The attraction basin of the New York City airports.

Marc Fuster Rullan, Riccardo Gallotti, Jose J. Ramasco
Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB)
Campus Universitat de les Illes Balears, E-07122 Palma de Mallorca, Spain

Abstract

2013 NYC taxi data has been cleaned and analyzed in order to understand and model the decision behavior of NYC citizens when facing the choice between the three airports that are serving the city. A comparison with Google data has proved that Taxi data is a good estimator to airport people flow. Visualization tools such as colormaps have been used to understand the human behavior. A discrete multichoice decision model has predicted the value of time for NYC citizens.

1 Introduction

Since the wide use of devices with localization such as smartphones or GPS navigators, large volumes of individual trajectory data have become open source. Due to this large amount of data, social topics such as mobility have emerged with a great impact among researchers. Airports in all around the world have a huge economic impact on society, thousands of people go to airports everyday both to take a plane and to work there. The transport to the airport is a problem that most of people face sometime. Moreover, citizens of populations with more than one airport, such as NYC, Paris or London, have to choose between the airports. Tickets price, trip time and trip cost are the most important but not the only factors that citizens have to consider to choose between airports. The main objective of this project is to understand and try to model the decision behavior of NYC citizens.

NYC has three near airports that are John F. Kennedy International Airport (JFK from now on), LaGuardia Airport (LaG) and Newark Liberty International Airport (NEW). JFK is the international airport and the one with most passengers (60 M per year), it also has a good public transport connection by train with Manhattan. LaG allows only national flights (24 M). LaG is also the closest to Manhattan but the public transport connection is by bus. Newark is both national and international but is has less passengers than JFK (35M). Before continuing reading I strongly recommend to look for a NYC picture to truly understand the airports' properties.

2 Materials and methods

For this analysis it has been used the dataset of NY taxis extracted from [1]. It contains all taxi trips in NY state in from 2010 until 2013. It has only been used the 2013 dataset of over 173 million trips to reduce the computational time without loosing much precision. According to [1], roughly a 7.5% the dataset has errors such as GPS coordinates of (0,0) and impossible velocities. The dataset is divided in two types, the *trip_data.csv* type and the $trip\_fare.csv$. Both types of file were ordered by the pickup date. The dataset's fields that have been used for the analysis are:

- **pickup datetime** in mm-dd-yyyy hh24:mm:ss EDT
- **dropoff datetime** in mm-dd-yyyy hh24:mm:ss EDT
- **pickup longitude**
- **pickup latitude**
- **dropoff longitude**
- **dropoff latitude**
- **tip amount** in USD
- **total amount** in USD, includes fare, taxes, tolls and tips.

The language used to analyze the data has been Python 2.7 due to the existence of the library Pandas [2] that provides an easy use of data structures. Both Numpy [3] and Matplotlib [4] have been used to make easy calculus and plots.

The dataset has a field that corresponds to the trip time. However, it is often stored in minutes and others on seconds. Therefore, the safest way to compute the trip time is by subtracting the pickup time to the drop off time as [1] explained. To convert the datetime format of the dataset to a useful format it has been used the library Datetime [5].

The first step of the analysis must be the cleaning of the data. The most clever idea is to create a new copy of the data filtering the GPS coordinates from the pickup and the drop off at least in the NY area. From now on, the criteria to delimit areas will be rectangles due to both its simplicity and the less computational time. The rectangles used in the analysis are described in 1.

	Longitude right (º)	Longitude left (º)	Latitude up (º)	Latitude down (º)
NY	-73.7353	-74.215056	40.915532	40.5936
JFK	-73.7672	-73.798118	40.6451629	40.63548
LaG	-73.858559	-73.883579	40.778471	40.765763
NEW	-74.169956	-74.194675	40.701687	40.681513

Table 1: Rectangles GPS coordinates of different locations. NY does not cover all the state, just the NYC proximity.

To geographically distribute the properties of taxi trips it has been used two different methods. The hexbins python function [6] which automatically divides the space in regular hexagons and count how many events are located in each particular hexagon. The second method is the squared bins method. This second method consists in dividing manually the space and organizing the data by the bin using python dictionaries. The hexbins function is the best option when you only have to count events because it has more precision to diagonal patterns. However if it is need to analyze other features such as the average trip time or the average cost, one must use the manually square bins gridding.

To get the index, using the manual method, of a certain position x between a maximum value (max) and a minimum value (min) with a certain number of linear divisions of space (div) it is need to use the formula 1

$$index = floor\left(div\frac{x-min}{max-min}\right) \tag{1}$$

This particular project needs a 2D division of space. Therefore each point has an X_{index} and a Y_{index}. In order to easily store the data, it has been used a $div = 100$ to create a dictionary of 10000 entries saved by the format $'8703'$ that corresponds to $X_{index} = 87$ and $Y_{index} = 03$.

