# From law of large number to the culture computation

Xiaoke(Jimmy) Shen

June 1, 2017

## 1 Introduction

Culture is important to everyone. Better understanding a culture can help the people with different background and culture know each other in a better way. How to do the culture mining or computation? Thanks for the theory from mathematics and also thanks to the modern computer technologies, it seems we do can find some ways to explore the culture by doing the data analysis, such find the sleeping time of a city.

## 2 Law of Large Number

Law of large number is one of the most wonderful law in this world(another one is from my personal's point of view). It shows us that after we get enough repeat observation or sampling , we can have some insight about the object we are observing. The theory will not be proved here as I am not from mathematics department. However, the simulation will be done here to demo the basic idea about the law of large number.

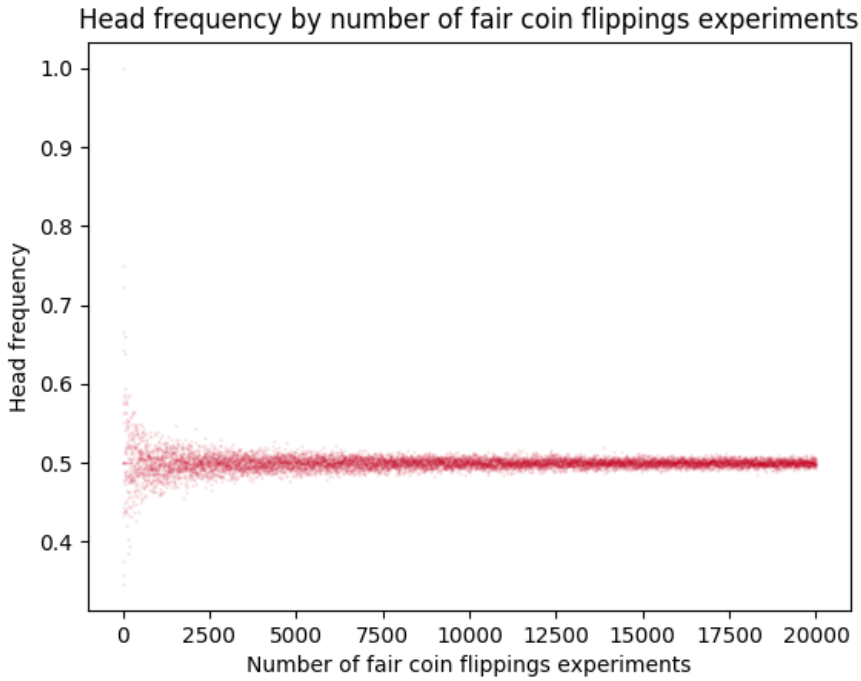


Figure 1: An illustration of the law of large numbers by using a fair coin to do the flipping experiments.

The illustration of the law of large numbers using a fair coin to do the flipping experiments. As the number of experiments increases, the average of getting the head ratio of all the results approaches to 0.5.

# 3 What is the sleeping time for a city?

Since the law of large number tells us with enough observations, we can get some stable observations. I am going to do something based on this law to see whether we can exploring the sleeping time of 5 cities(The data is provided by the Cultural Analytics Lab of the Graduate Center, CUNY). Those five cities are Buenos Aires, London, Los Angeles, Nairobi and Tokyo.

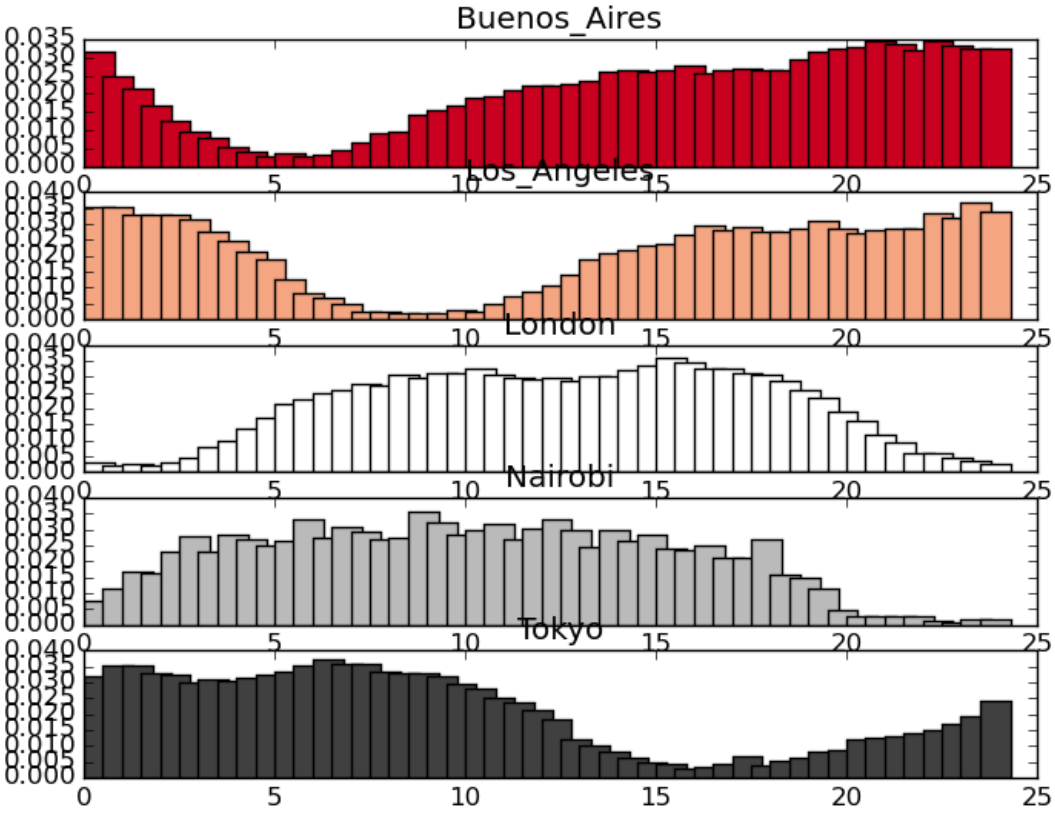


Figure 2: Final result of the sleeping time analysis of 5 cities. The x axis shows the time period of a day from 0:00 am to 23:59 pm. And each period is 30 minutes. The y axis is the activation ratio of the specified time period.

The final result is shown in figure 2. The x axis shows the time period of a day from 0:00 am to 23:59 pm. And each period is 30 minutes. The y axis is the activation ratio of the specified time period. From the result we can see different cities has totally different work and rest style. If we take the small value of activation ratio as the sleeping time of the city, each city has different sleeping time which strongly related to the life style and culture.

How can we get it? Following sessions are the step by step explanation to show how finally we find some insight about the sleeping time of a city.

### 3.1 Observe the twitter post every minute

If we observe the twitter post every minute, the visualization results are shown in the following figures. From the histogram of the 5 cities twitter post per minute we can see this value is mainly between 0 and 10. The twitter post number per minute for each city is visualized by using a gray leve image. When we get a pure dark point, it means that point has the maximum number of the twitter post. When the gray level is smaller, then it means the number of post is decreased. When there is no post, it will become 0 which is totally blank. In the data argumentation part, a threshold is set to make sure the result only has two status 0 and 1 to help us find a better insight behind the data.

From those result we already see some structure of the active level for each city based on time slots.

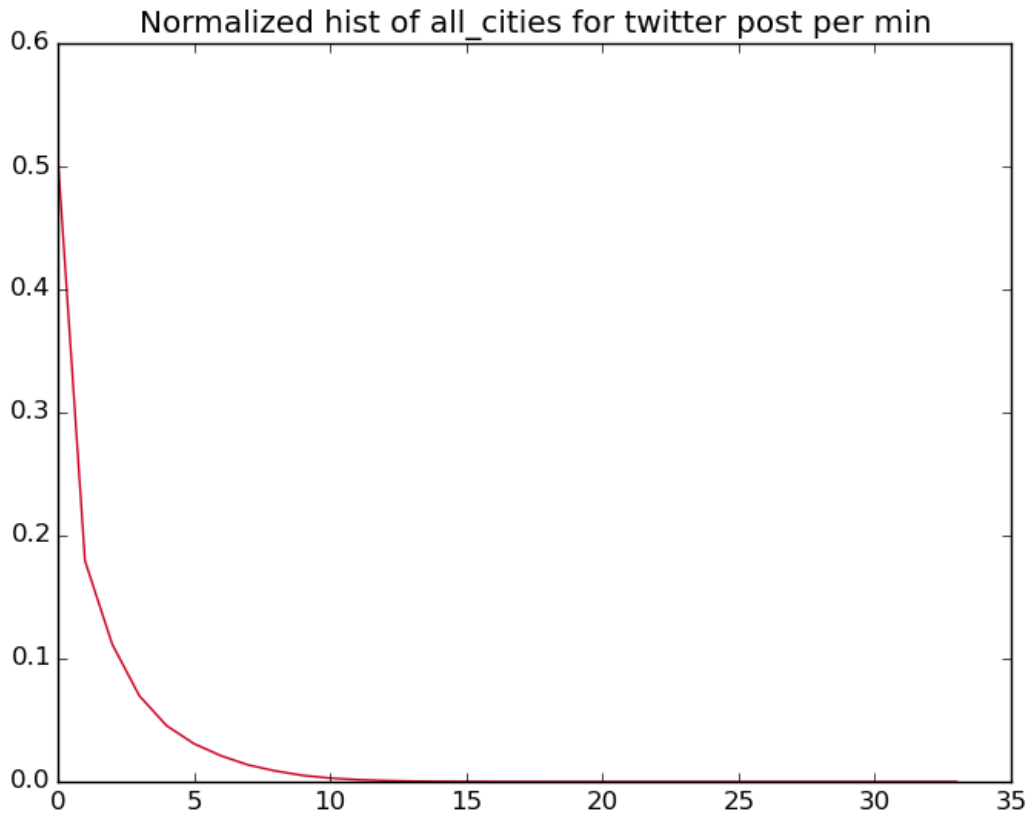


Figure 3: Histogram of five cities when the observation time slot is every one minute

