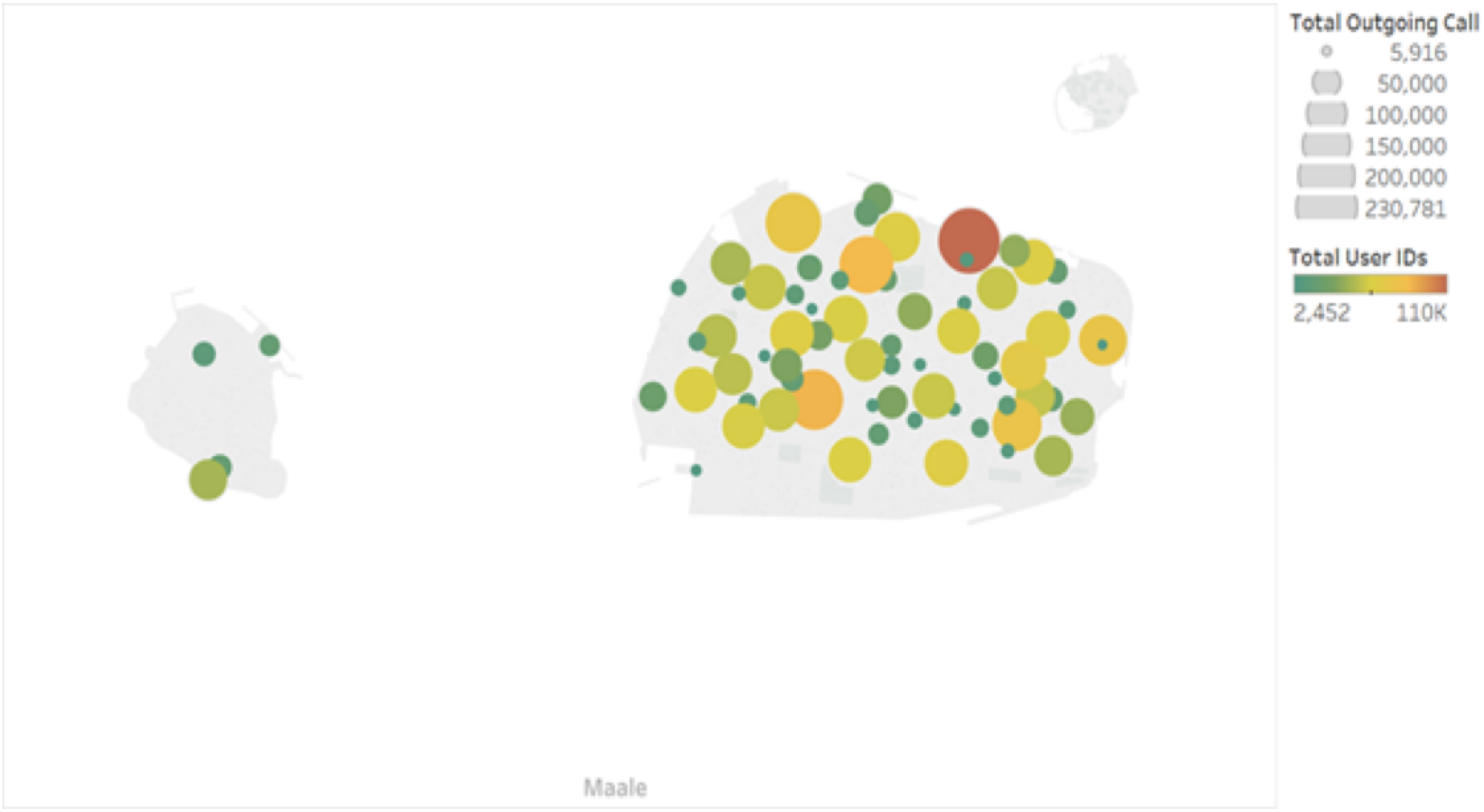
12.4) 7:00 PM - 12:00 AM



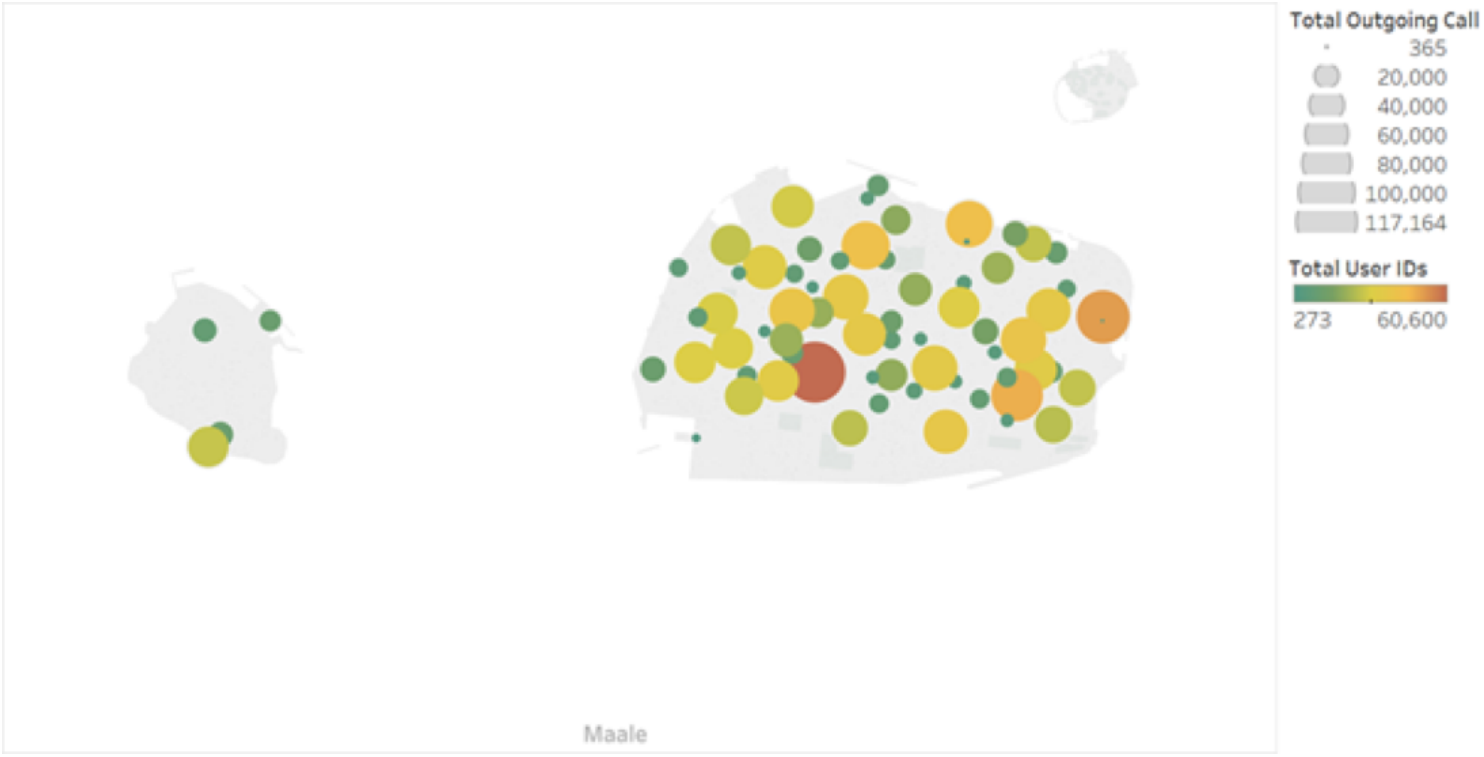
To distinguish the clearer location of residential and work neighborhoods, we show the outgoing call maps in four different weekday and time windows: (a) Sunday – Thursday, 7:00 AM – 6:00 PM, (b) Sunday – Thursday, 7:00 PM – 6:00 AM, (c) Friday – Saturday, 7:00 AM – 6:00 PM, and (d) Friday – Saturday, 7:00 AM – 6:00 PM. During weekdays, we detect the movement from north of Henveiru district in the morning to the middle of Machchangolhi during the night. The pattern is less obvious during the weekend.

Figure 13: The total outgoing calls and total number of users per cell phone towers divided by hours of the day and weekdays in Malé, Maldives

