Bay Area's Happiest Bicyclists: A Study on the 2011 Implementation of GreenWaves in San Francisco

By: Zachary Wong

Background, Goals, and Thought Process:

In Spring 2011, the San Francisco Municipal Transportation Agency implemented GreenWaves in the SF Mission district to encourage residents to bicycle instead of drive. The GreenWaves are synchronized traffic lights that let bicyclists encounter a series of green lights, so they experience fewer stops in the particular block. While there is community support for this initiative and the SF government has used these GreenWaves as evidence for their city's success in becoming more environmentally conscious, I evaluate whether these GreenWaves indeed increased the number of bicyclists in SF or whether their claims are only based on intuition.[1] If these claims are true, SF and other Bay Area policymakers should invest in implementing more GreenWaves - the increase in bicyclists may decrease traffic congestion in the SF area, increase workers' productivity, and help make Silicon Valley a greener, healthier space for all its residents.

My Research Question:

Did the 2011 implementation of GreenWaves in SF's Mission District increase the number of bicyclists in SF?

My Approach:

My research design is based on a regression discontinuity in time design, which spans from 2006 to 2017, chosen based on available data. I focused particular attention on 2011, when the first GreenWaves were implemented. My research specification:

$$COUNT_{SF} = \alpha + \beta 1(POST)_{SF} + \beta 2(RACK)_{SF} + \varepsilon_{SF}$$

COUNT is my outcome variable: the manual count of bicycles in SF for each year. POST is a dummy variable recorded as 1 if the year is 2011 or after and 0 if pre-2011. RACK is a covariate for the number of bicycle rack spaces in SF, accounting for this case of omitted variable bias: SF implemented more bicycle racks after 2011 and led more people to bicycle.

I used three data sets, all from the SF open data portal. The first data set is called "Bike Volume Manual Counts," which provided me with data for my outcome variable of the number of bicyclists counted every year in September (counted after GreenWaves implementation in March 2011). The second data set is called "Bicycle_Green_Wave_Streets," which helped me determine which streets the GreenWaves were implemented and confirm that they were in the Mission District. Lastly, "Bicycle_Parking" gave me the number of rack spaces for bicycle parking with years available from 2006 to 2017. Figure 1 in the appendix shows the cleaned

[1] Background researched primarily from this news article: https://www.sfmta.com/press-releases/sfmta-unveils-new-and-enhanced-bicycle-friendly-green-waves

data in preparation for analysis in STATA. Note: There is no bicycle count data for 2012, so the data skips 2012 and picks up from 2013 to 2017.

Findings:

I found a dramatic increase in the number of cyclists after 2011. Looking at Figure 2, even though bicycle counts were rising before 2011, we still see a stark discontinuity at the 2011 cutoff where there were 3844 more bicyclists or a 58% increase in bicyclists after the GreenWaves were implemented. Additionally, post-2011 is relatively smooth, with little bumps in bicycle counts. In Figure 3, I plotted the RACK covariate against 2006-2017, and the number of bicycle spaces had little discontinuity in 2011. There was even a 47% *decrease* in the bicycle rack spaces from 2010 to 2011. This graph provides more substantial evidence that the jump was caused by the GreenWaves and not an omitted variable like increased spaces for secure bicycle parking that may have increased bicyclists around 2011. Furthermore, in Figure 4, I overlay both the outcome variable and covariate against years, and we can see the smoothness of the bicycle rack spaces versus the discontinuity in the bicycle counts at the 2011 cutoff.

Statistical Significance:

Figure 5 shows the ANOVA table for the complete regression. Even though the coefficient of interest is statistically significant, even above the 1% confidence level with a t-stat of 5.43, there was only 1 sample for each year. Similarly, even though it would typically be reassuring that the coefficient on RACK is not statistically significant with the 95% confidence interval intersecting 0, the sample size is just too small enough to rely on these standard errors. The main reason for the single sample size is that the data for SF is only collected once a year and not multiple times throughout the year. As a workaround to this limitation, I wanted to divide the data by districts to increase the study's statistical power. I would have separated the Mission district, which had GreenWaves, from the other districts which didn't implement GreenWaves, but there were too many missing values for the other districts to make that approach feasible. Additionally, I did not have the means to account for the spillover effects of Mission District's GreenWaves with this district-level approach: residents from other districts would likely bike more frequently because they can also use the GreenWaves in Mission District.