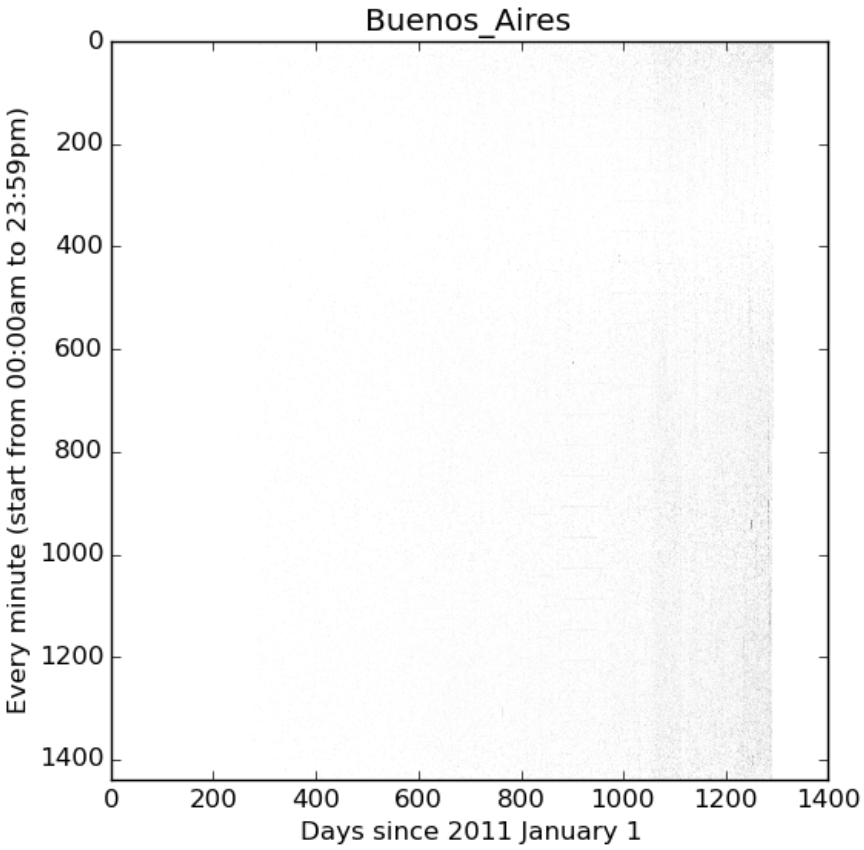


Figure 4: Buenos Aires every minute without data argumentation

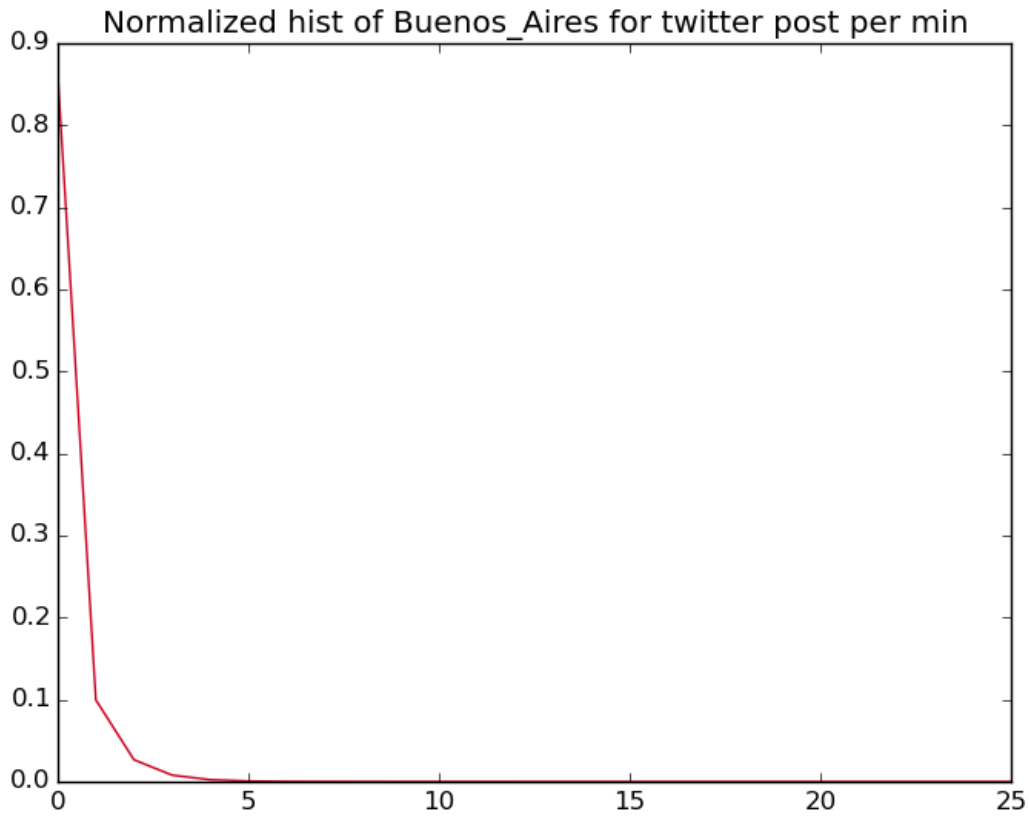


Figure 5: Histogram of Buenos Aires

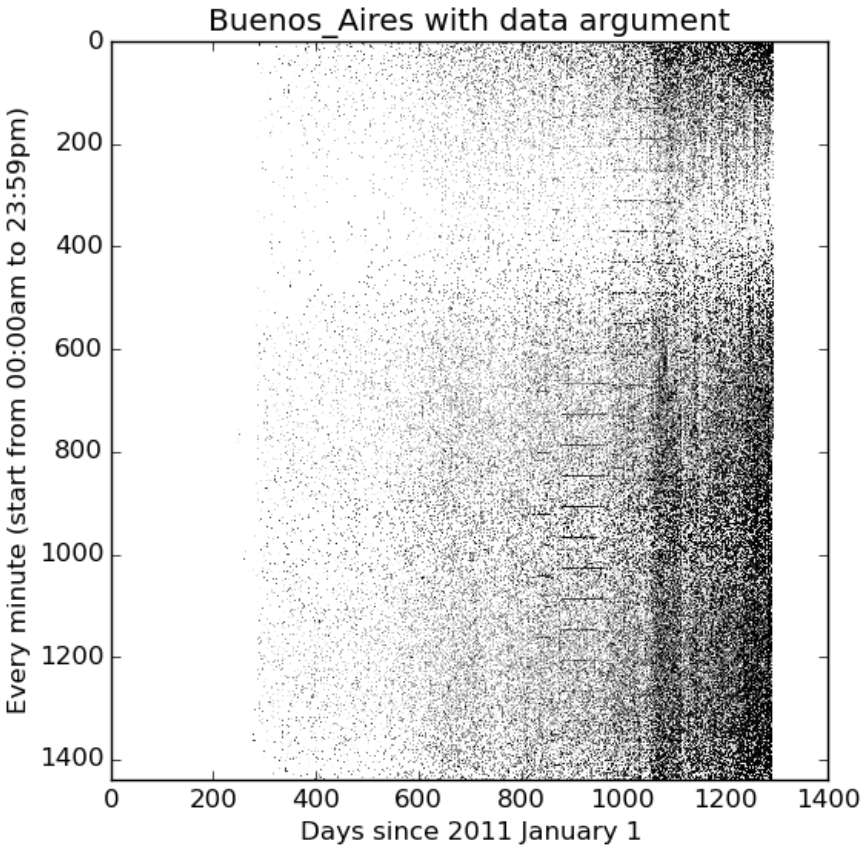


Figure 6: Buenos Aires every minute with data argumentation

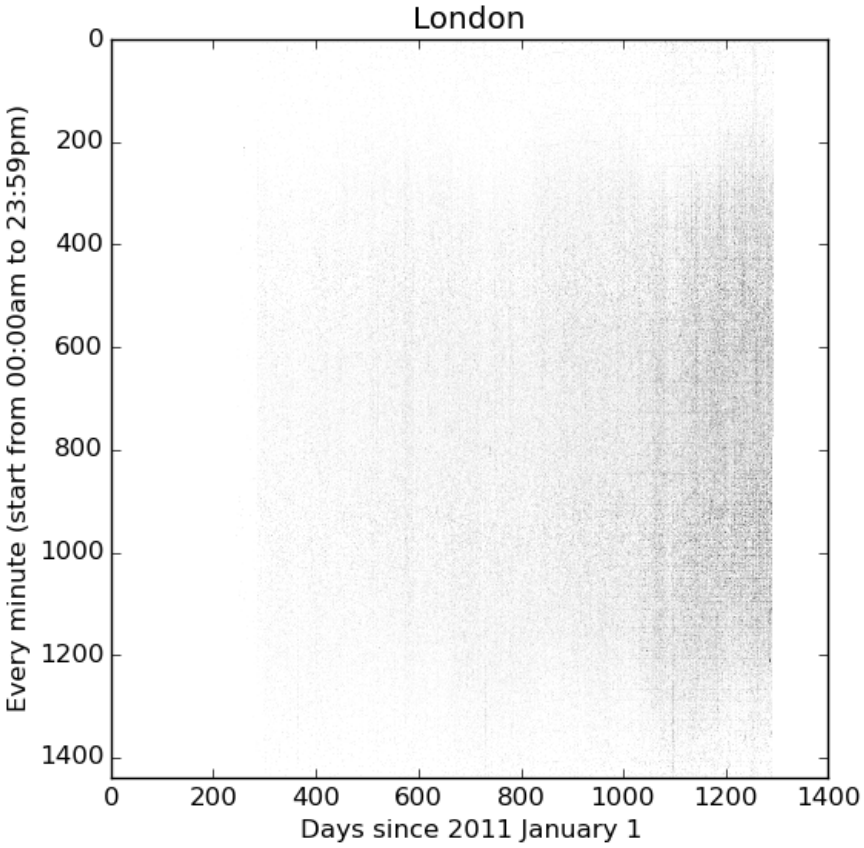


Figure 7: London every minute without data argumentation

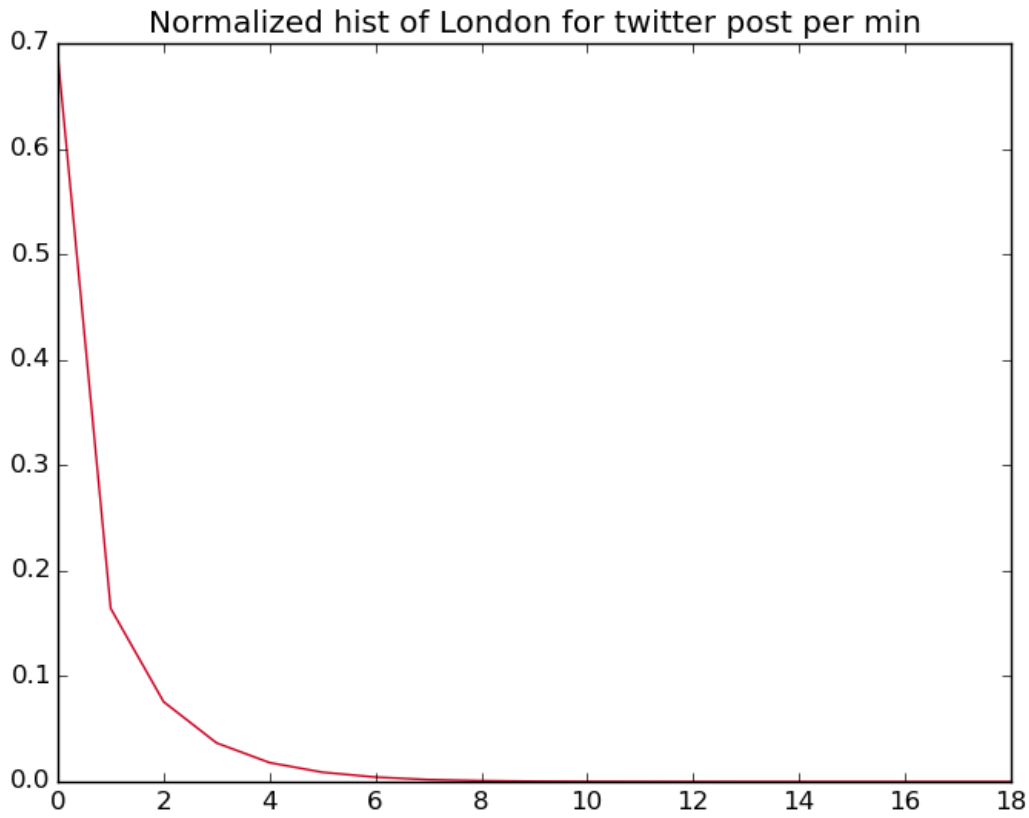


Figure 8: Histogram of London

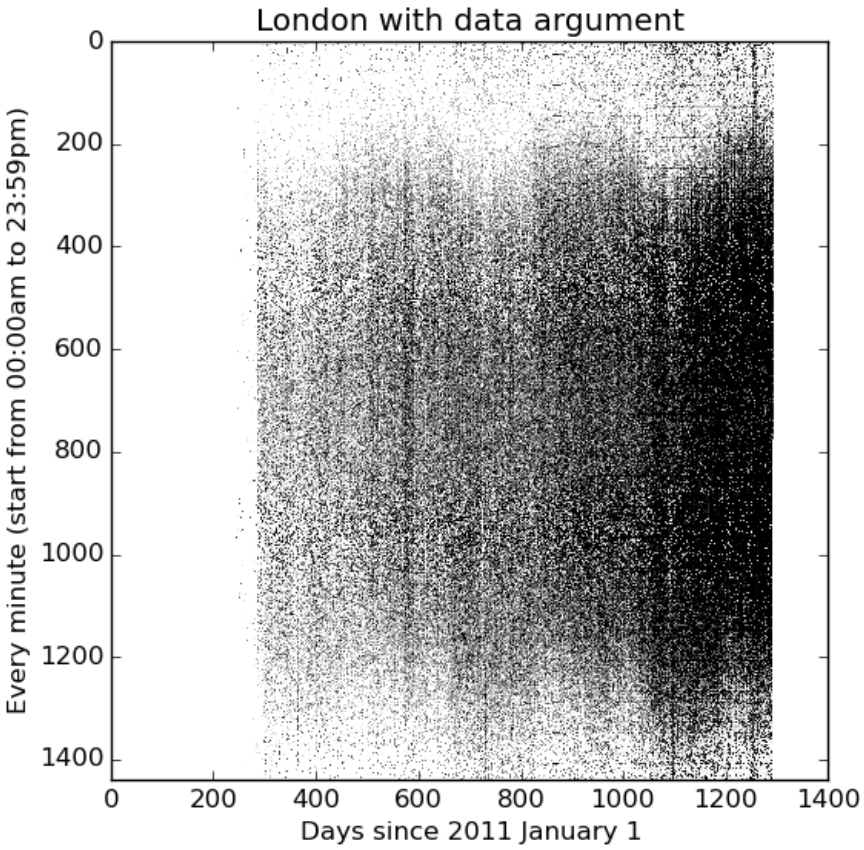


Figure 9: London every minute with data argumentation

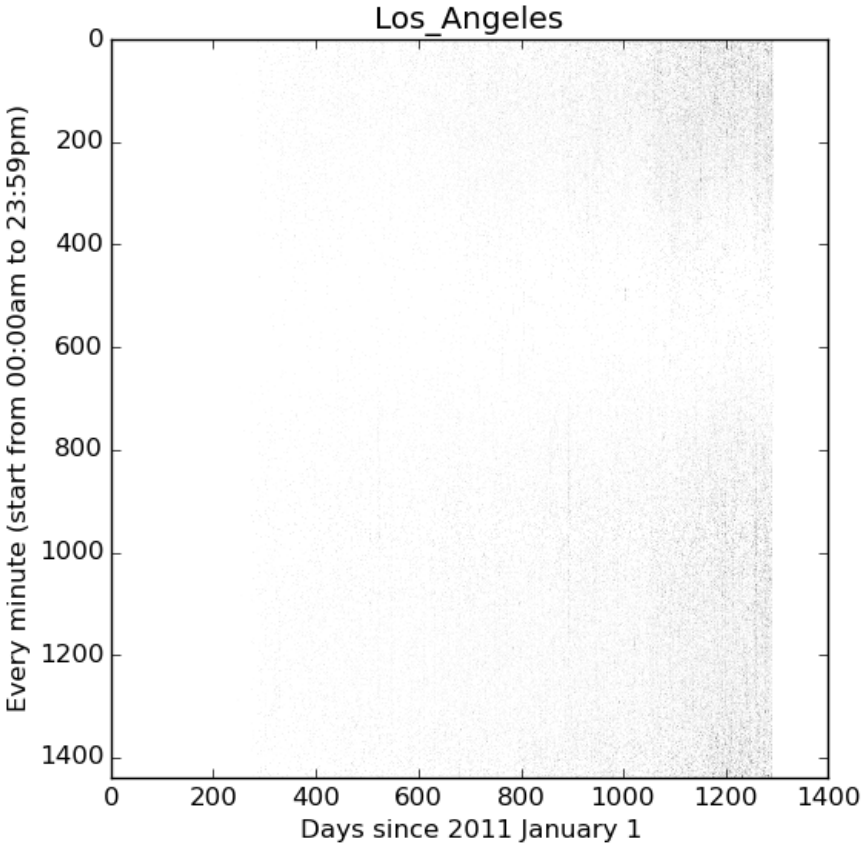


