Understanding Impact of London Olympics on Crime

Dhiraj Hinduja, Tarun Newton, Yanlun Ren, Chandrakanth Tolupunoori, Venkata Krishna Chaitanya Rachapudi

March 04, 2019

Executive Summary

Tourism is a billion dollar industry for many countries especially for the UK. Tourism in the UK is expected to be nearly 10% [1] of its GDP. Hence, it is important for the UK to establish a healthy and safe environment for the tourists especially in London which is one of the largest tourist attractions in the world.

An influx of tourists is generally driven by large scale events. Crime rates could go high in crowded areas and a negative experience for tourists could change their viewpoint on London. Hence, our aim is to study the true effect of London Olympics on the crime rate during that period. Knowing the effect would help the London government decide on the level of police force and awareness to forestall crimes and establish a safe environment for tourists. To achieve this, we analyzed the crime rate trends across the UK by police force districts where Greater London is taken as an area of treatment and the rest of the areas are considered as areas of control (excluding areas surrounding Greater London to avoid interference problems)

Y(greater London) -> X (rest of the areas in the UK excluding areas surrounding London)

We used two methods to determine the true effect of change in crime rates on London namely difference in differences and synthetic control. We cannot rely on the results of difference in differences as the data violates some of the assumptions of the method. The other method that we used is called synthetic control which showed a reduction in crime rate after London Olympics with large standard errors. The results showed that the crime rate changes anywhere from 5% to -65% but under the assumptions and limitations of our analysis we are unable to find statistically significant causal result on the effect of London Olympics on crime. One key component of this involved security at the Games, with the inability of G4S, the private security firm awarded the contract for Games security, to fulfill its obligations, resulting in a major shift in security provision in the weeks leading up to the Games. In this circumstance it can be considered as a commendable feat for the London police that the analysis did not know an increase in crime rate during the olympics games.

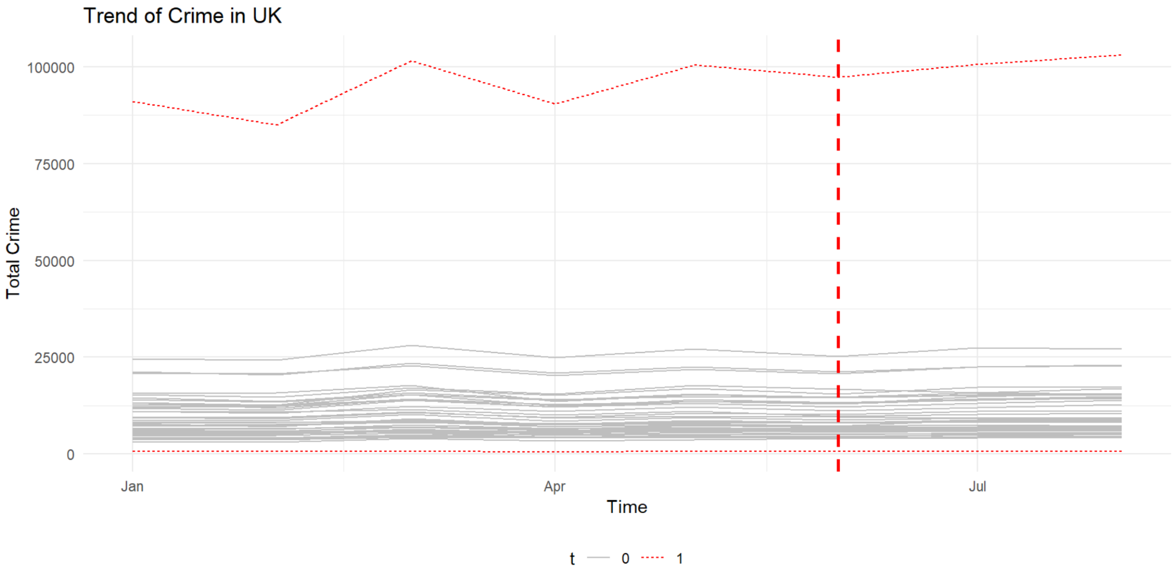
The document further explains the challenges with the data set used and the next steps that can be taken to improve the analysis

Background

The build-up to the 2012 Olympics was fraught with problems and attracted considerable negative publicity in both the British and international press. One key component of this involved security at the Games, with the inability of G4S, the private security firm awarded the contract for Games security, to fulfill its obligations, resulting in a major shift in security provision in the weeks leading up to the Games. Considering the already established importance of tourism to London and the entire UK as a whole understand if the London Olympics caused an increase in crime is an important question to answer.

