Gravity Model

Twitter Data from NYC

Overview

1. Probabilistic Weight

Construct a network with the probabilistic weight of link(a,b) defined as

$$link\ weight(a,b) = \Sigma_c \frac{t(c,a) \cdot (t(c,b) - \delta(a,b))}{T \cdot (t(c) - 1)}$$

where

- *t(c,a)* denotes the total number of tweets that user *c* has posted at location (in our case, zipcode) *a*
- $t(c) = \Sigma_a t(c,a)$
- $T = \Sigma_c t(c)$

2. Gravity Model

On the other hand, the weight of links can be modeled as

$$link\ weight(a,b) = k \cdot w^{out}(a) \cdot w^{in}(b) \cdot f(d(a,b)),$$

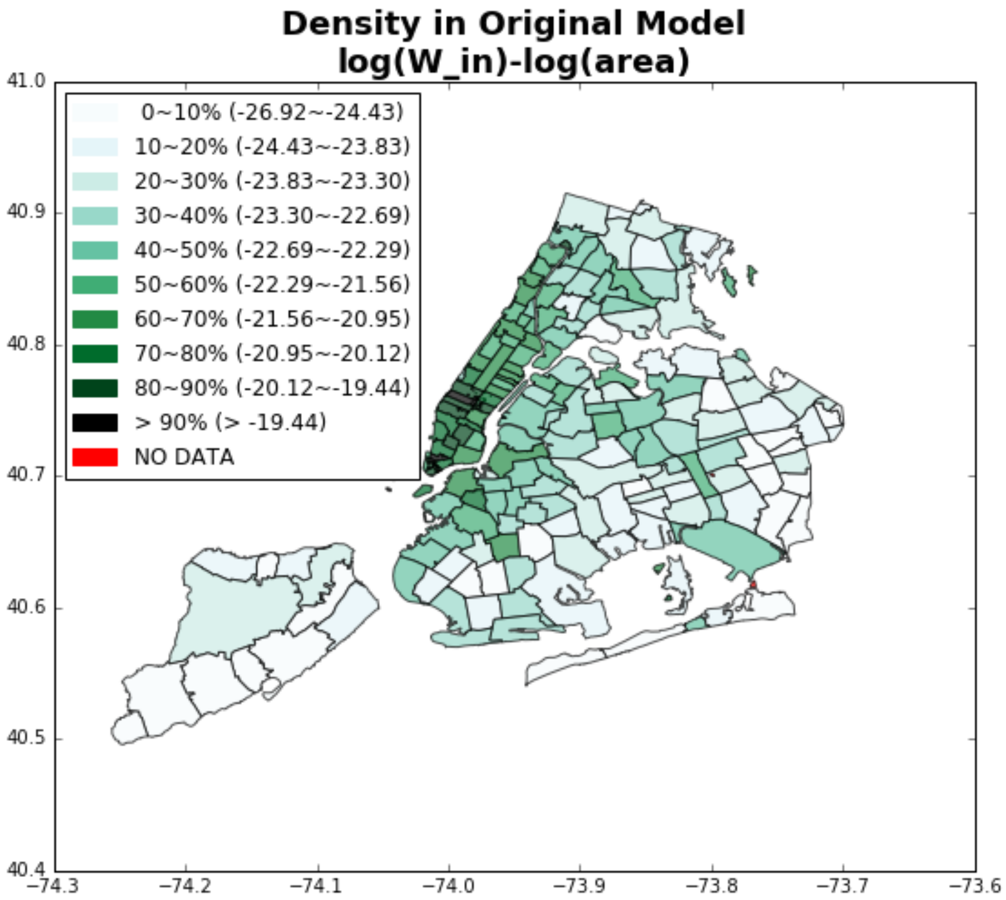
where

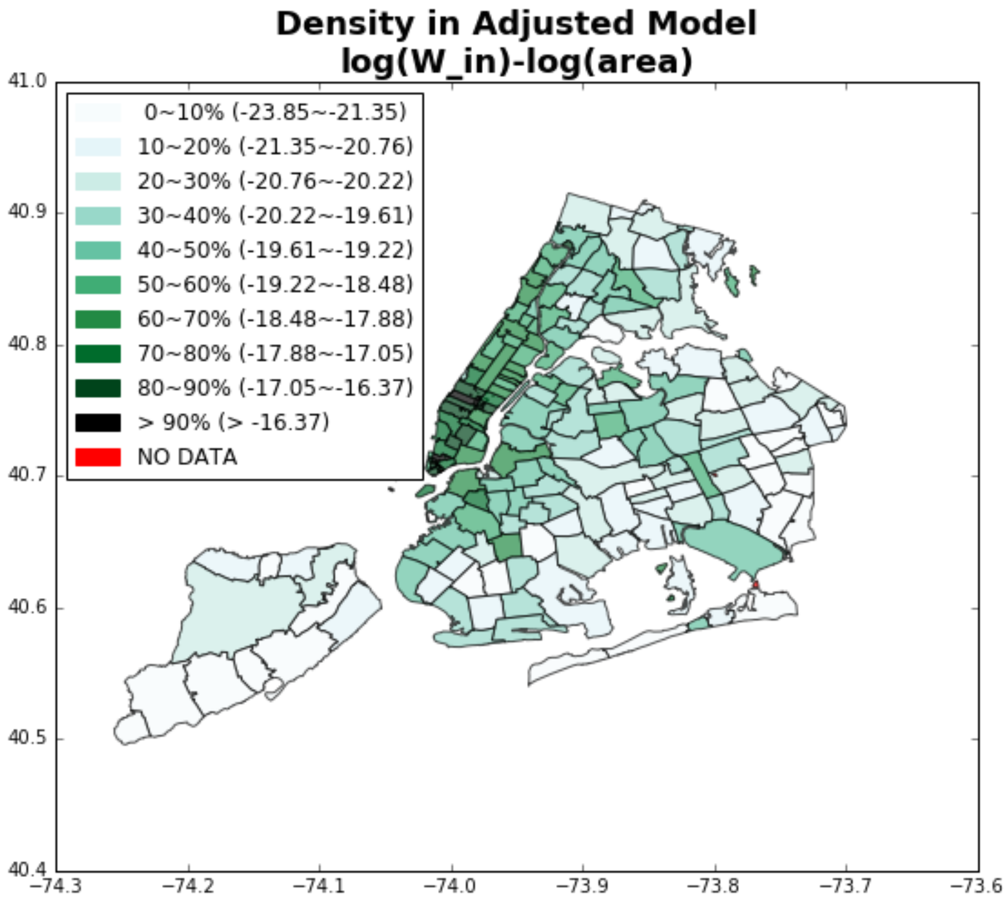
- $w(x)$ is the weight or customized centrality of the node x
- $f(d(a,b))$ denotes function with respect to distance between location a and b, usually it dacays as distance increases.

W_b^{in}:

Divide by the area

Not showing much difference if we add the log of number of tweets

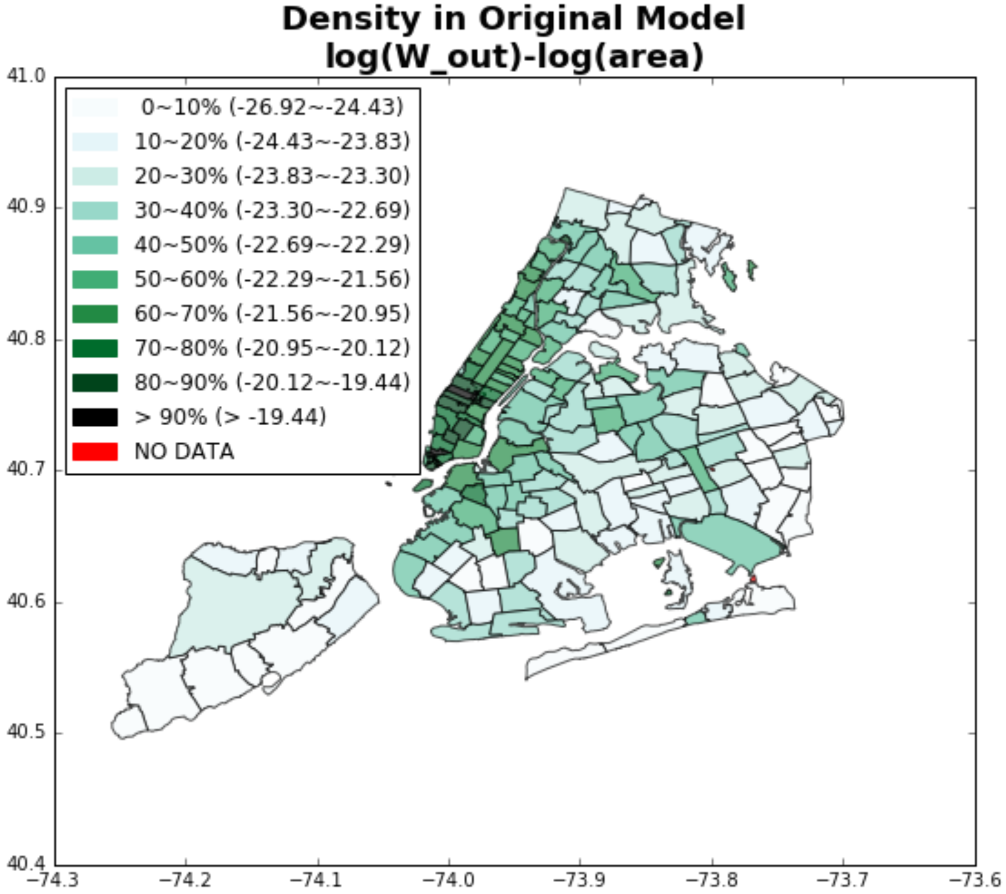


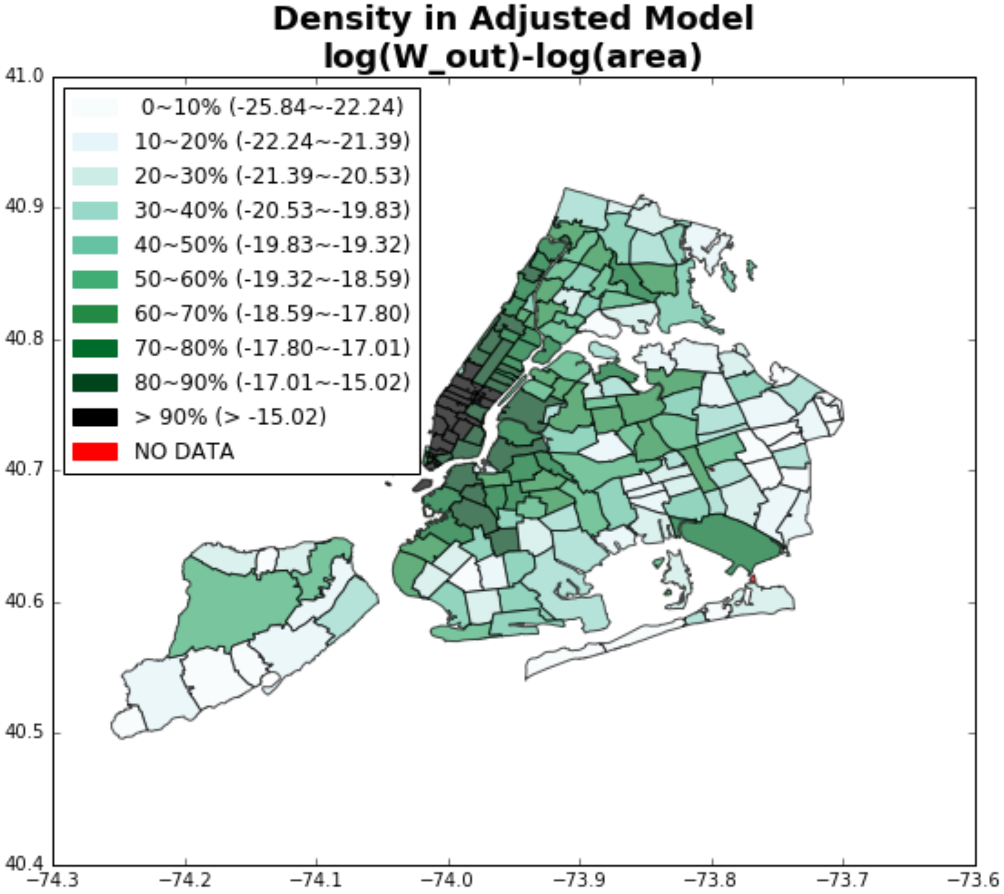


W_a^{out}:

Divide by the area as well

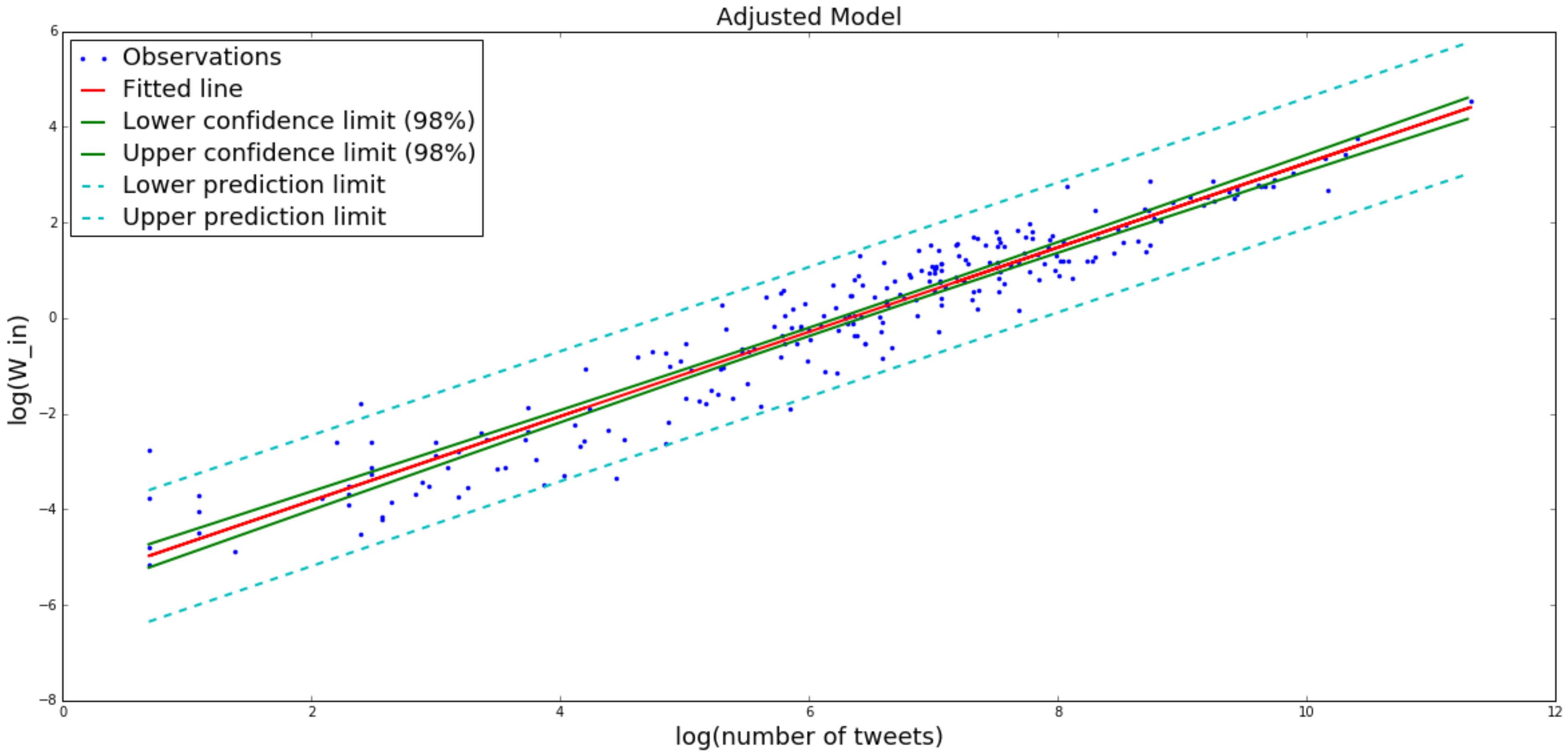
The absolute quantities don't change significantly but the distribution does, resulting in more darker coverage all over the city on different level





The relationship between W_b^{in} and #Tweets (Adjusted Model)