Figure 10: Los Angeles every minute without data argumentation

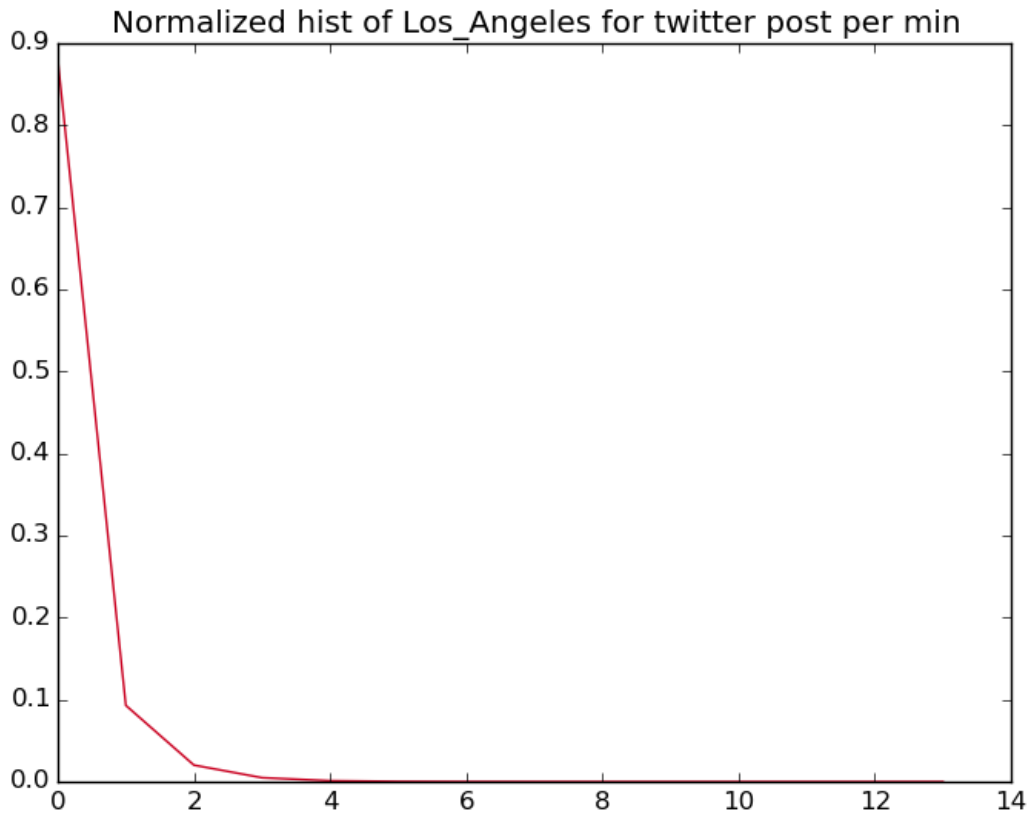


Figure 11: Histogram of Los Angeles

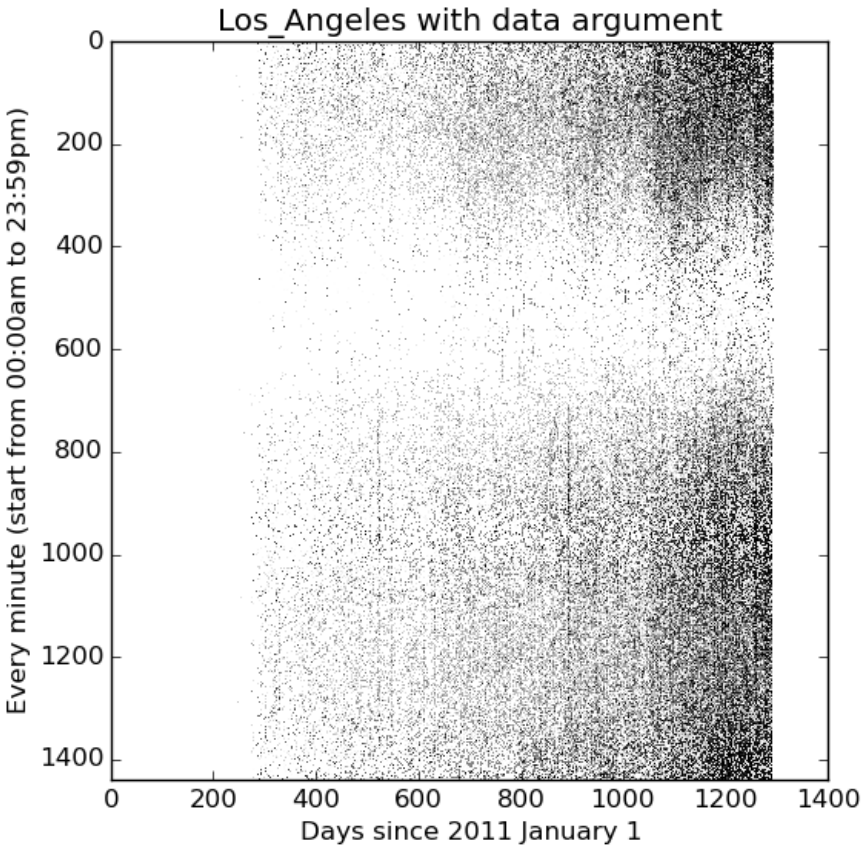


Figure 12: Los Angeles every minute with data argumentation

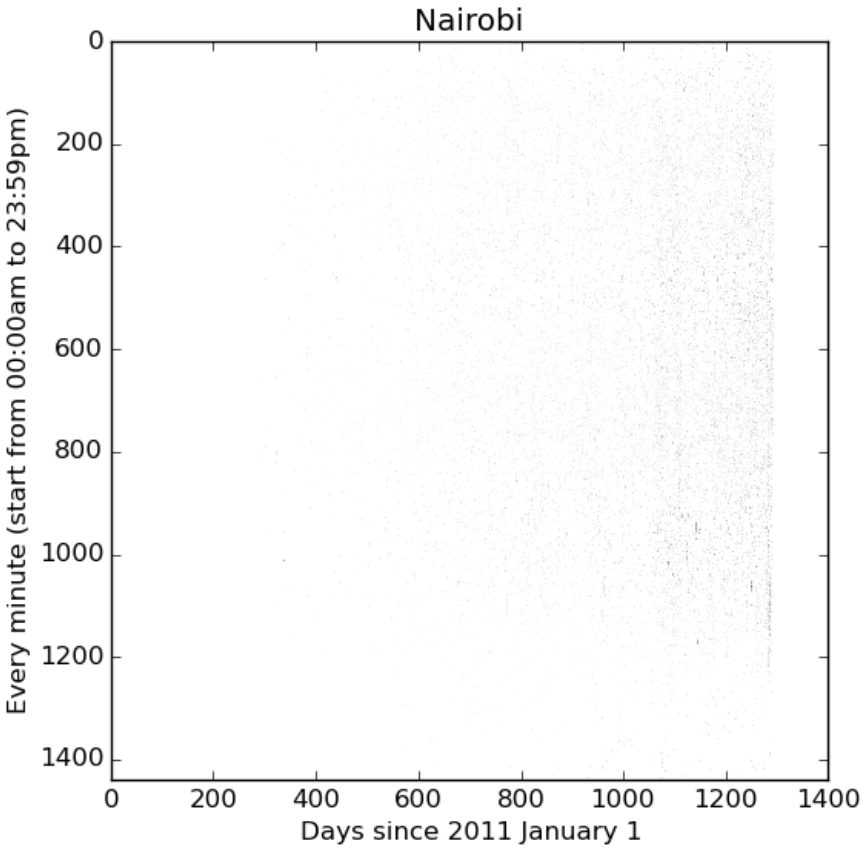


Figure 13: Nairobi every minute without data argumentation

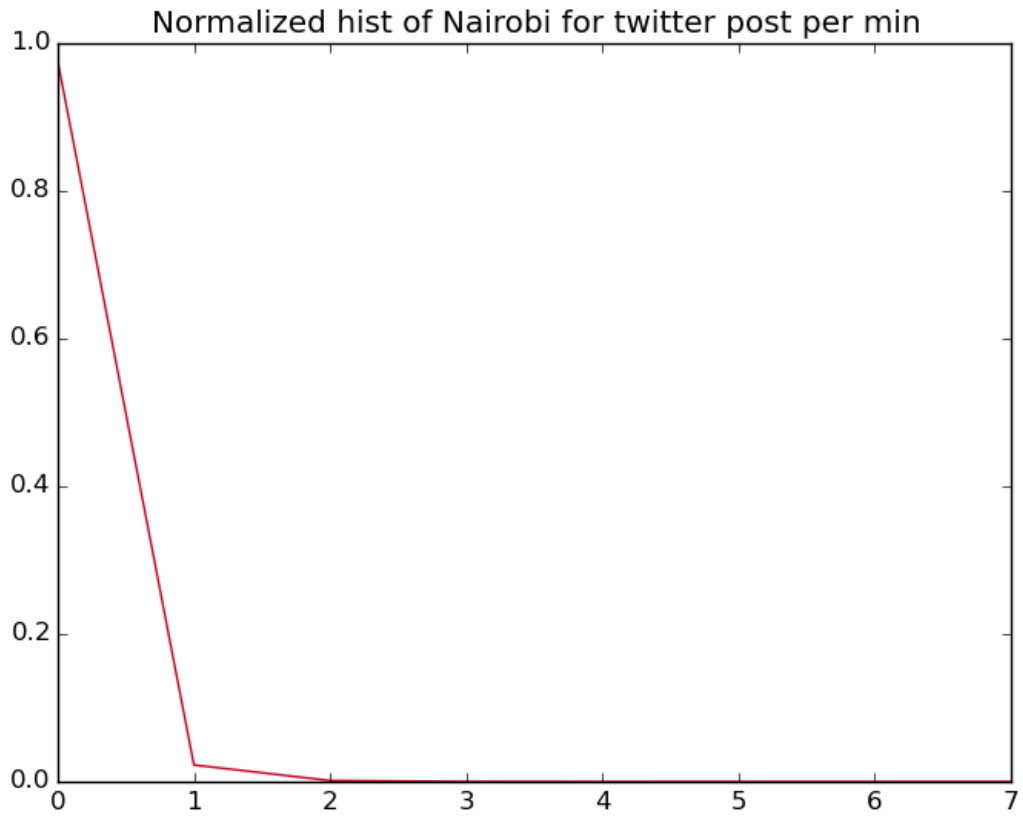


Figure 14: Histogram of Nairobi

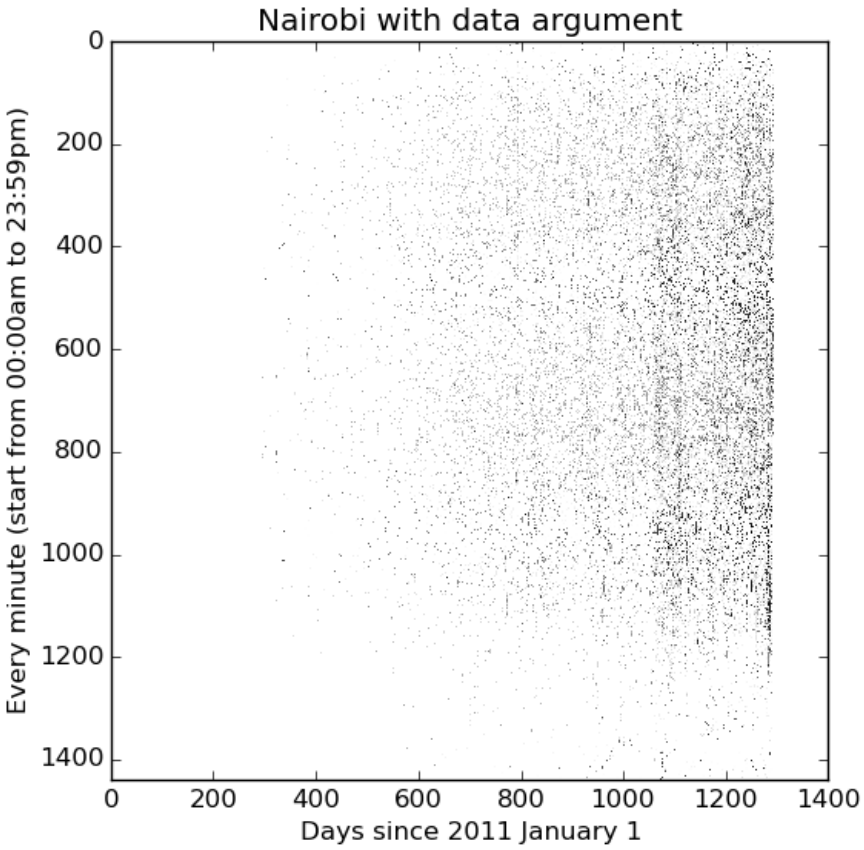


Figure 15: Nairobi every minute with data argumentation

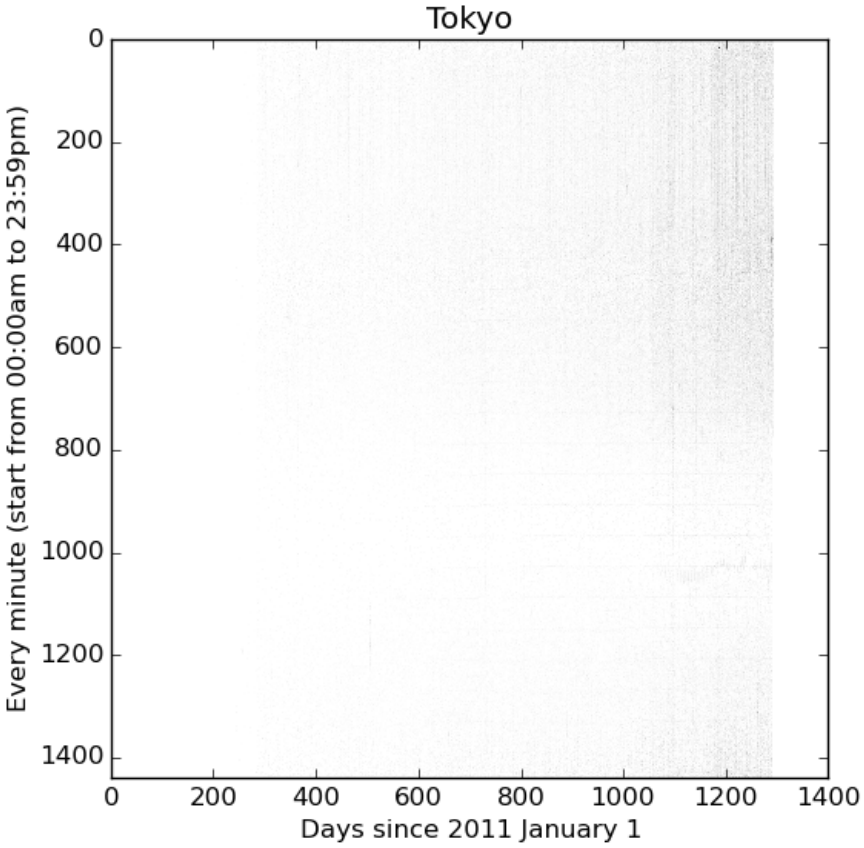


Figure 16: Tokyo every minute without data argumentation