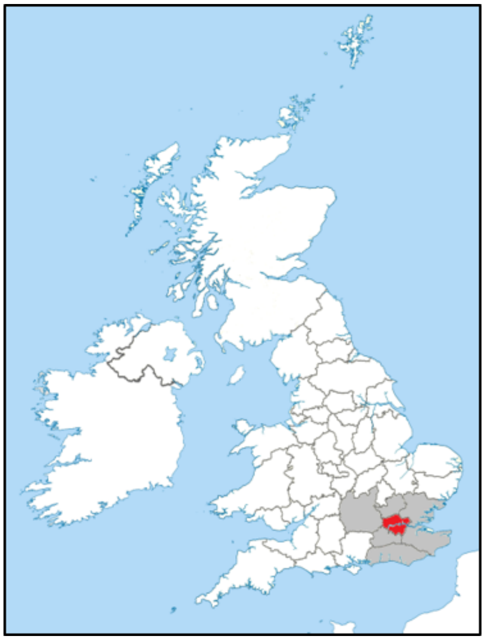
Data Description

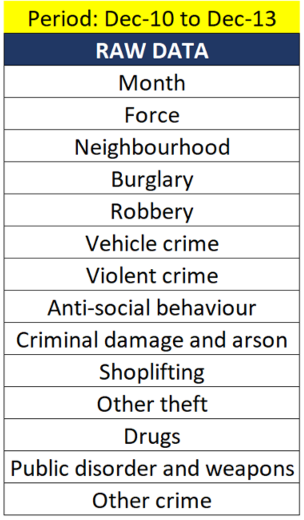
The data used for the analysis comes from UK's police website. This is the site for open data about crime and policing in England, Wales and Northern Ireland. We took data recorded about monthly criminal activity at a police force level from 2011 through 2013 for Great Britain. It is separated into a variety of types of offenses. These are violence against the person, sexual crimes, robbery, burglary, theft, fraud, damage, drugs and other. Each of these categories provided can be parsed out into further subcategories. However, we combined the various categories into a total crime value as we were interested to understand the impact of the Olympics on total crime (we also did separate analysis trying for different crime groups). The games began in late July 2012 and concluded in mid-August. Although the games were held over a short period, it is fair to assume that the time period over which effects that the games might have had could be observable over a longer period. Hence we are trying to see if crime increased during the months of July to October 2012. That is we assume that the treatment start date for our analysis is July 2012. The geographical unit which we considered is a police Force. The data has 44 Force districts in the United Kingdom.



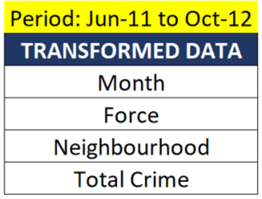
Based on our research Metropolitan Police Service and the City of London Police are the police forces where the Olympic games of 2012 were held in London. Therefore in our analysis these two forces were considered as the treatment unit.

To avoid spillover impact the areas surrounding the test force districts were removed because we require that the potential outcome observation of one unit should be unaffected by the particular assignment of treatments to other units. Interference amounts to violating this assumption. Hence we removed the Force areas geographically closest to the location of the Olympics from the analysis which in our case is Surrey Police, Sussex Police and Kent Police. Therefore we consider the remaining Forces as out control units.





Period: Dec-10 to Dec-13
RAW DATA
Month
Force
Neighbourhood
Burglary
Robbery
Vehicle crime
Violent crime
Anti-social behaviour
Criminal damage and arson
Shoplifting
Other theft
Drugs
Public disorder and weapons
Other crime



Period: Jun-11 to Oct-12
TRANSFORMED DATA
Month
Force
Neighbourhood
Total Crime

Ideal Experiment

The gold standard to determine the causal effect of the Olympics on the crime rate would be to randomly organize the Olympics events at different boroughs in London and different times of the year. Location and time are potential confounders as the Olympics might only be organized in crowded or affluent areas thus biasing the crime rate as different boroughs have different socio-economic parameters. Due to random assigning of the event to various boroughs would take care of these possible confounders.

However, implementing an experiment like this is not feasible as there are huge monetary costs involved in organizing the Olympics and the sales could also be affected by the location of the event.