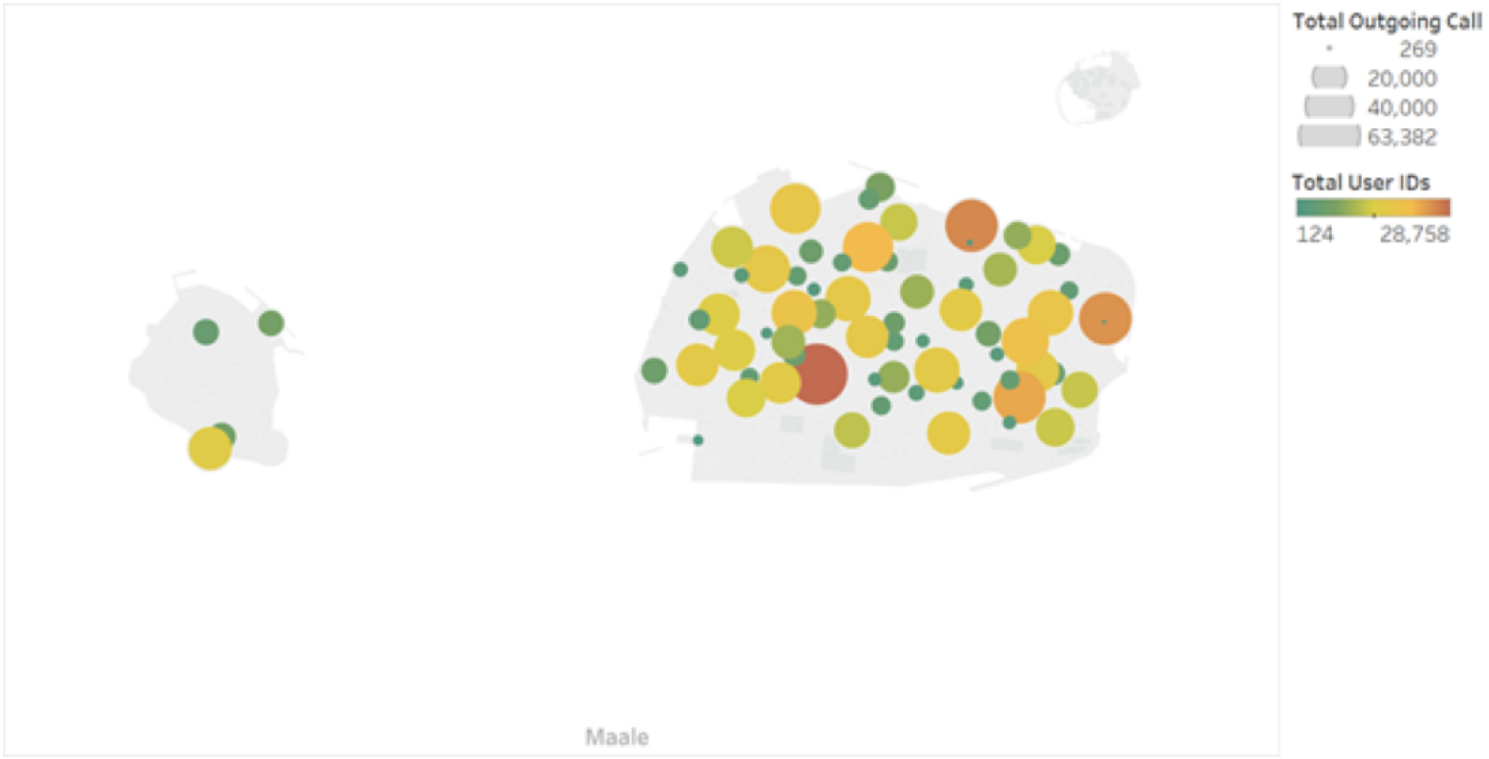
13.1) Sunday – Thursday, 7:00 AM – 6:00 PM



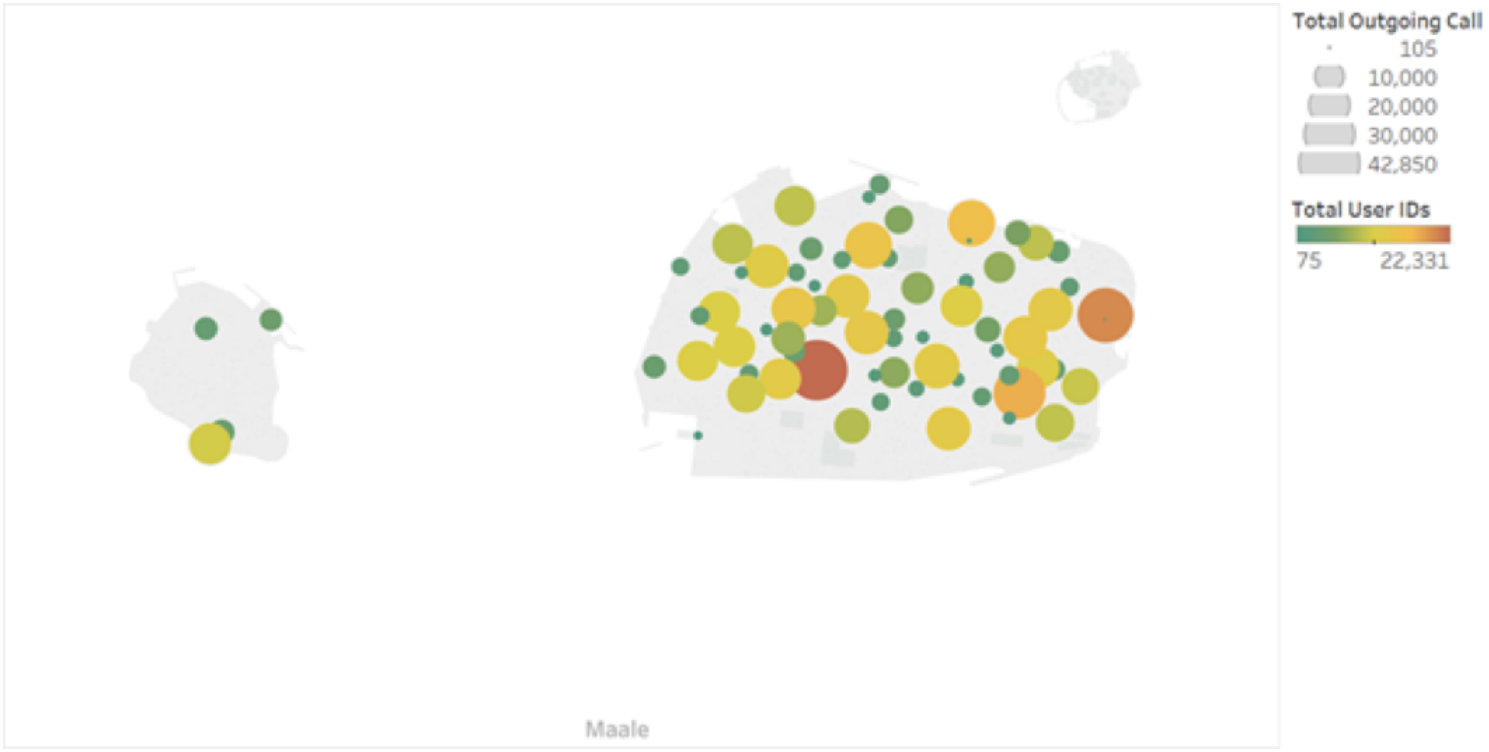
13.2) Sunday – Thursday, 7:00 PM – 6:00 AM