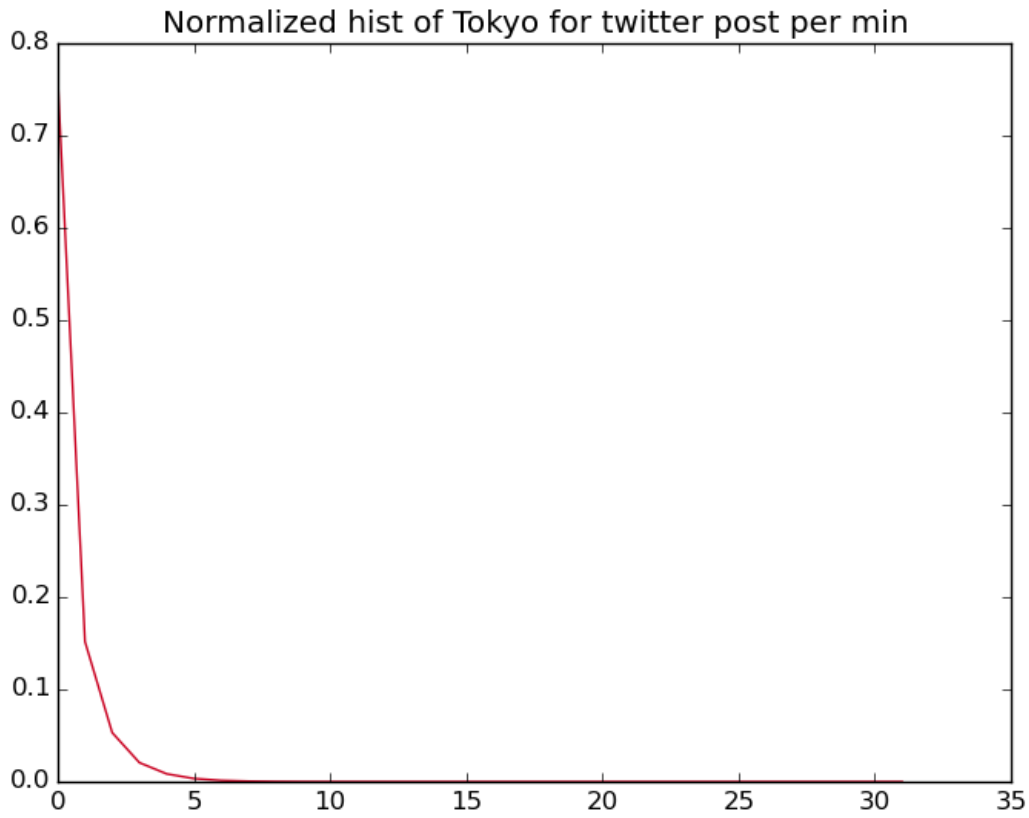


Figure 17: Histogram of Tokyo

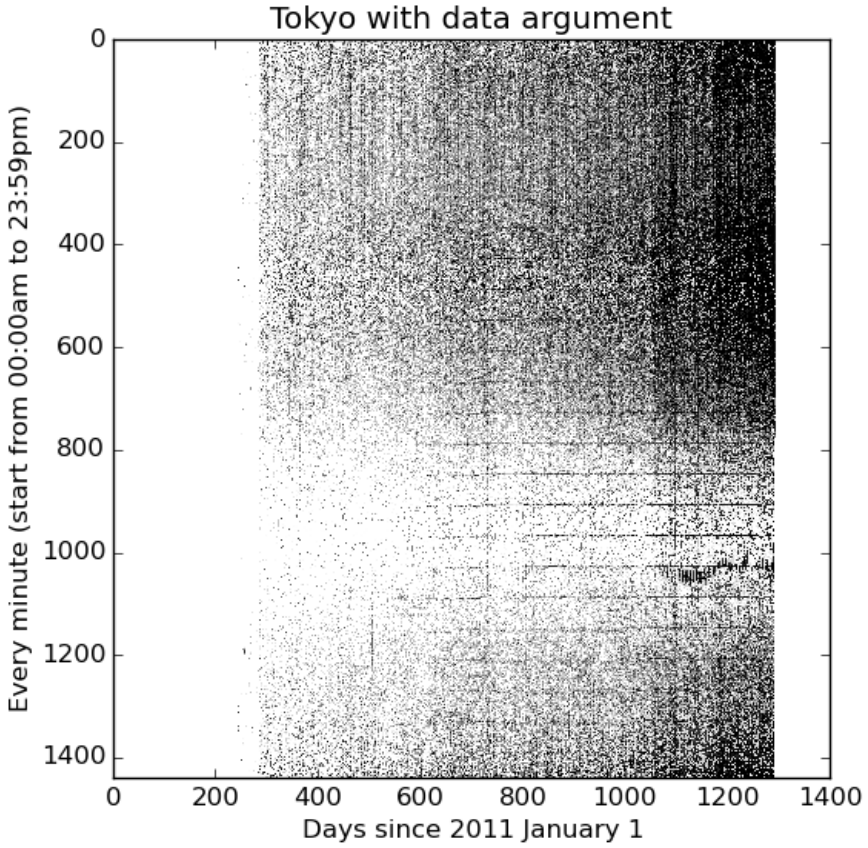


Figure 18: Tokyo every minute with data argumentation

## 3.2 Observe the twitter post every 10 minutes

The observation results based on 30 minutes time slot are shown in this section. From the argumentation result we can see the pattern of each city is even more clear. It is not done yet. A further analysis is shown in the following section.

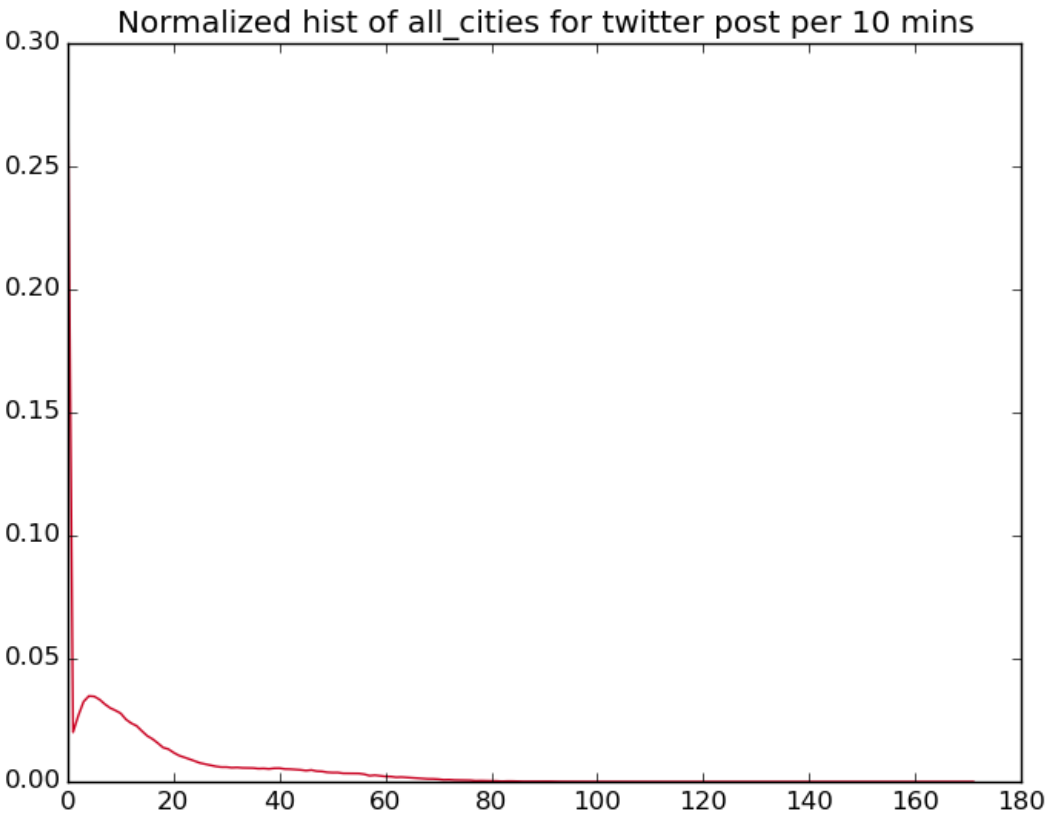


Figure 19: Histogram of five cities when the observation time slot is every one minute