Methods

Difference in differences:

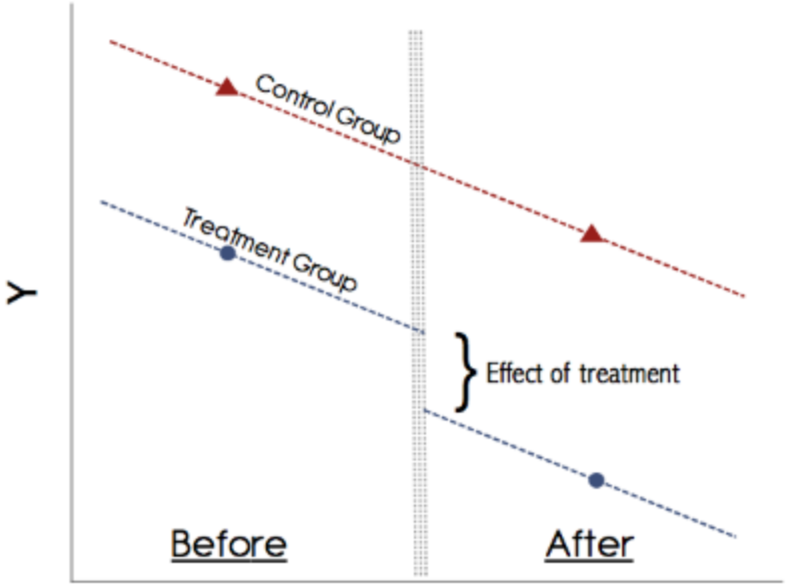
To understand if Olympics in London has an impact on crime we cannot do a simple pre post difference because there could be time-invariant confounds within groups ie Group Fixed Effect or inter-temporal confounds that apply to everyone ie Time Fixed Effect or both. Therefore we have to use a technique called Difference in differences. In short, in difference in differences we,

- First we take the after-before difference in each group.

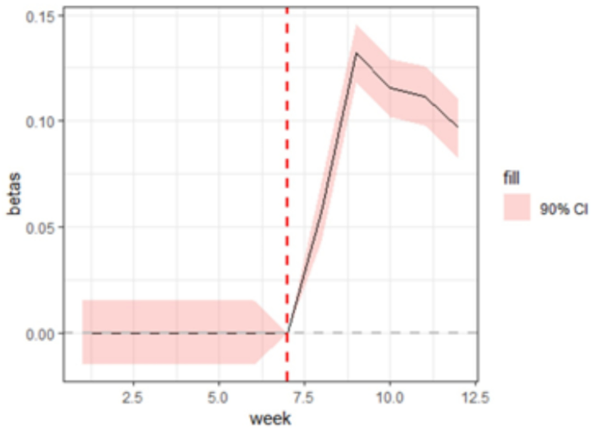
- Next, calculate the difference in those inter-temporal differences.

Difference in differences (DiD) will help determine the causal effect of a treatment (London Olympics) for a location (Greater London) by comparing trends with similar other areas (rest of the areas in the UK excluding the abutting areas of Greater London)

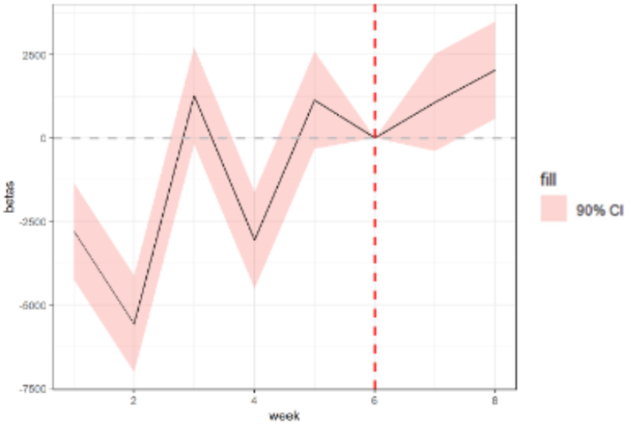
The assumptions for Difference in difference are: * We assume parallel trends between the two groups on the outcome variable. * We also assume that treated subjects are not influencing control subjects.



The method requires both the treatment and control to have parallel trends so that the change in the slopes of the lines (in blue) after the treatment happens can be considered to be the effect of the treatment. However, in our case, the crime rate trends did not match well. The normalized slopes for both treatment and control group should equate to 0 as shown in the left graph (expectation). However, the right graph is the reality for our data because of which we could not use DiD



Expectation



Reality

Parallel trend assumptions is very important for difference in differences because the control group needs to be a good counterfactual for the treated group. As our data during the pre period has violated the parallel trends assumptions we are now trying another method called synthetic control.