13.3) Friday – Saturday, 7:00 AM - 6:00 PM



13.4) Friday – Saturday, 7:00 PM – 6:00 AM



4. Population density estimation with Voronoi Diagram

As previously described, we selected 68 antennas that are located on the Malé main island since that is the only island where the data contains dense cell phone towers. Unlike in the previous analysis, we removed the antennas in Villingili island as we do not have enough data to provide significant results in terms of density, i.e. too many few antennas and connections of users. A Voronoi diagram (https://en.wikipedia.org/wiki/Voronoi_diagram) was created for that island based on the location of these antennas. The map layer is applied on the diagram to separate the island into regions covered by antennas. We calculated the area of each region in square kilometers (km^2).

In our analysis, we found that the population is distributed differently during the daytime and nighttime. The possible explanation is that people stay at home in the evening and go to work during the daytime. For this reason, we calculated the population density separately during daytime (7:00 AM – 6:00 PM) and nighttime (7:00 PM – 6:00 AM) in the Voronoi Diagram. We chose only calling activities data on weekdays (Sunday – Thursday) because that is the period in which the daytime/nighttime distinction is clearly reflected.

User locations are identified from the antenna that is mostly used, regardless of activity types, since the antenna location indicates where the user was on that timestamp.

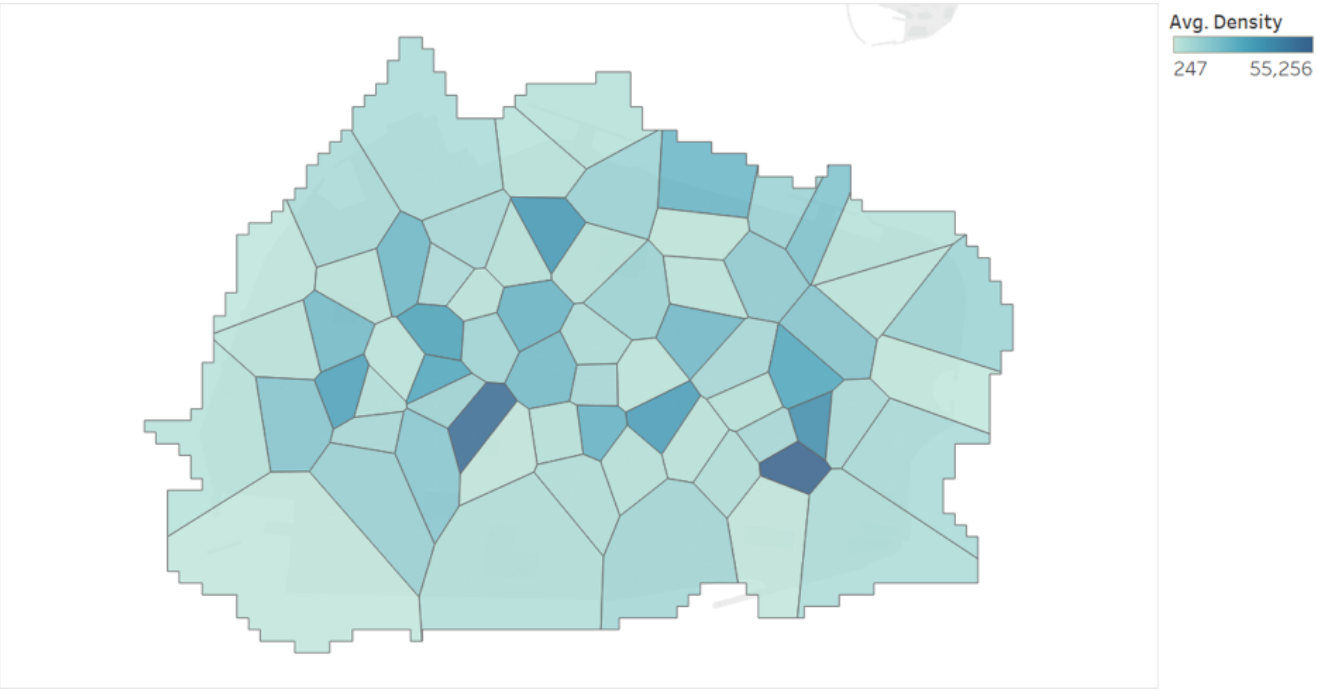
For each region in Malé island, the population density was estimated from:

$$density = \frac{The\ total\ number\ of\ users\ at\ duration\ t}{The\ total\ area\ of\ region\ r\ (in\ km^2)}$$