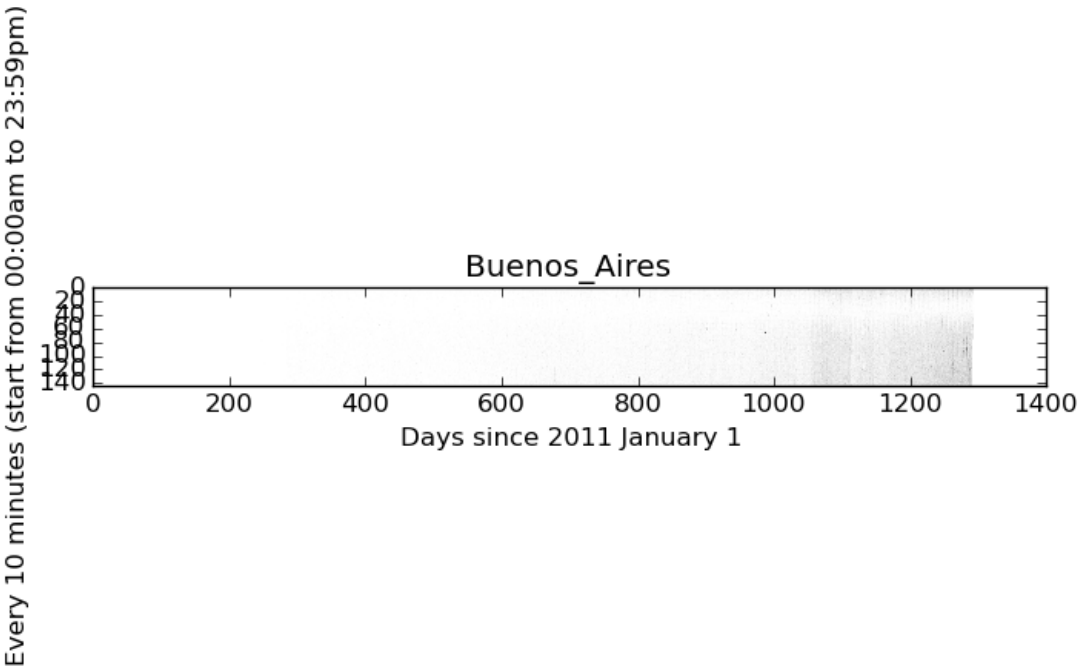


Figure 20: Buenos Aires every 10 minutes without data argumentation

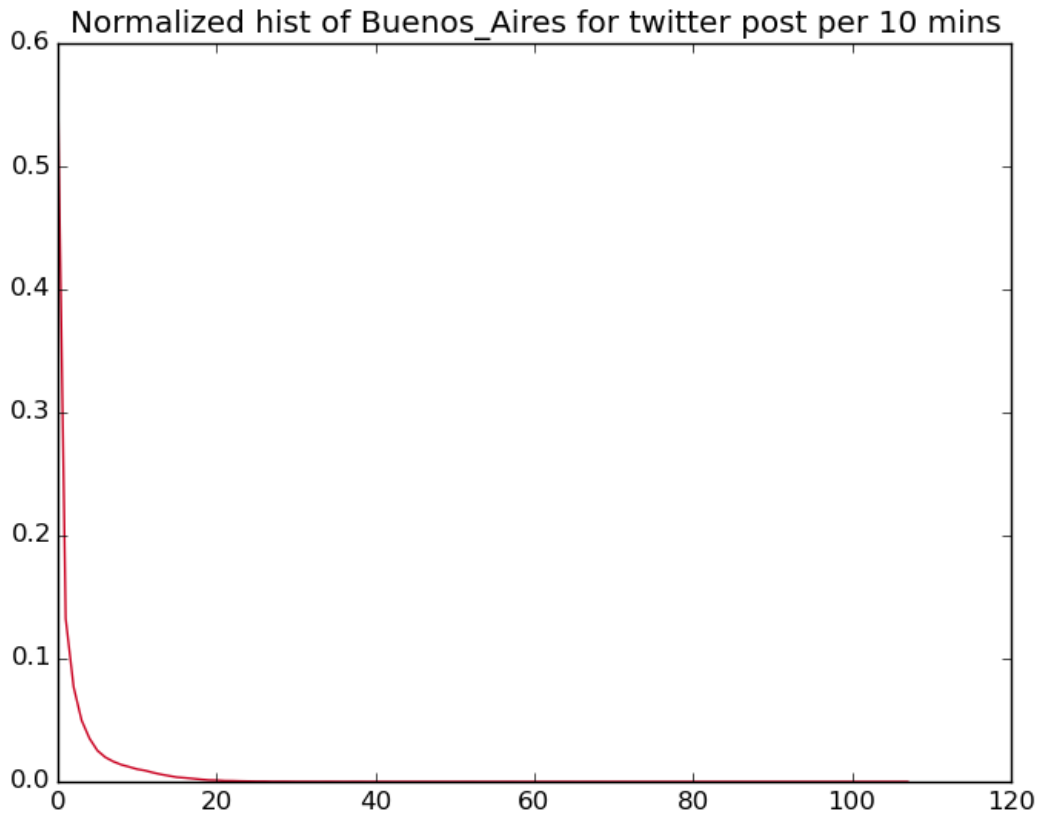


Figure 21: Histogram of Buenos Aires

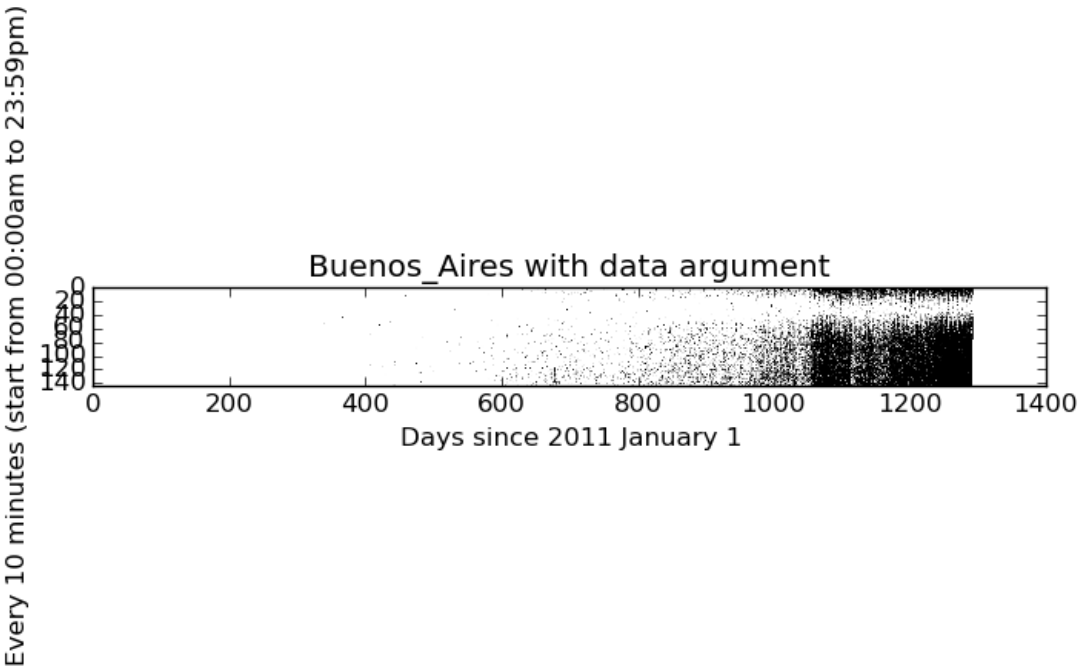


Figure 22: Buenos Aires every 10 minutes with data argumentation

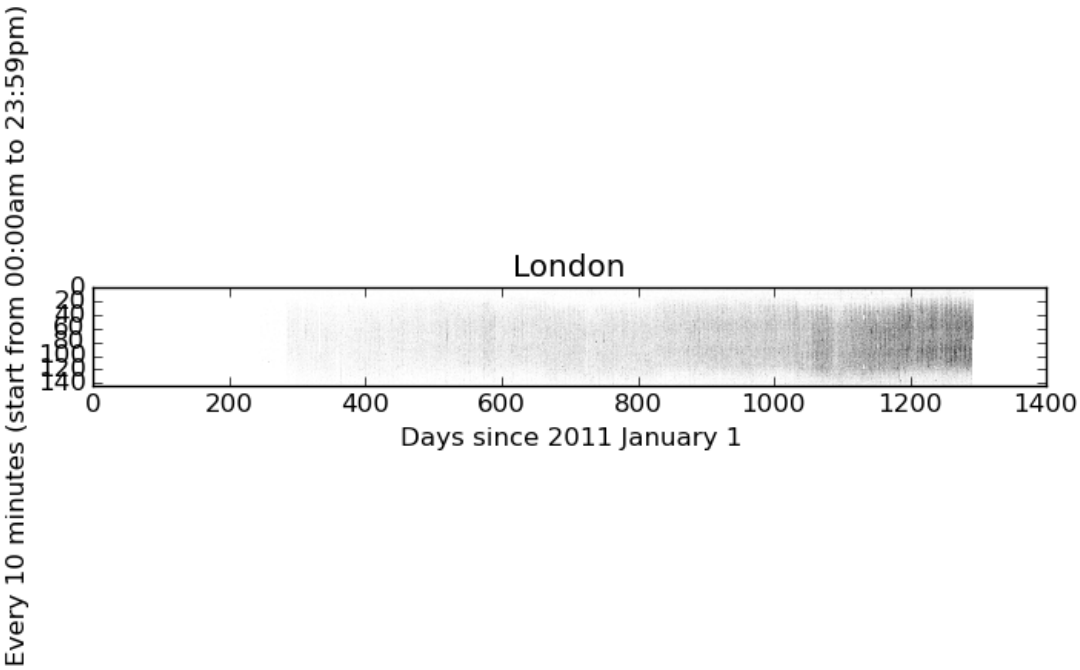


Figure 23: London every 10 minutes without data argumentation

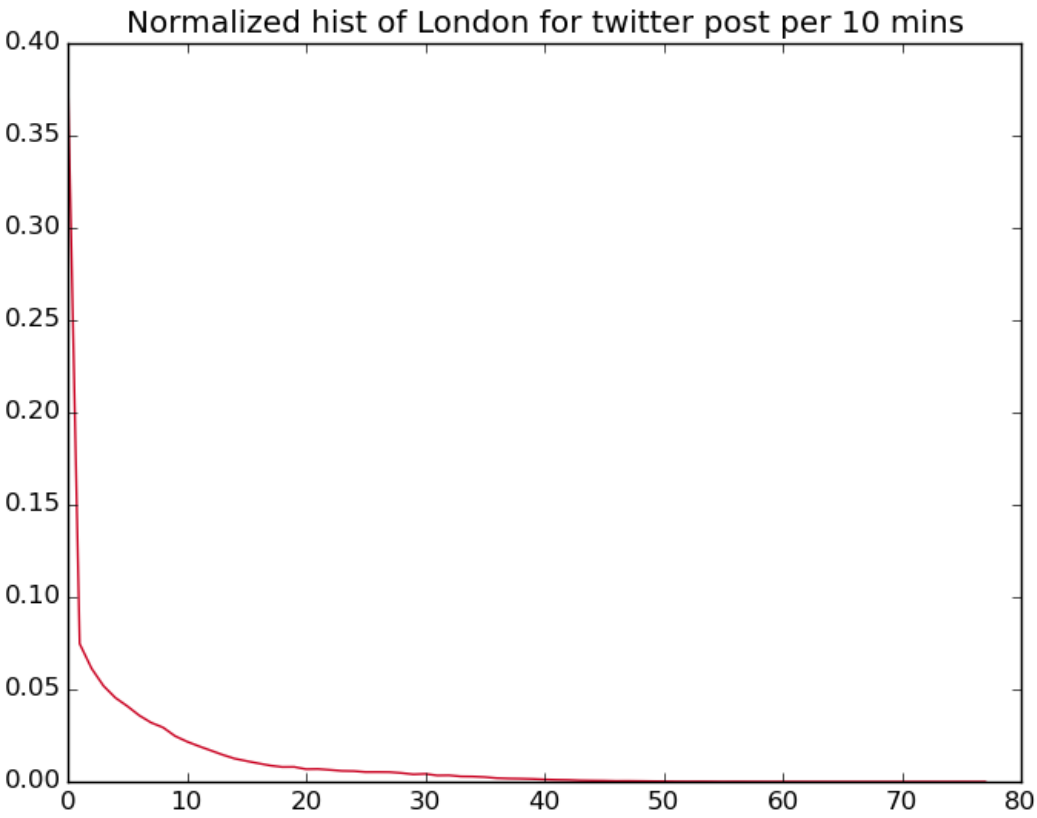


Figure 24: Histogram of London

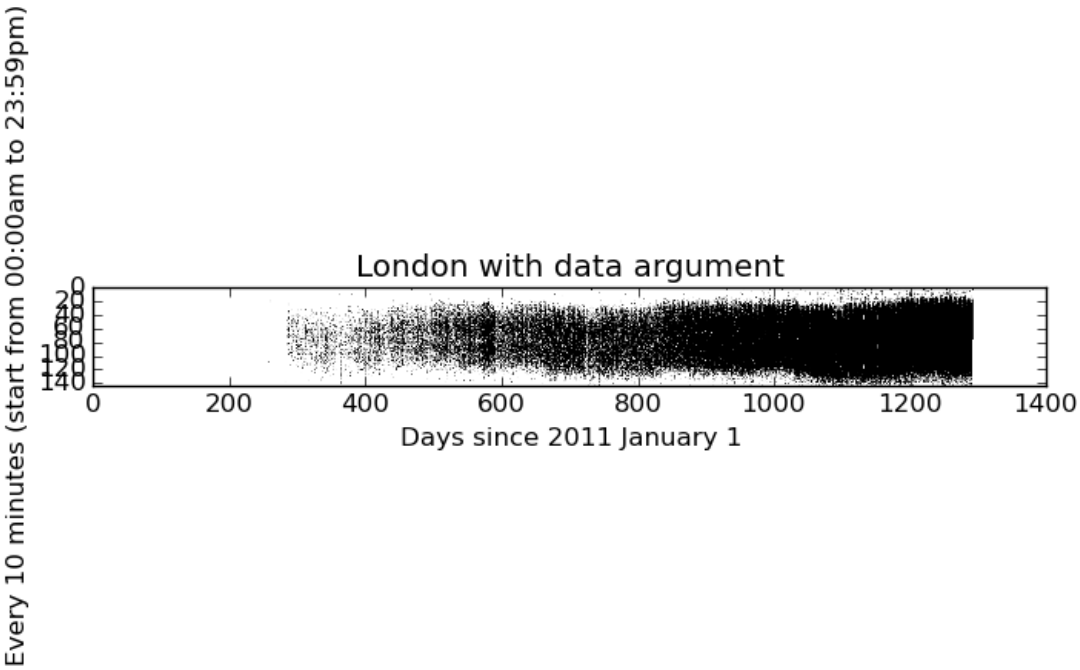


Figure 25: London every 10 minutes with data argumentation