In order to model people's decision it has been used the discrete multichoice theory developed by the Nobel laureate Daniel McFadden [7]. The basic model used contains an Utility variable 2.

$$U = -a(b \cdot T + C) \tag{2}$$

where T is the trip time (in minutes) and C is the trip cost (in USD). b is a parameter called "The cost of time" that will be estimated. The parameter a will be explained in the following paragraph. The Utility is used to compute the decision probability. For a three option choice, the probability can be computed as 3.

$$P_i = \frac{e^{U_i}}{e^{U_1} + e^{U_2} + e^{U_3}} \tag{3}$$

The equation 3 is affected by the order of magnitude of U. If $|U|$ is small, the probability gets closer to 0.33 and the map will be smoother. If $|U|$ is high, the probability increases or decreases faster.

In order to estimate the parameters a and b, it has been created an error per bin $e(x, y; a, b)$ obtained by subtracting the real probability and the model probability. This error per bin has

been weighted by the number of events of the cell (N_{xy}) by the equation 4 in order to give more importance to bins with more trips.

$$E(a,b) = \frac{\sum_{x,y} N_{xy} e(x,y;a,b)}{\sum_{x,y} N_{xy}} \tag{4}$$

The parameters a and b that minimize the error will be the final results.

3 Results and discussion

3.1 Is taxi data a good estimator to airport flow?

There is not many data amount on private cars or public transport. Taxi data is fairly easy to get and well organized. To check if taxis describe the flow of people to airports, it has been compared with the Google presences data of the JFK airport. This data has been extracted by inspecting element on the browser and copying it manually. The result is in figure 1.

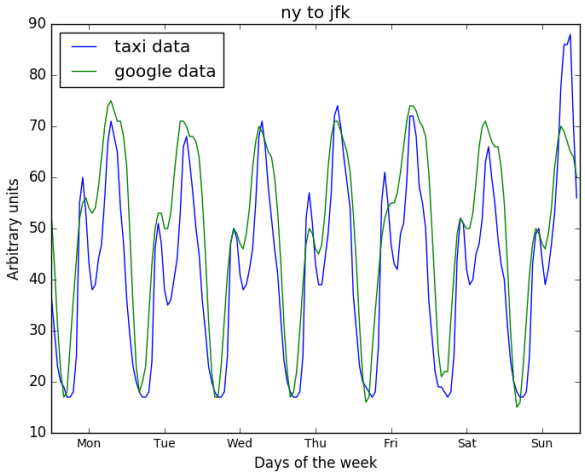

Figure 1: Not scaled comparison between Google data presences and JFK airport taxi arrivals.

It is important to say that Google data had no units, to plot the figure it has been scaled. However, the figure shows that taxi data is a good estimator. There is a small difference on Sundays that may be due to the fact that public transport have lower frequencies on Sunday so more taxis are taken.

3.2 A first approach to taxi trips

The first steps of the projects are focused on understanding the main features of taxi trips in NYC. First of all it is important to understand the location difference of pickups and dropoffs. The figure 2 shows the difference in the number of pickups minus the dropoffs for each location. This figure has been obtained by a 1 million random trips.

The red zones show more pickups than dropoffs. It is easy to see that pickups tend to be more located in important streets and dropoffs are more distributed. This fact is mainly due to the more density of free taxis driving in main streets.

A different general feature of taxi trips that has been analyzed is the tip percentile distribution. It has been used two scenarios. The left one has been obtained from a 1 million random sample. The right one has been obtained from all taxis that have gone to JFK airport with a flatrate of 58.8 USD. For both datasets, roughly a 40% of the data contained no tip. This number may not reflect reality because it is possible that some taxi drivers do not record their tips. That is why all non tip trips have been deleted for both datasets. The results are shown in the figure 3