Synthetic control method:

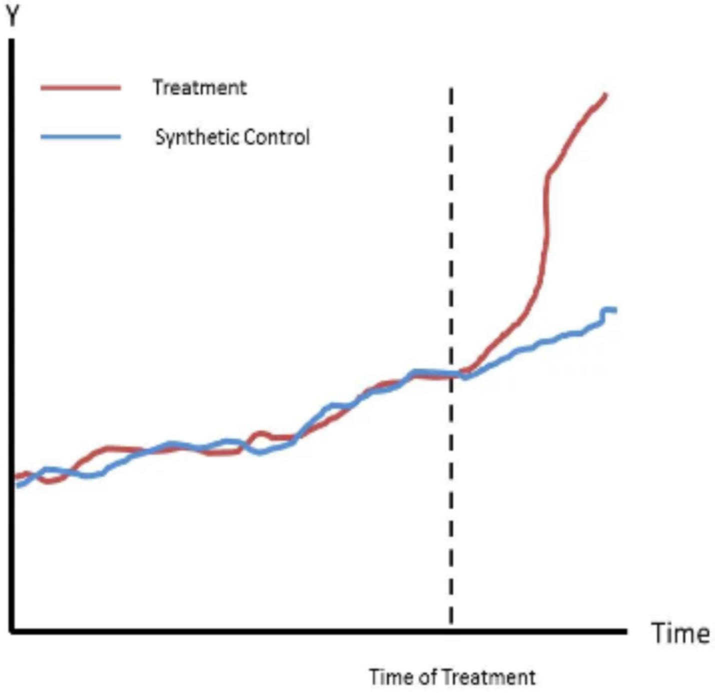
Synthetic control method is generally used when pre-treatment Trends are Violated, ie our set of controls is not useful for DiD, so we cannot just trust the DiD estimate of the treatment effect.

The synthetic control method is a statistical method used to evaluate the effect of an intervention in comparative case studies. It involves the construction of a weighted combination of groups used as controls, to which the treatment group is compared. This comparison is used to estimate what would have happened to the treatment group if it had not received the treatment. Unlike difference in differences approaches, this method can account for the effects of confounders changing over time, by weighting the control group to better match the treatment group before the intervention. Another advantage of the synthetic control method is that it allows us to systematically select comparison groups.

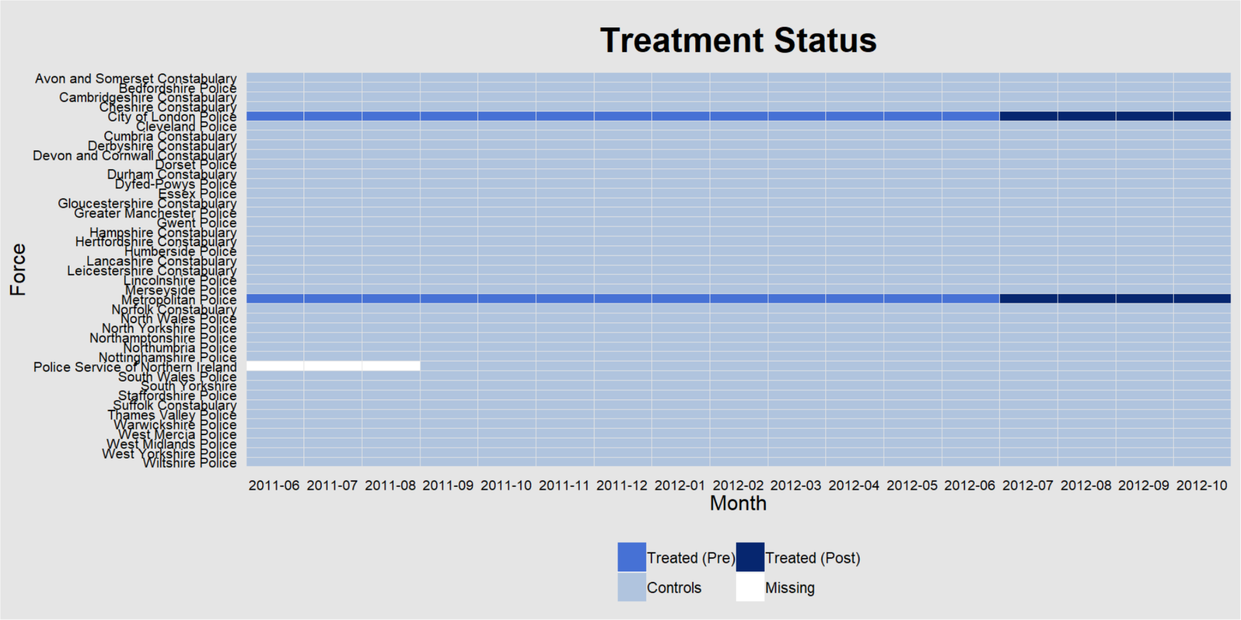
To be able to perform Synthetic control we need the below conditions to be met * Need Repeated Observations of Treated and Control, in the Presence and Absence of Treatment. * No Interference,we must not have any interference between treated and control units, else the counterfactual will be wrong.