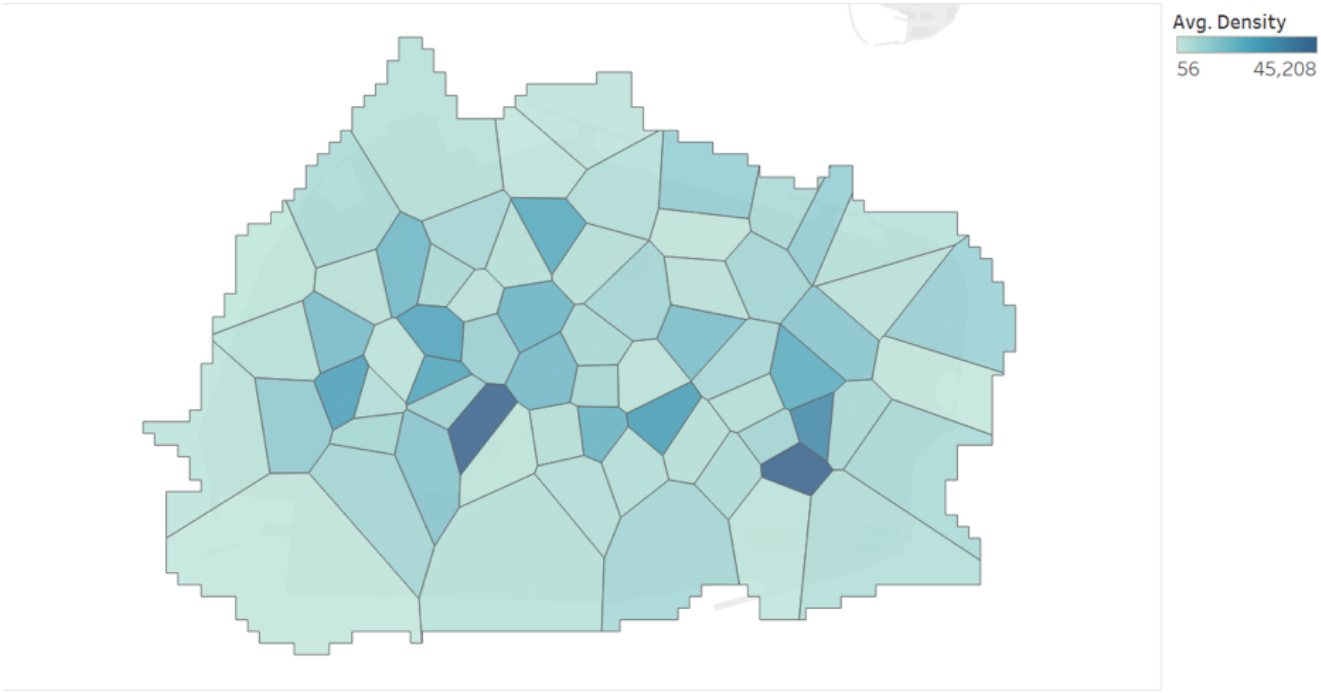
Where r is a region separated from the Voronoi diagram, and t is a time duration, either daytime (7:00 AM - 6:00 PM) and nighttime (7:00 PM - 6:00 AM).

Since the user IDs of the dataset were changing over the different days, we looked at the population density on a daily basis, calculating the average density of each region as a result.