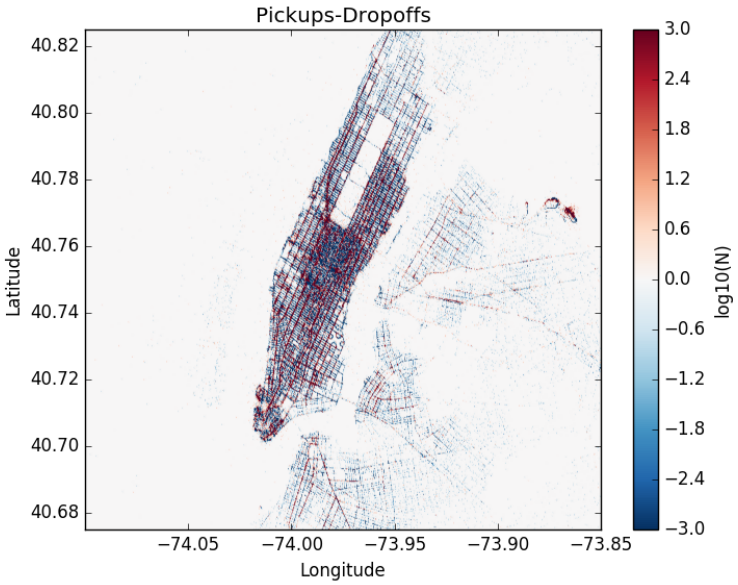

Figure 2: Number of pickups minus number of dropoffs. In logaritmic scale.

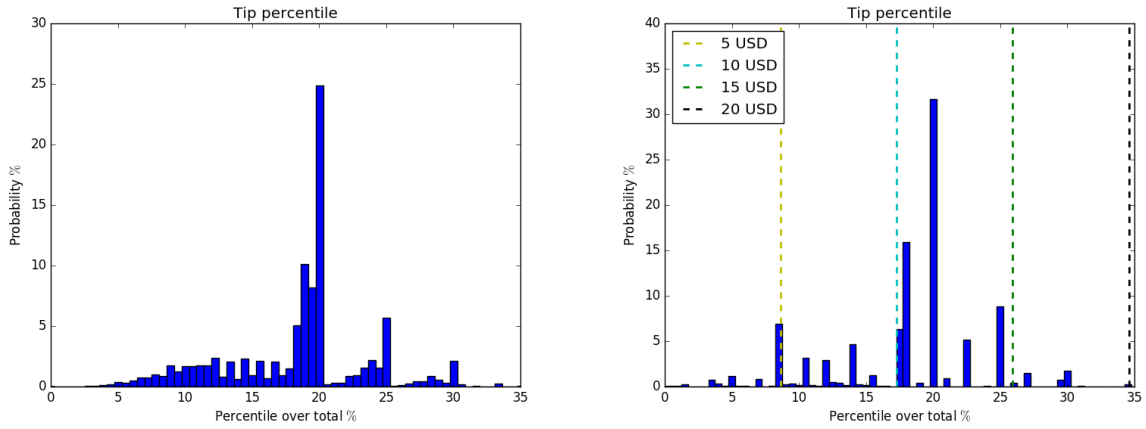

Figure 3: Tip percentile and for the right one the equivalent on USD due to the existence of the 58.8 USD flat rate.

Data shows that around 25% of trips leave a 20% tip and that number even increases in JFK trips. Some of the columns on the left image that have not a round percentile may have their explanation in making the fare amount plus tip a round number. Sometimes people also pay by card and leave a round tip as it can be seen on 5 and 10 USD lines on the right figure.

3.3 Understanding mobility to NYC airports

The dataset contains the information needed to make a person decide the best airport option related to his departure location. Applying the second algorithm described in 1 it has been created different maps that describe the avarage cost and the avarage time for the trips. Only were considered bins where at least had 3 trips to the corresponding airport. The figures obtained are in 4

As 4 shows, the best option for most of Manhattan and Brooklyn is LaGuardia airport for both its lowest time trip and lowest fare trip. This result is probably is due to the existence of a Manhattan flatrate to JFK and Newark's remoteness. Only nearest points to Newark or JFK do not have LaGuardia as the best choice.