Further Limitations/Assumptions

I attempted to identify and control an important omitted variable: secure bicycle parking availability. I also needed to aggregate all SF data, which naturally accounted for other state-level policy changes that may have increased bicyclists since these policies applied to every SF resident. However, I would have liked to introduce a comparison group, specifically Berkeley, to support my results further. Berkeley has no GreenWaves, and it is located across the bay so the distance and water body limit spillover effects of bicyclists from SF to Berkeley. However, Berkeley is near enough to SF where other variables like weather conditions or general economic fluctuations would somewhat affect both areas. Thus, showing that Berkeley and SF are not systemically different except for the increase in bicyclists in SF versus Berkeley after 2011 would further uncover the causal effect of the GreenWaves on increasing bicyclists.

Besides omitted variable bias: I want to address assumptions I made to account for 2 other problems common with regression discontinuity in time designs:

1. **Time-varying treatment effects:** GreenWaves could have been de-installed or never fixed after a malfunction after 2011. This study assumes that the 2011 GreenWaves are working and installed after 2011 until at least 2017.
2. **Sorting effects**: People who love biking could have chosen to move or live in the Mission district because of the GreenWaves implementation, leading to the increase in bicyclists. This study assumes that the GreenWaves are not enough motivation for bicycle-loving people to relocate only a few blocks or districts.

Conclusion/Reflection:

I thought that finding a unique problem to evaluate worked out well. Homelessness and housing are some of SF's biggest challenges and hundreds of researchers have already studied these topics but picking a specific year and government intervention to analyze helped narrow down my scope. However, I found that the process of finding sufficient data to answer my research question was difficult. I had to change my approach a few times to account for empty data in the Bicycle Count dataset. Even then, I could not come up with enough samples to rely on the standard errors for statistical significance analysis. However, the graphs and data still point towards GreenWaves making a difference by encouraging motorists to bicycle more often. More mayors should invest in this technology, as its impacts span sectors from safety, health, to the environment. On the technical side, I thought that the combination of SQL, Pandas, and Stata worked quite well together. I had fun practicing with these tools throughout this process, integrating them in Jupyter notebooks, and trying to leverage the best of each of these tools.

Appendix:

Bicycle and Rack Count per Year

year	bicycle_count	spaces	post
2006	4282	284	0
2007	4623	41	0
2008	6170	88	0
2009	6469	46	0
2010	6639	1292	0
2011	10483	688	0
2013	11010	1127	1
2014	11010	405	1
2015	10655	2101	1
2016	11547	706	1
2017	10772	1032	1