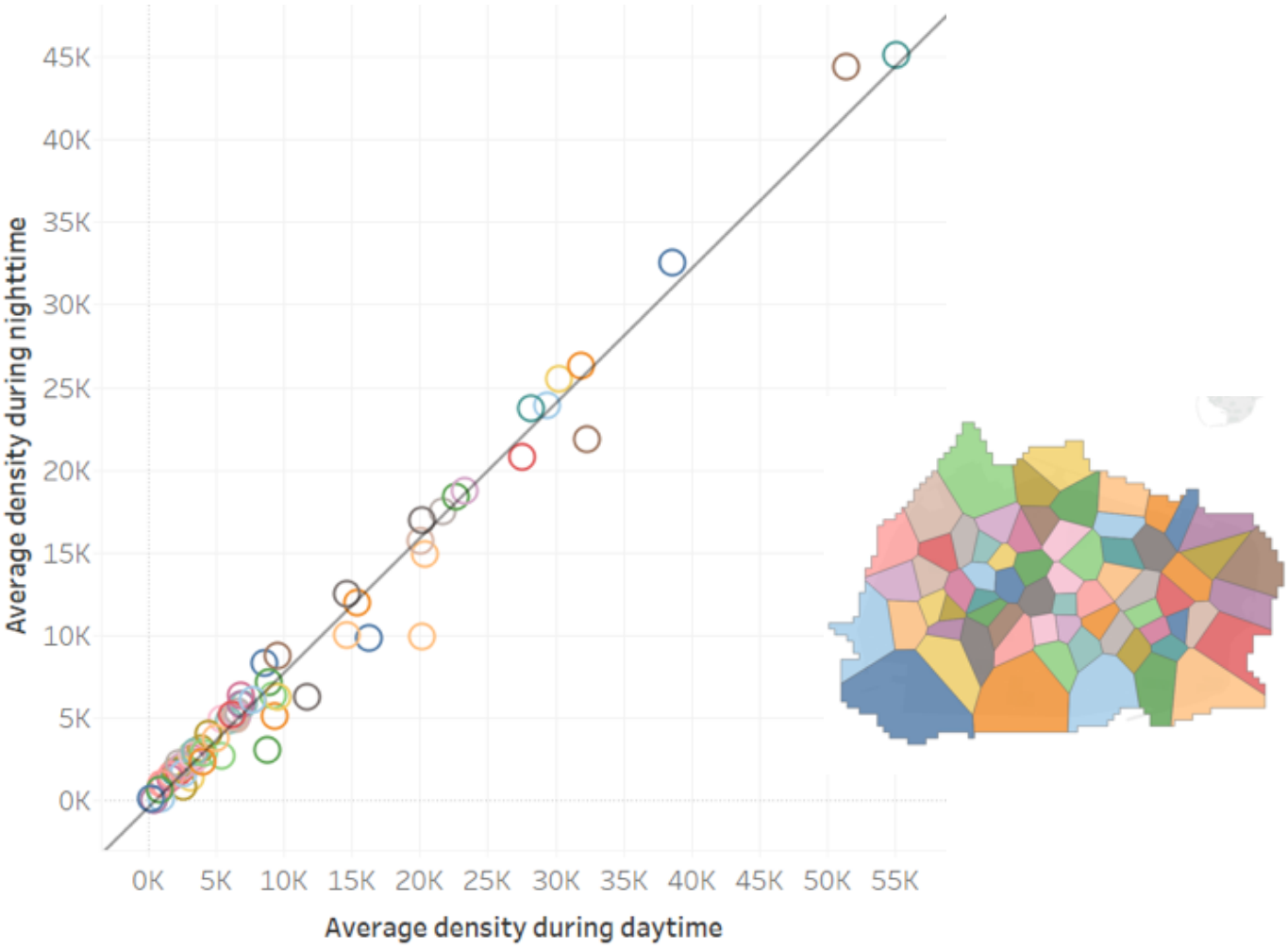
Average density on weekday, daytime (7:00 AM - 6:00 PM)



Average density on weekday, nighttime (7:00 PM - 6:00 AM)