In our context we use the rest of the areas of the UK (except for the ones abutting Greater London) shown as a blue polynomial curve below (synthetic control) to predict the crime rate trend in Greater London shown as a red polynomial curve in the graph below.

The prediction, if estimated accurately, can be then considered to the crime rate that would have been had there been no olympics shown as continuing blue polynomial after the treatment (dotted vertical line in the graph). The difference between the predicted crime rates and the true crime rates is assumed to be the effect of the London Olympics on crime rates.

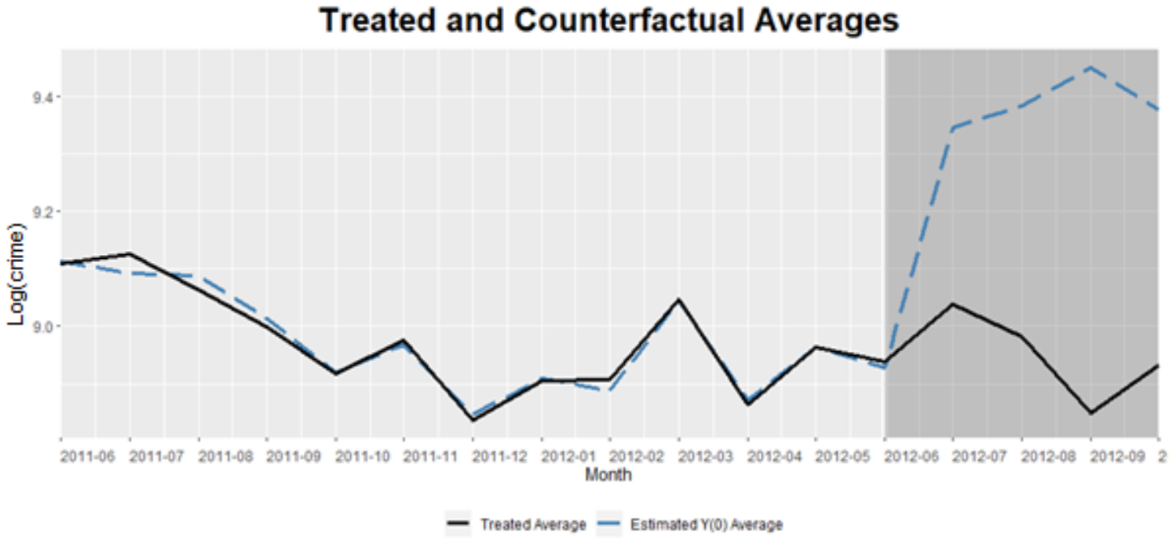


Before we conduct any statistical analysis, it is helpful to visualize the data structure and/or spot missing values (if there are any). We can easily do so with the help of panelView. The following figure shows that: (1) there are 2 treated units and 42 control units; (2) the treated units start to be treated in period 2012-07; and (3) there are missing values for one Police force.