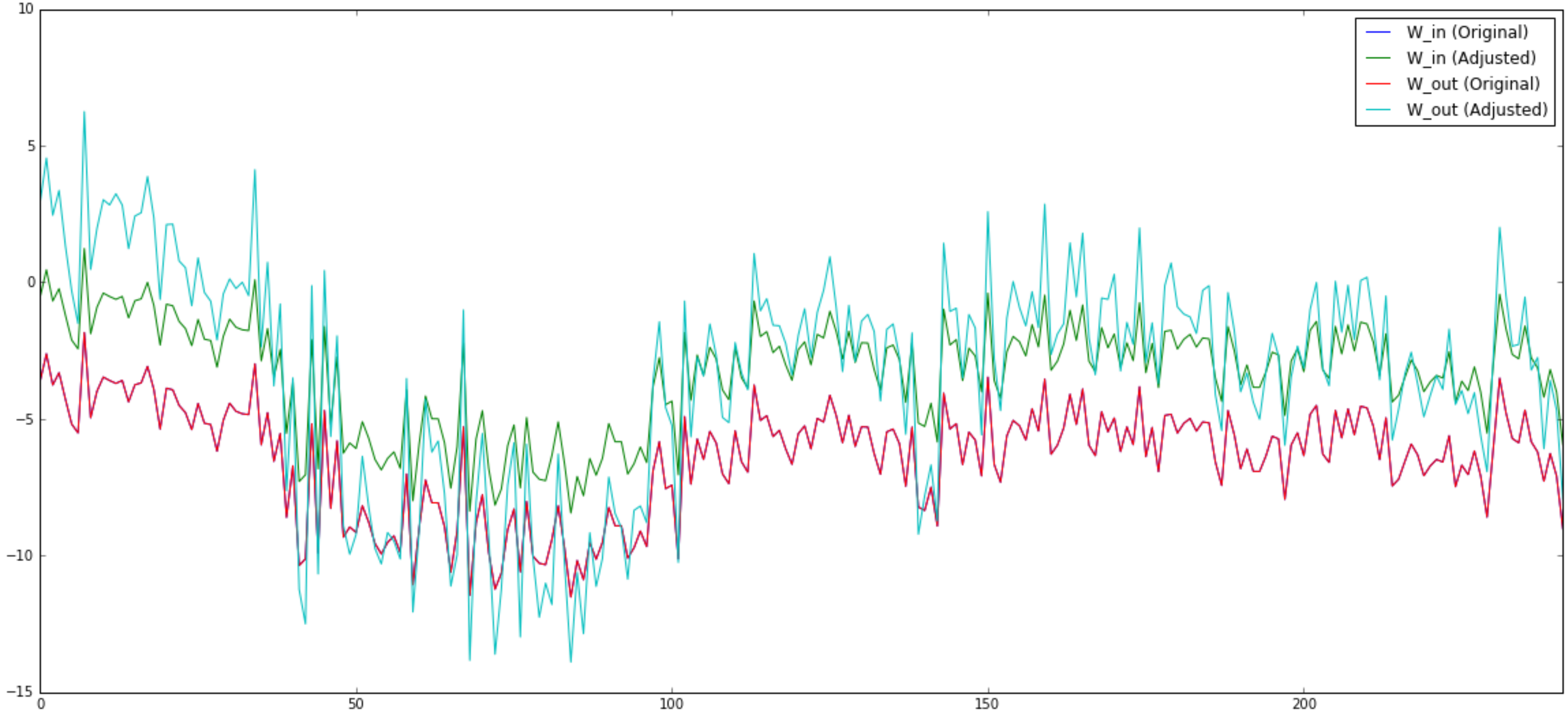
$\log(W_b^{in})$ = 0.894480• log(#Tweets) - 8.982664 or W_b^{in} = 0.0001256• $(\#Tweets)^{0.894480}$
The R square of this model is 0.918590
The model is under linear as the exponent is approximately 0.883218, with 95% CI upper limit 0.928



This plot is for comparing and showing the following facts:

- Red and Blue are practically identical, which makes sense since original model is symmetric
- Green and Cyan are from the adjusted model where we add log(#Tweets). They are clearly not identical anymore.
- Green, as W_{in}, is just a parallel shift of Red/Blue, while Cyan, as W_{out}, changed significantly as expected—they "absorbed" the multiplicative factor from distance function.

(Please ignore the x-label as it should be all the zip codes in NYC instead of 50, 100, etc. but displaying them would make the plot a little messy)



Further exploration

We've focused on the dataset before Dec. 19, 2015 to avoid massive empty data frames between Dec.19, 2015 and Feb. 1, 2016

- 1. Drop the multiplicative factor and examine distance function model

$$link(a,b) \sim W_a^{out} W_b^{in} D^{\gamma}$$

  or equivalently, we try to fit

$$\log\big(link(a,b)\big) \sim \log(W_a^{out}) + \log\big(W_b^{in}\big) + \gamma \cdot \log(D)$$

  where D denotes the distance (measured in miles)

- 2. Check the exponent γ for each day, see if it varies significantly over time

So for the exponent of distance, we've observed:

	Mean	Std
Weekdays	-0.5934	0.1252
Weekends	-0.7411	0.0491

To sum up, **weekends have lower γ and less fluctuation**