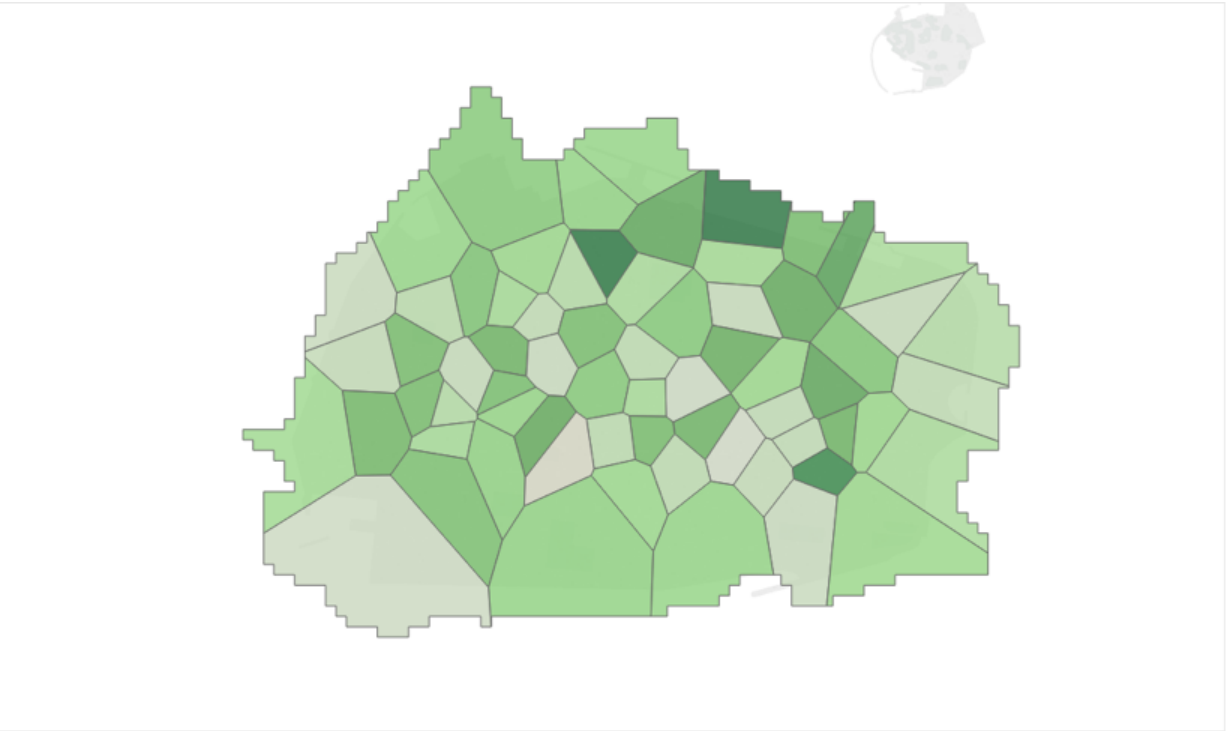


The correlation between density during daytime and nighttime

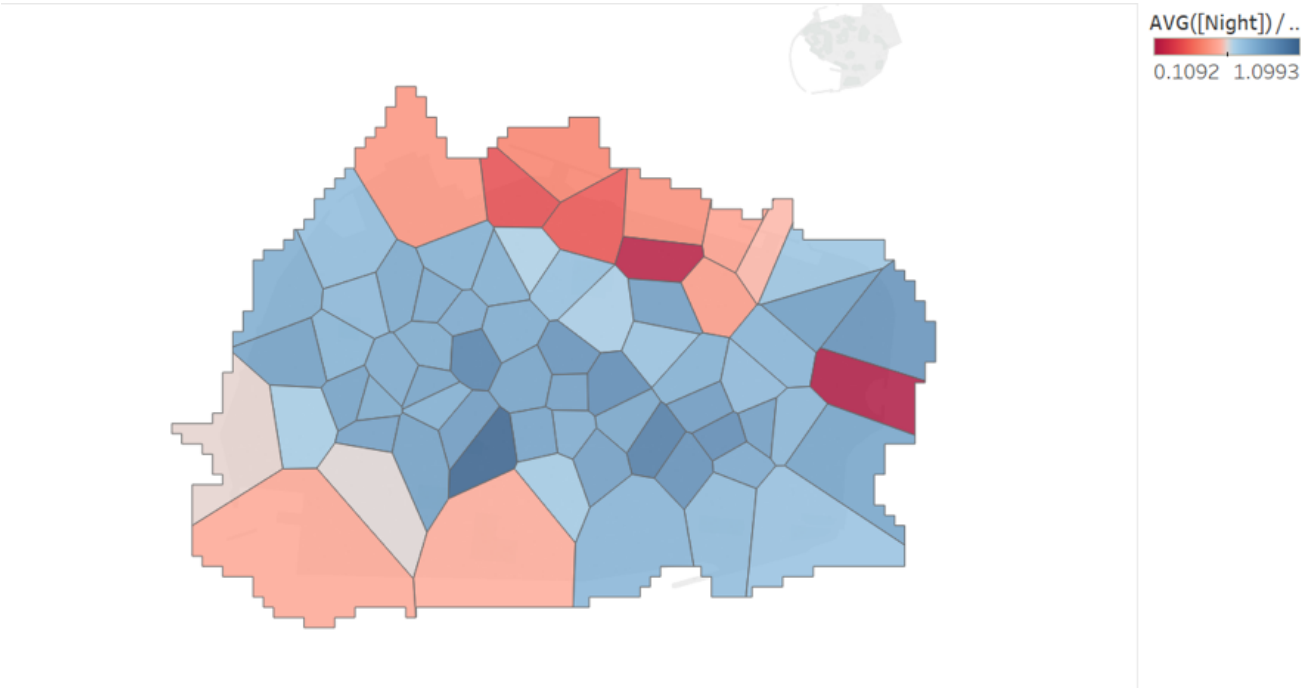


Mobile phone usage is higher during daytime than nighttime, but the usage in each region tend to be consistent in time, i.e., the percentage of users connecting to each antennas during day time and night time is consistent in most of the region. The correlation coefficient between the density during daytime and nighttime is very high: R^2=0.989994.