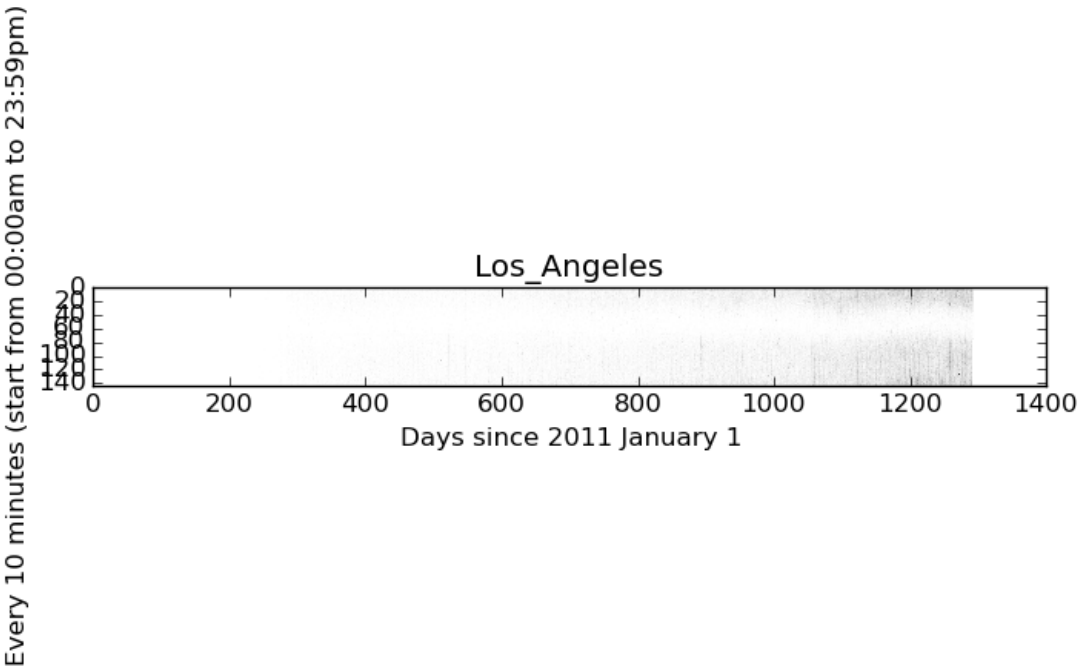


Figure 26: Los Angeles every 10 minutes without data argumentation

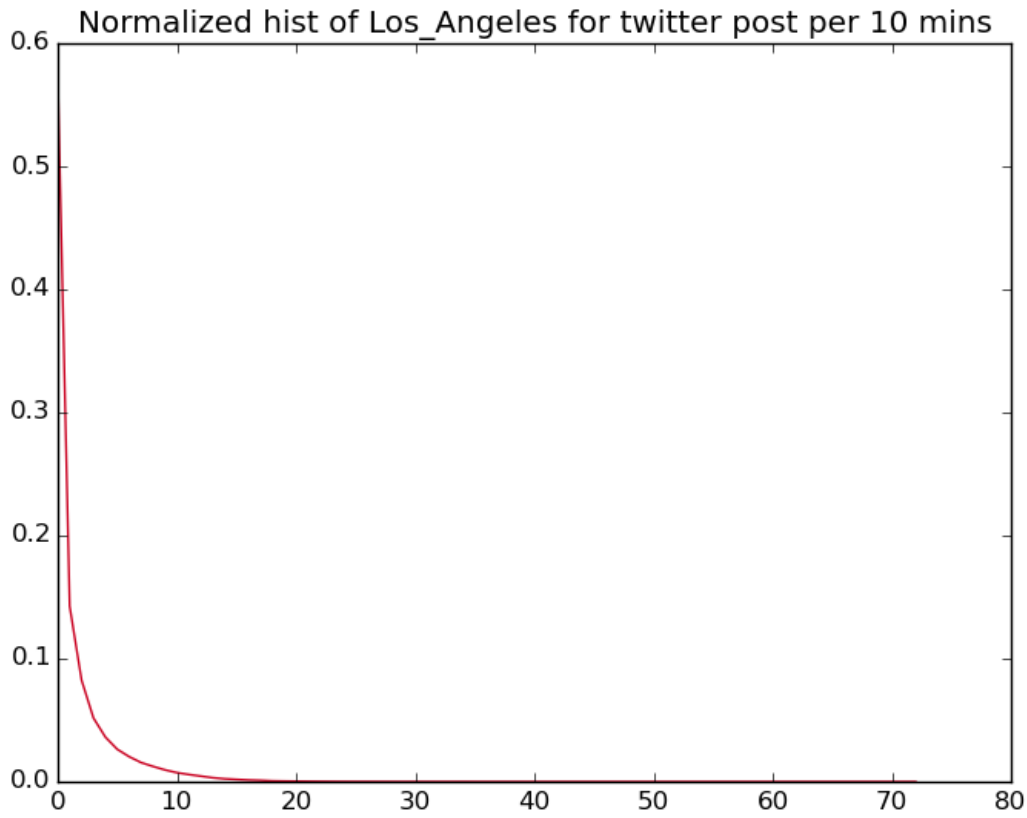


Figure 27: Histogram of Los Angeles

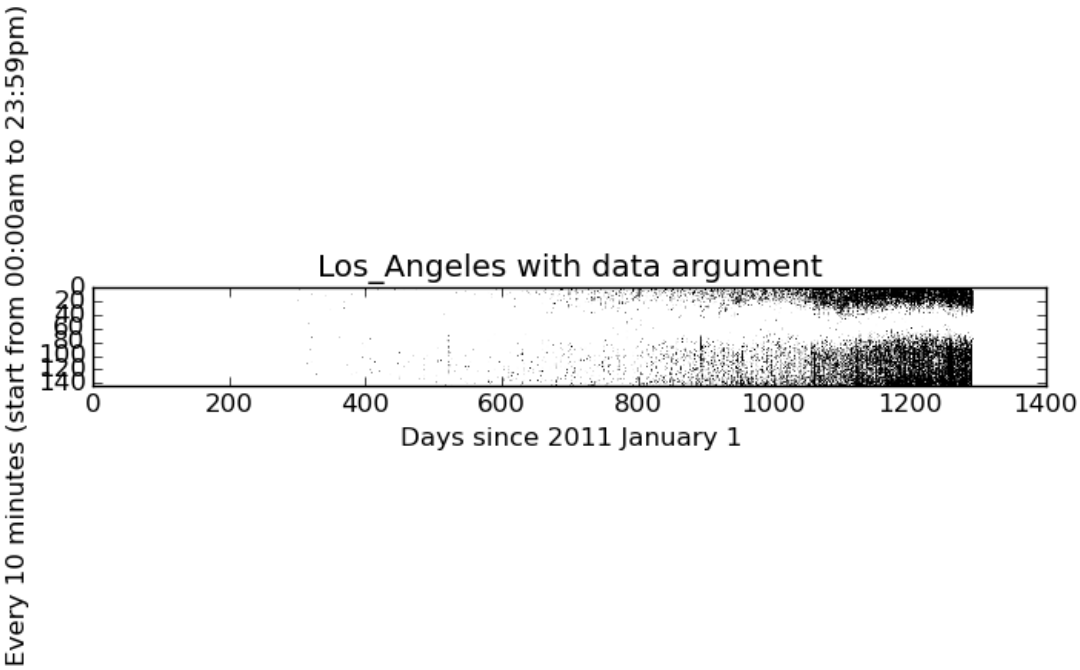


Figure 28: Los Angeles every 10 minutes with data argumentation

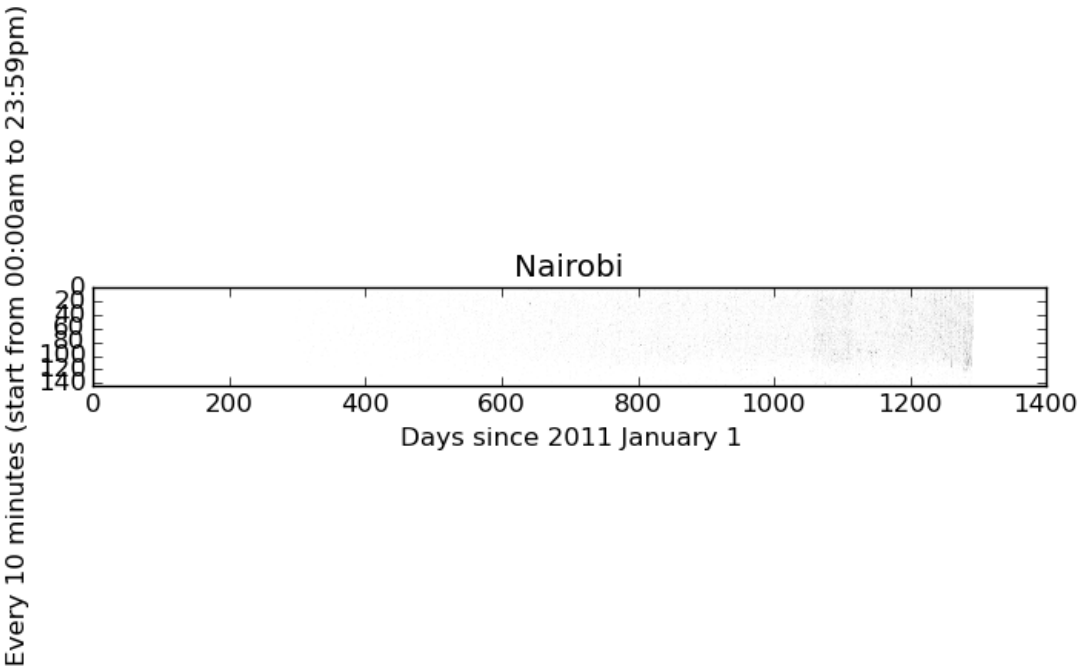


Figure 29: Nairobi every 10 minutes without data argumentation

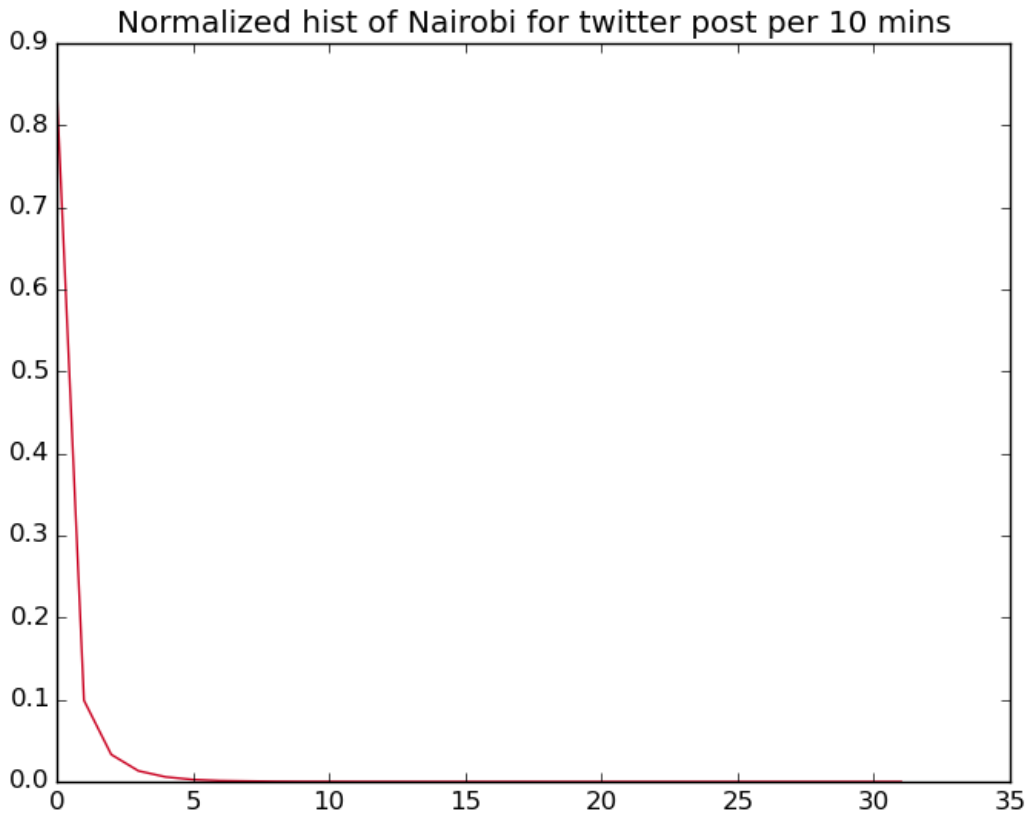


Figure 30: Histogram of Nairobi

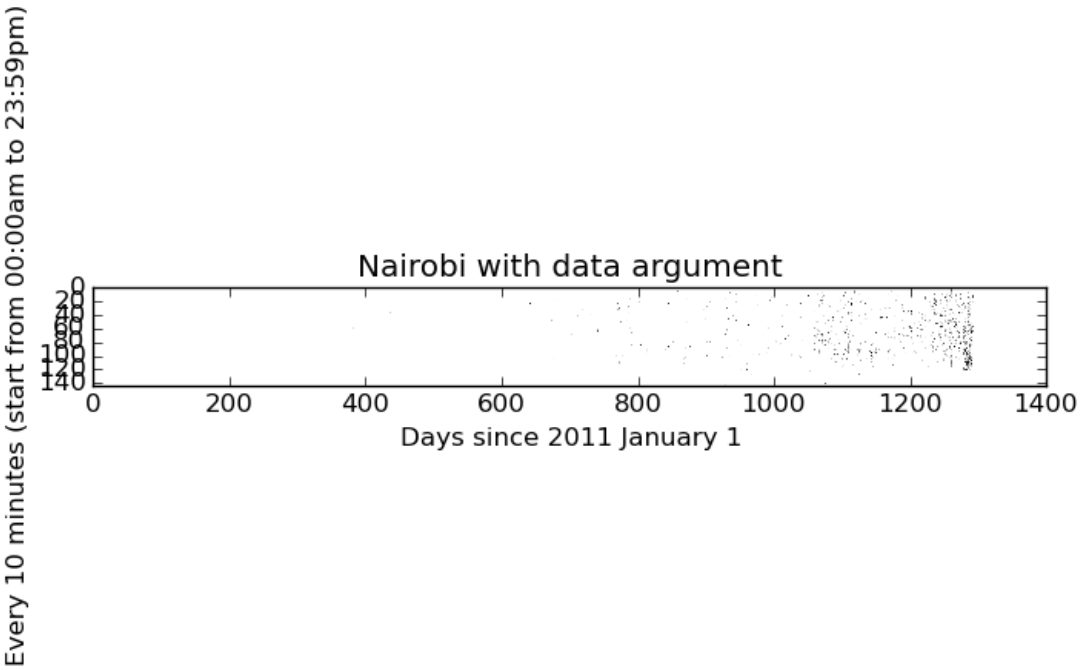


Figure 31: Nairobi every 10 minutes with data argumentation