Figure 1: Bicycle and Rack Count in SF with generated "post" dummy variable

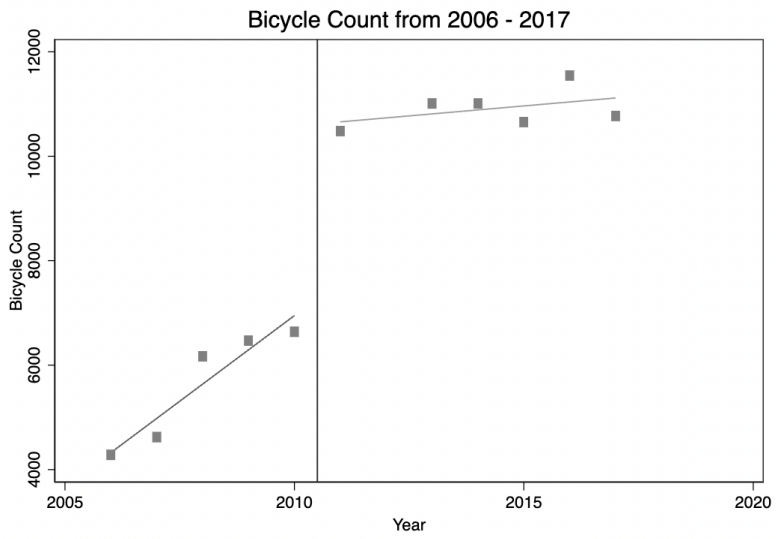

Figure 2: Bicycle Count in SF with a Discontinuity at the Cutoff 2010-2011

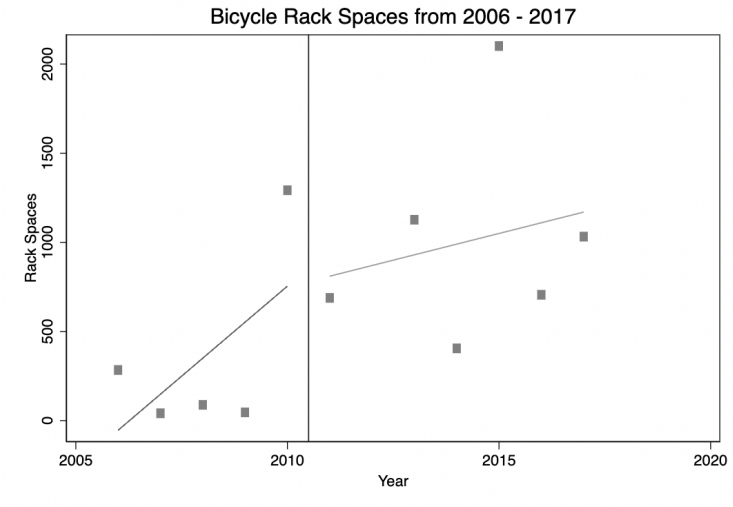

Figure 3: Bicycle Parking Availability in SF without much Discontinuity at the Cutoff 2010-2011

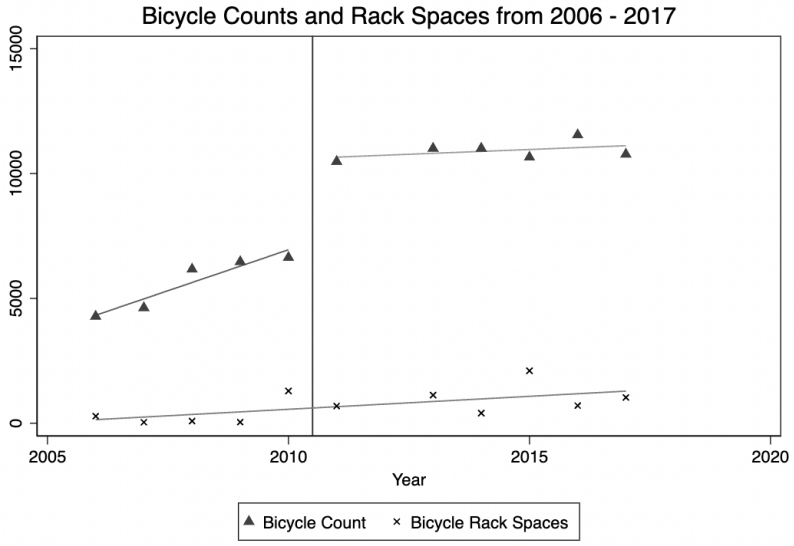

Figure 4: Comparison of Bicycle Count and Parking Availability Discontinuity in SF at the Cutoff 2010-2011

```
      Source |       SS           df       MS      Number of obs   =        11
-------------+----------------------------------   F(2, 8)         =     22.92
       Model |  69357570.9         2  34678785.4   Prob > F        =    0.0005
    Residual |  12102643.8         8  1512830.48   R-squared       =    0.8514
-------------+----------------------------------   Adj R-squared   =    0.8143
       Total |  81460214.7        10  8146021.47   Root MSE        =      1230

------------------------------------------------------------------------------
bicycle_co~t | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        year |   707.6565   130.2095     5.43   0.001     407.3929     1007.92
      spaces |  -.1030588   .7681104    -0.13   0.897    -1.874324    1.668207
       _cons |   -1414831   261577.2    -5.41   0.001     -2018029   -811633.2
------------------------------------------------------------------------------
```

Figure 5: ANOVA table showing coefficient on post and spaces. Refer to the "Statistical Significance" section for further discussion of this study's sample size and standard errors.