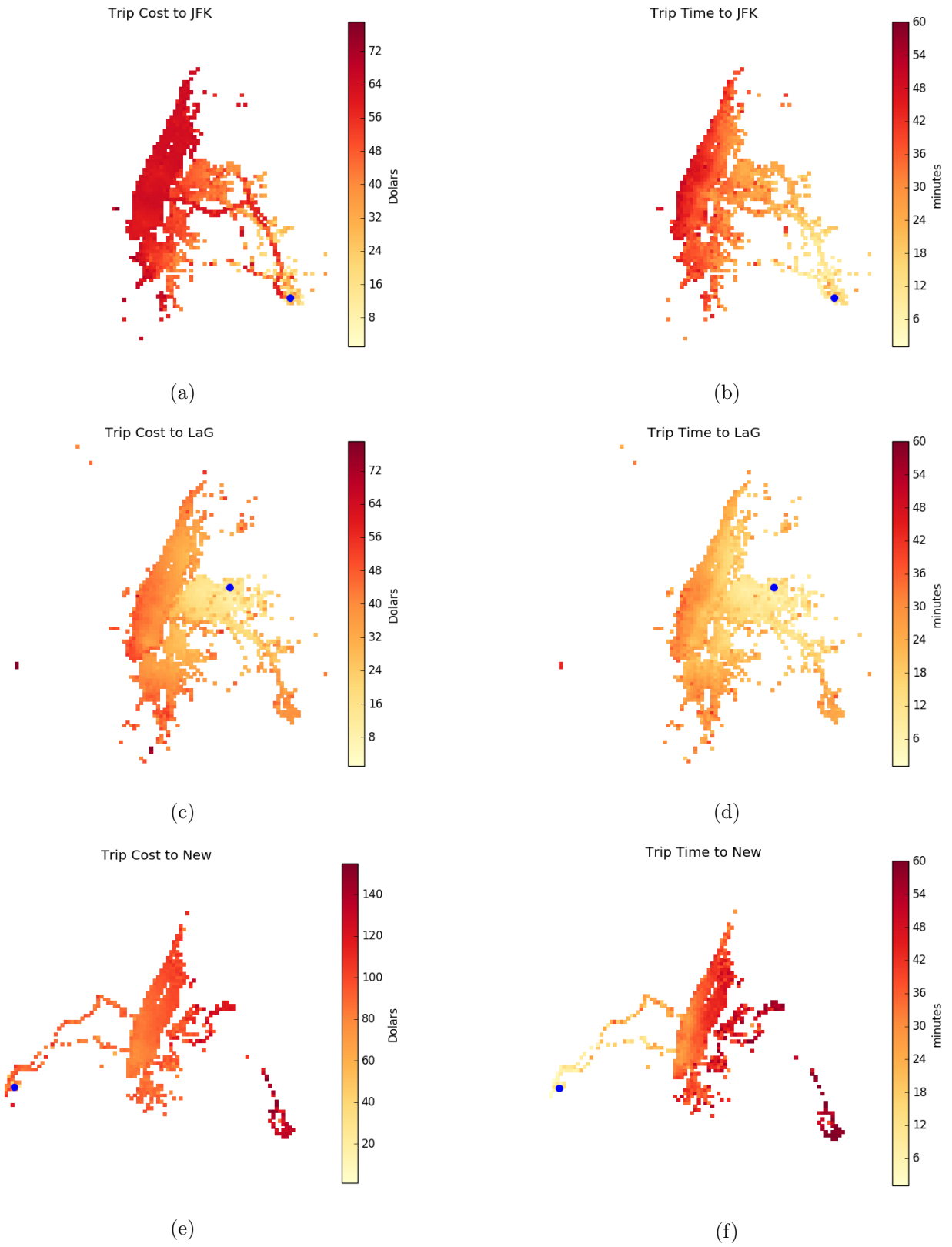

Figure 4: Time and cost maps

3.4 Estimation of the cost of time.

Proceeding as explained in methodology, our model has created a decision map over Manhattan and proximities. The model has been fitted with the real data. The fitted decision map is shown with the real data map in figure 7.

Figure 5: Real data decision on the left side and model predictions on the right one.

The model fits well most points of the map. There is a huge difference on the points near JFK airport. While the real data shows that exactly near JFK it is better to travel to LaGuardia, the model predicts that JFK is a better option. The model prediction is more logical than the data. A possible explanation may be that there are many trips that have a pickup in the main road to JFK and charge the Manhattan flatrate as it is seen in 4a. According to Google researches, the flat rate should not be applied outside Manhattan. It is possible that taxi drivers forget to start the data recording at the beginning of the trip. Therefore they remember while driving on the main road and they start the data recording creating data biases.

The fitting method has given a result of $a \approx 0.042$ and the important result of a Value of time $b \approx 0.35$ USD per minute. That means that we are willing to pay 21 USD more if the transport method reduces the time by at least one hour. To check the validity of the results, it has been scattered the real probabilities on the X axis and the model probabilities on the Y axis. The ideal result should be a straight line of slope 1. The linear regression of the scatter has a slope of $m = 0.991$ with an $R^2 = 0.975$.

4 Conclusions

To sum up, taxi data has described Google data presences successfully, proving its validity as a fine population estimator. Furthermore, taxi data has shown that pickups are more located in the most important streets while dropoffs are more distributed. Another general feature of human behavior is that 20% is by far the most frequent taxi tip.