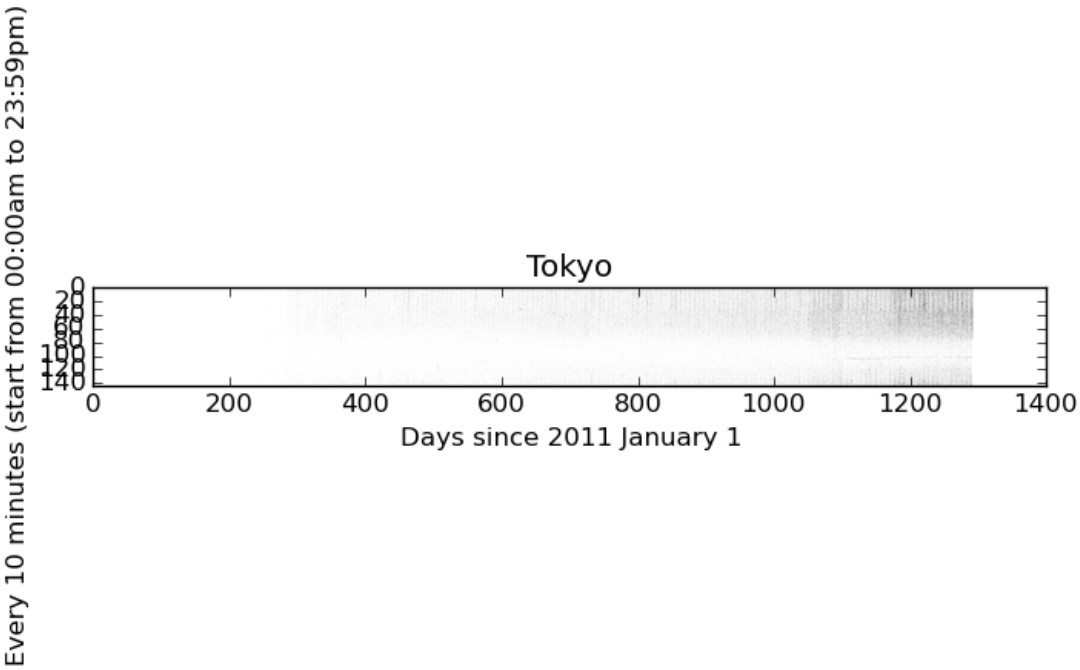


Figure 32: Tokyo every 10 minutes without data argumentation

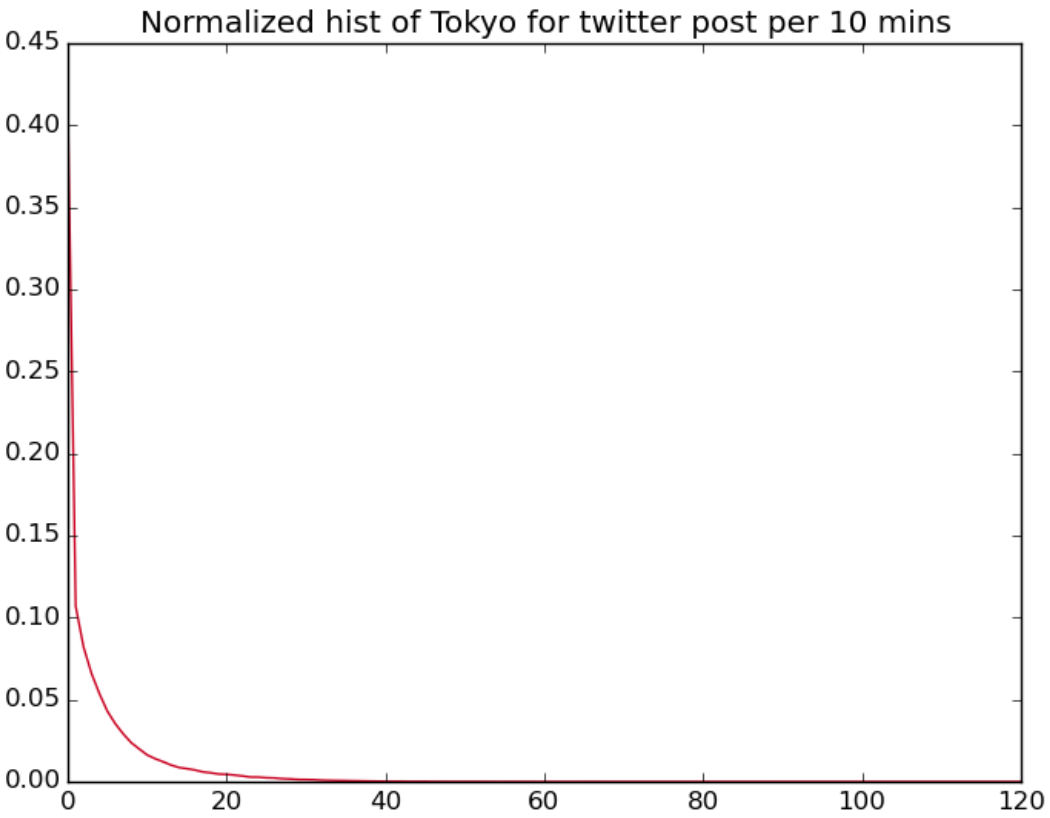


Figure 33: Histogram of Tokyo

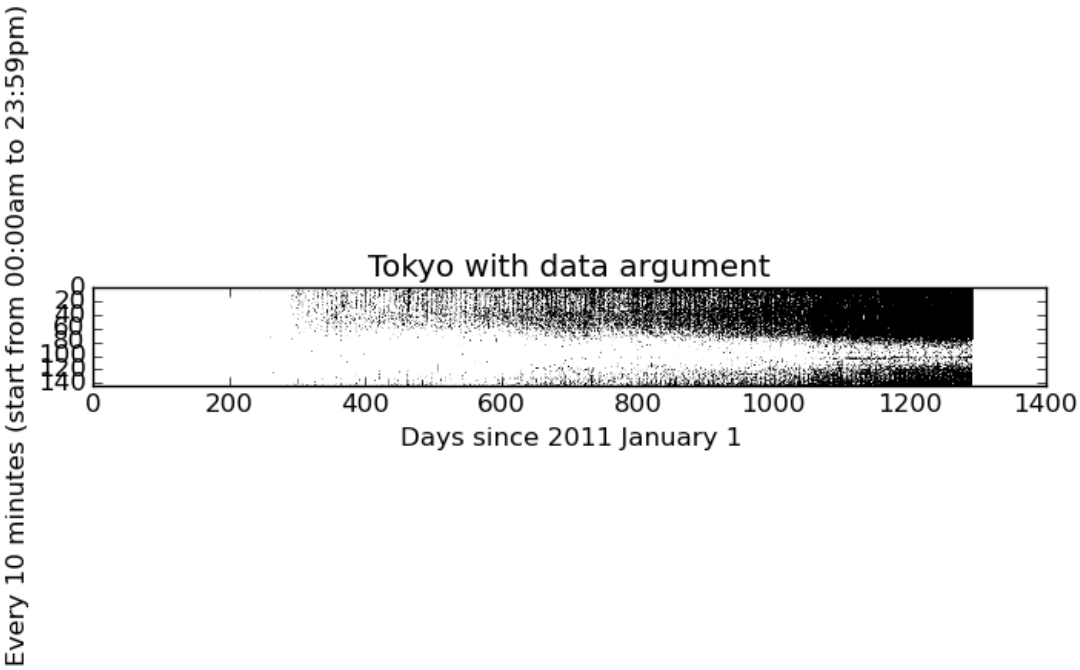


Figure 34: Tokyo every 10 minutes with data argumentation

## 3.3 Observe the activity level of a city based on the twitter post number every 10 minutes

The activity level is defined as: $twitterPostNumberPerTimeSlot/totalPostNumber4theWholeL$
A specified threshold is chosen to do the data argumentation. The results are given in the following figures.