The difference between density on daytime and nighttime

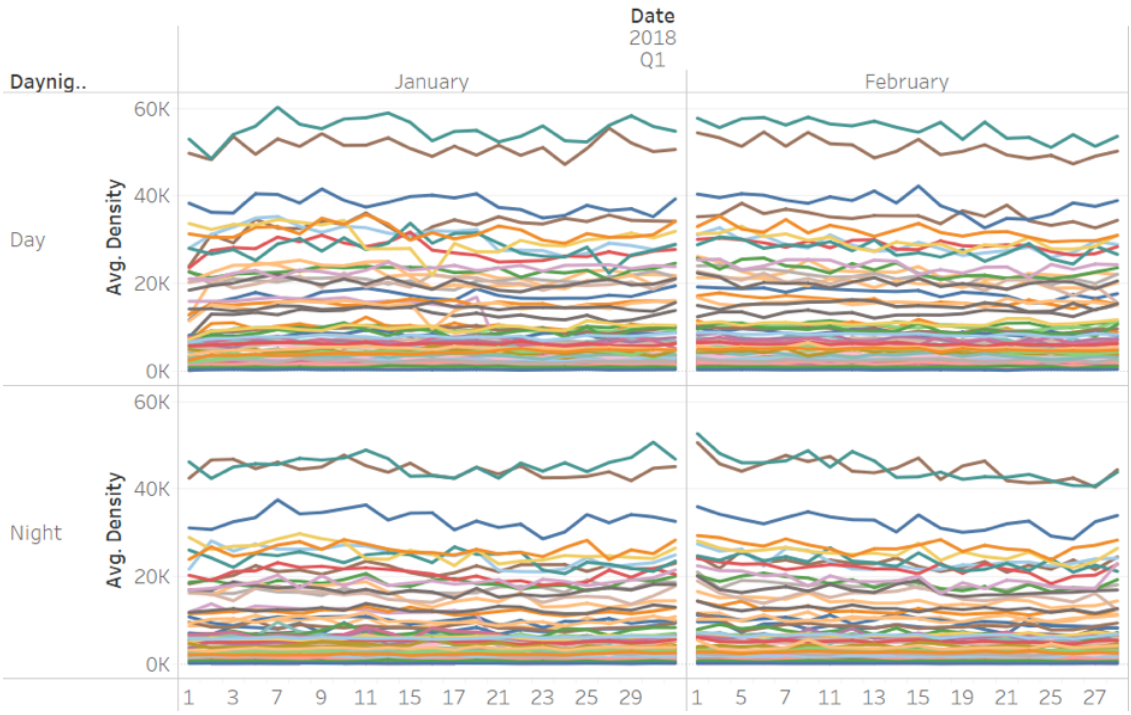


The ratio between density of nighttime over daytime



The ratio between the density of nighttime over daytime: the higher ratio (darkest blue color) indicates that the region has a higher population density during the night. The lower ratio (darkest red color) indicates that the population is denser during the day.

It is worth noticing that the density estimated from our approach is consistent over two months. This results confirm how the mobile phone data are a reliable and sustained source for addressing questions related to the population.



5. Conclusions

This study is an initial pilot aimed at demonstrating how mobile phone data can be used to estimate the population in the Maldives, particularly in the capital city of Malé. This small nation consists of more than 1,000 islands and, as a result, its dispersed population is difficult to measure. Therefore there are strong incentives to implement new methodologies to produce higher quality statistics. The ambitious 2030 Agenda set forth by the United Nations in 2015 and endorsed across the globe by nearly every member nation will be all but impossible to achieve without greater availability of quality, timely data which governments may leverage to make evidence-based decisions, and for citizens to hold them accountable.

In order to realize the stated goal of 'leaving no one behind', there is also an urgent need to obtain disaggregated data along different dimensions such as gender, age and socio-economic status. The 2017 Report on Sustainable Development by Secretary General Antonio Guterres emphasized this point in particular, underscoring "the need for reliable, timely, accessible and disaggregated data to measure progress, inform decision-making and ensure that everyone is counted". The prospects of actualizing such a goal are strong; current technological capacities already allow for such data to be gathered, which is a significant improvement over what national traditional statistics can offer.

The innovative methodology used in this pilot project may be applied to the national census process. The preliminary results generated illustrate how mobile phone data can estimate population size and density and enable timely analysis. Further studies could leverage CDRs to analyze gender, movement patterns, population density and activity in almost real time. These granular insights are critical in the work towards achieving the SDGs and help ensure that 'no one is left behind'.

Although there were a few challenges in this study, mainly pertaining to the type of data and geographical coverage provided (as explored further in Annex I), we still see a potential to scale up this project and expand the analysis beyond the island of Malé. We therefore would propose to discuss opportunities to look further into innovative methodologies for future population estimates.

Annex I: Data improvement recommendations

In order to improve the current analysis generated we would suggest to focus on the following changes: (i) extend the duration of a single user ID per user, (ii) include internet usage, and (iii) expand the geographical area to include more islands. With only one island in the dataset, the methodologies analyzed in this report are not comprehensive enough to be applied to actual policy as we do not have sufficient data points. By applying these three improvements we would be able to provide a more detailed analysis on the population in the Maldives and, more importantly, we could directly compare the results with the official census data.

Detailed descriptions of the suggested data improvements are as follows:

i. Extend the user ID: A complete analysis requires user IDs to apply for a greater time period than 24 hours, as was the case in this report. Each individual -- or, in effect, each phone number -- would be associated with a unique ID, which varies in duration depending on the mobile operator. Lacking the ability to perform a similar analysis on data containing unique IDs of a longer duration, it is not possible to conduct a comprehensive mobility analysis and the quality of the analysis of population density would also be affected. For privacy reasons, however, it is customary to limit the size of the time window to three months. We also suggest to provide at least 6 months of CDR data, whereas only 6 weeks were available in this pilot study. The longer the window of time, the greater the quantity of data available for understanding user behavior.

ii. Internet usage: Another important aspect is related to the type of events (e.g., call, sms and internet usage) that are present in the dataset, each of which is usually associated with the location of the antenna used. Collecting a greater number of possible events per user allows for the identification of where a user is living while preserving their anonymity. Internet usage in particular would provide significant insight into the mobility of users and determine in which area they are living. By analysing only calls and SMS, we are limited to getting the position of the user very few times a day, which in turn limits the ability to estimate population density.

Iii. Expand the geographical area: The analyses reported in this document have been conducted with data from Malé island where a large percentage of the residents commute daily. The limited geographical range of this study introduces noisy data as we are not able to differentiate between residents and commuters working in the area. Including additional islands in our analysis would improve the ability to estimate population density of residents and eliminate commuters from the dataset.