Using discrete choice theory (2) and (3) with taxi data, a decision map has shown that LaGuardia airport is the best choice for taxi trip from Manhattan for both its lower cost and time. Fitting the model with the data, the results show that NYC citizens have an estimated value of time of $b \approx 0.35$ USD per minute with an $R^2 = 0.975$.

Acknowledgments

This work was supported by the SURF@IFISC fellowship. Financial support from FEDER is gratefully acknowledged. Marc Fuster thanks R. Gallotti and J.J.Ramasco for their dedication and patience.

References

[1] Brian Donovan and Daniel B. Work *New York City Taxi Trip Data(2012-2013)*.1.0 University of Illinois at Urbana-Champaign. Dataset. `http://dx.doi.org/10.13012/J8PN93H8`

[2] `http://pandas.pydata.org/`

[3] `http://www.numpy.org/`

[4] `https://matplotlib.org/`

[5] https://docs.python.org/2/library/datetime.html

[6] https://matplotlib.org/examples/pylab_examples/hexbin_demo.html

[7] McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf

Appendix A: The importance of data cleaning.

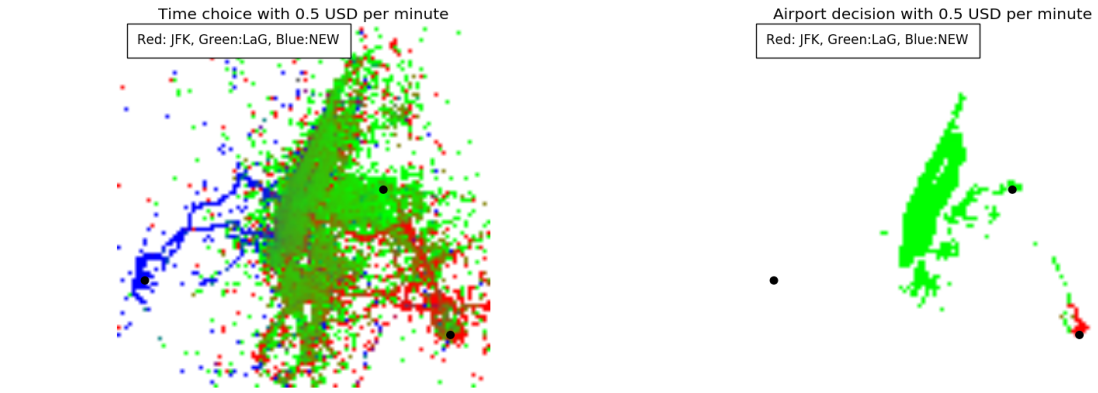

Figure 6: Uncleaned on the left while at least 5 trips to each airport on the right.

Some of the red and blue points on the uncleaned picture are not trustful because they do not contain trips to all airports. Therefore, it is important to delete those points because they may lead to biases.

Appendix B: The importance of logarithmic scale.

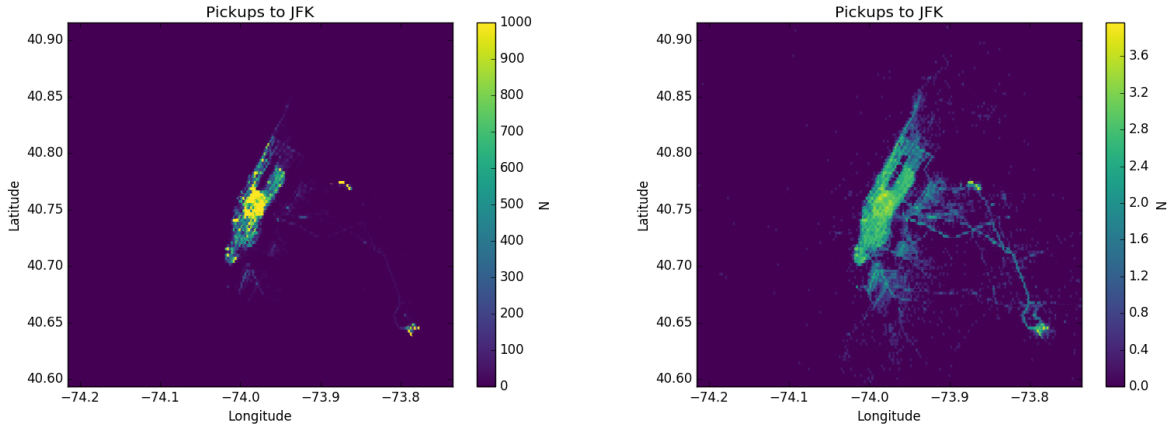

Figure 7: In linear scale on the left versus on logarithmic scale on the right side. Logarithmic scale shows more the change rate.