From those results, we can see when the observations number becomes a large enough value, the status is more stable which can be well explained by the law of large number. Based on this, the last step will be done to get the sleeping time of a specified city which is doing the analysis based on the last 50 days data. The last 50 days data is more stable as the data size is big

enough and also the activation model of a specified city will not be changed within 50 days. Then the data analysis based on the last 50 days generated the final result shown in figure 2. By this step, we can have a better understanding about the sleeping time of a city.

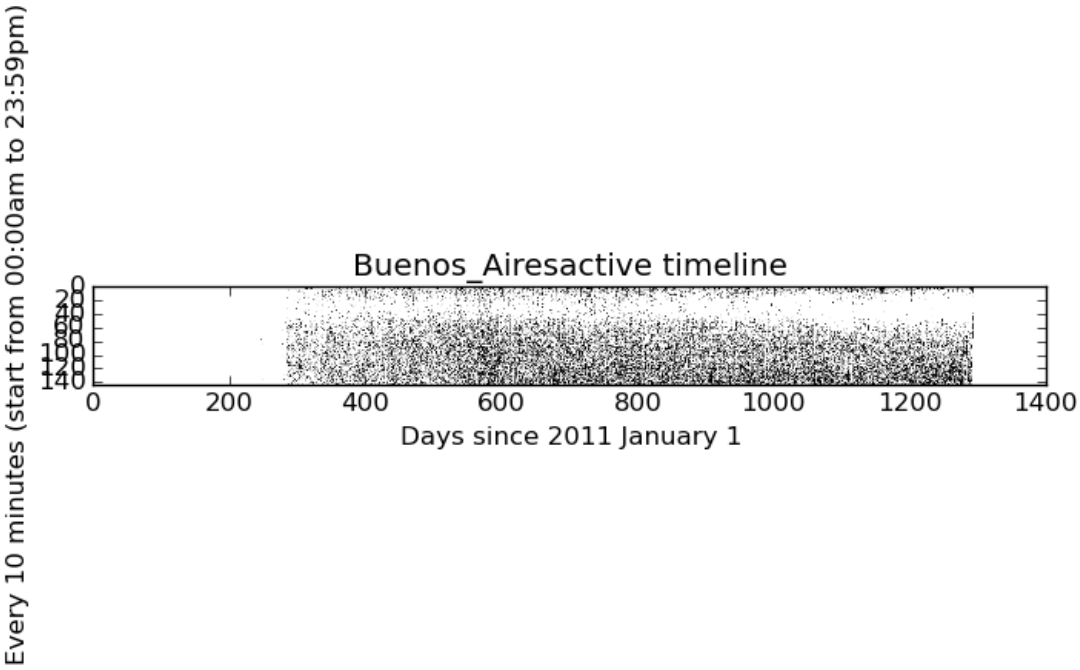


Figure 35: Buenos Aires activation ratio based on every 10 minutes twitter post data with data argumentation

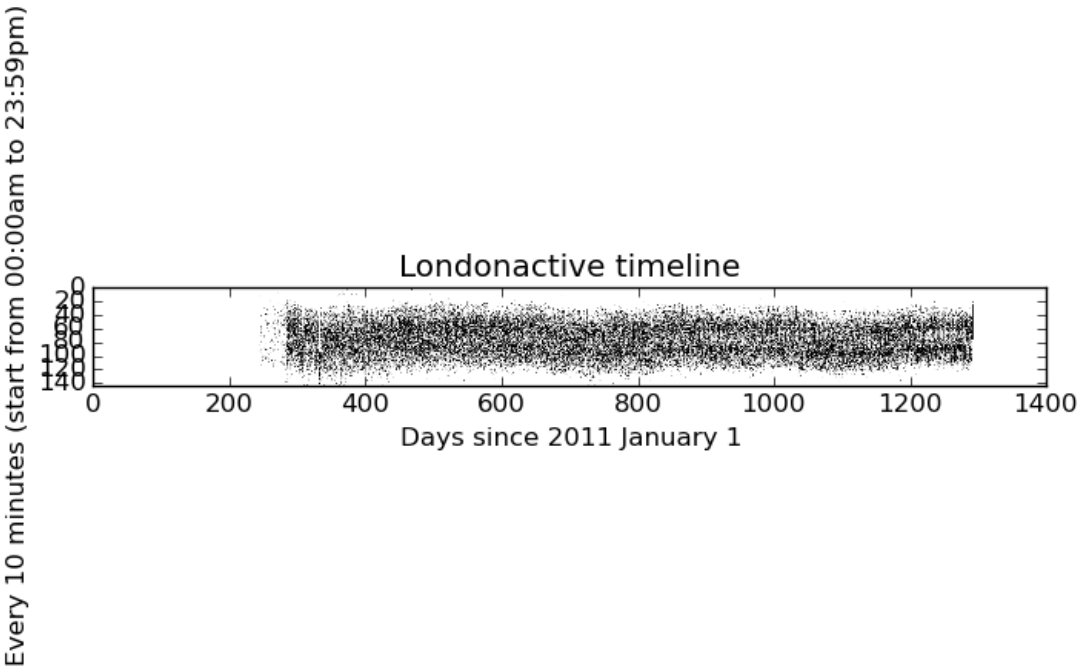


Figure 36: London activation ratio based on every 10 minutes twitter post data with data argumentation

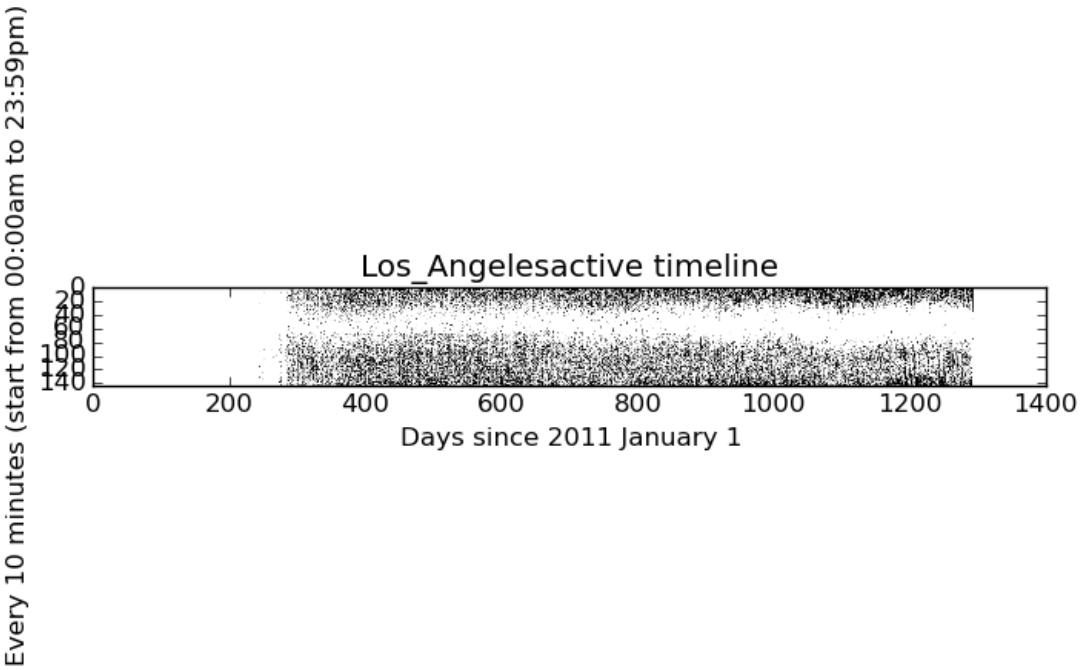


Figure 37: Los Angeles activation ratio based on every 10 minutes twitter post data with data argumentation

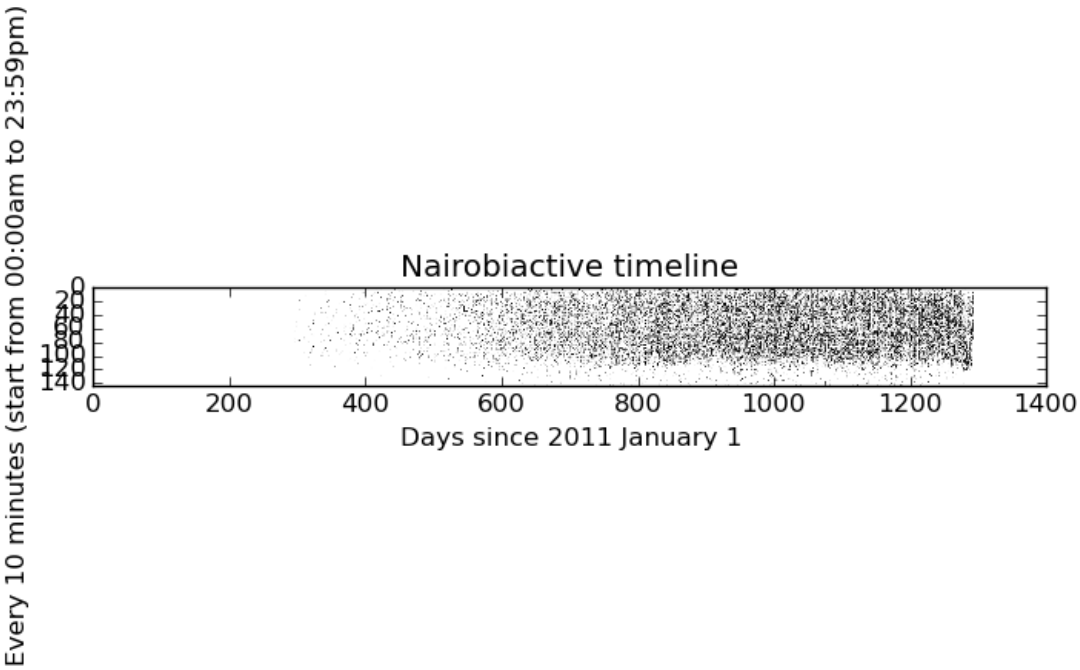


Figure 38: Nairobi activation ratio based on every 10 minutes twitter post data with data argumentation

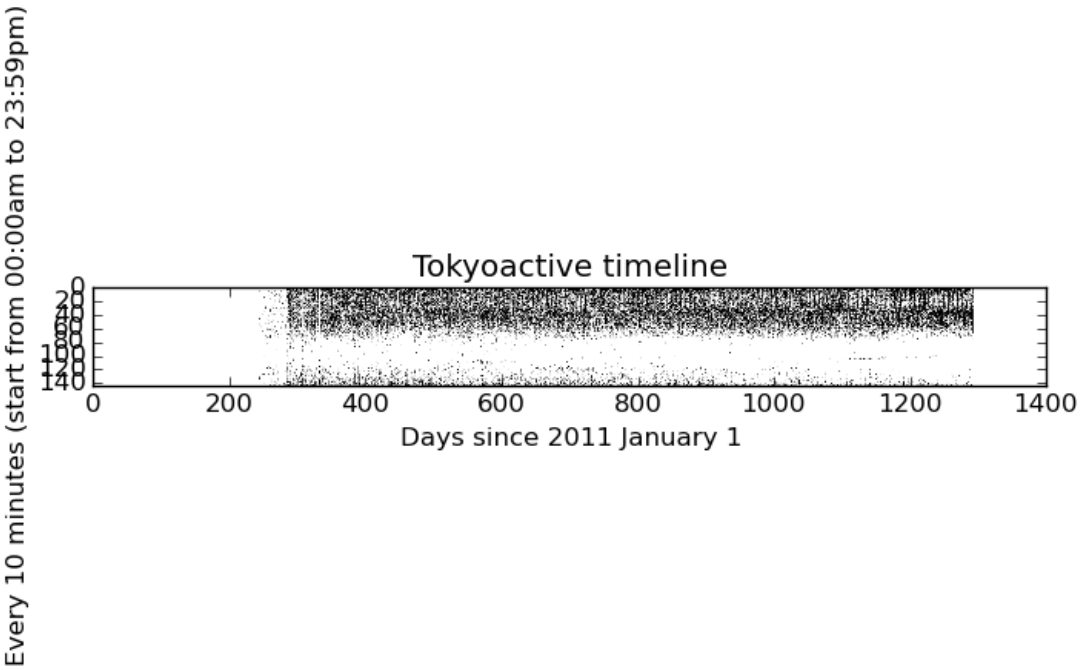


Figure 39: Tokyo activation ratio based on every 10 minutes twitter post data with data argumentation

# 4 Acknowledgement

Thanks for professor Lev Manovich sharing the twitter data from the five cities. This article is part of course work of professor Lev Manovich's "Data Visualization" course at the Graduate Center, CUNY, during Spring 2017. Thanks for Professor Lev Manovich's inspiration on building a good taste on data visualization such as how to choose the color, why does the detail of a good visualization arts matters and so on. Thanks again. As the twitter data belongs to professor Lev Manovich's Cultural Analytics Lab. This article will not be posted to public without the permission of the Cultural Analytics Lab.