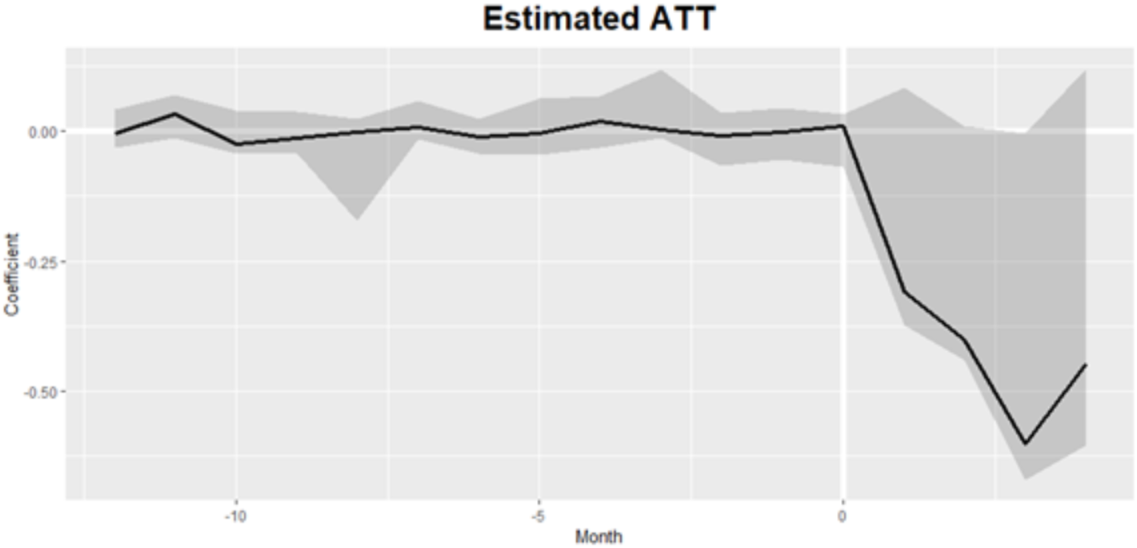
Using pre-period data from other states, we are building a model that assigns fixed weights to each control police force, and arrives at a weighted average that closely resembles crime rate in London before the Olympics.

We will then be able to use the resulting model to synthesize what crime in London would have looked like in post period, too (absent treatment). The result showed that there is actually a reduction in the crime rates in London after the Olympics.



The black line denotes the true crime rates in Greater London while the dashed line represents the predicted crime rates in London.

Although, the graph shows a decrease in the crime rates in Greater London, the results are not statistically significant as shown below:



From the graph above, the crime rate has very large confidence intervals meaning the results tell us that the true crime rate after the London Olympics can vary anywhere between +5% to -65% compared to the crime rates that would have been had there been no Olympics.

Since, the confidence intervals are so large, it is difficult to conclude whether the London Olympics caused a reduction or an increase in crime rates in Greater London.

Limitations

The limitations of our analysis is that we were not able to account for anticipatory role that the Government took ahead of time to prepare for the London olympics. We were not able to answer how does the strategic introduction of police influence the crime rate changes. We aren't sure about when the actual start date really occurred and if that could have caused a problem with our analysis. It is measurement error sort of problem because if police force was increased ahead of time by a few months only in the Olympic areas then out pre treatment effects are incorrect and it can cause a problem with our analysis. Based on our research we found out that over 23,700 security staff would be deployed for the games but the exact dates and overall trends in police force deployment is not available. Another point to note would be that the London police force has a budget of £2.5 billion while most other police forces budgets are in the range of £500 million.

Conclusions

The build-up to the 2012 Olympics was fraught with problems and attracted considerable negative publicity in both the British and international press. One key component of this involved security at the Games, with the inability of G4S, the private security firm awarded the contract for Games security, to fulfill its obligations, resulting in a major shift in security provision in the weeks leading up to the Games. Under the assumptions and limitations of our analysis we see that crime decreases but we are unable to find statistically significant causal result on the effect of London Olympics on crime. Based on our research we see that Olympics security bill soared to more than £1bn though we cannot with absolute certainty conclude that there was a decrease in crime we can say that London Olympics did not cause an increase in crime. Further analysis listed in the sections below might help bolster the claim that crime went down during the London Olympics. The evidence suggests that Olympic security was a success that there was no increase in crime that usual for the government and the London Organising Committee of the Olympic and Paralympic Games (LOCOG), despite the negative publicity attached to G4S and the private security sector.

Next Steps

To bolster the analysis we recommend doing either one or both of the following analyses

Strategy 1:
The first approach we could take would be to enhance the existing approach by collecting more information about London and other control cities in the UK. As mentioned in the above sections if we are able to gather more data in terms of the monthly police force size or other economic indicators our synthetic control might be able to give us more confident results and we might be able to bolster our existing analysis enough to give get a conclusive answer. Further analysis on the deployment of extra police in different geographies will also help isolate the impact.

Strategy 2:
Analyze London's crime rate in comparison to similarly populous cities across the world for better control. In the current method one of the drawbacks is that the crime rate of the London which is the test unit is very different from the crime rate in the other parts of the UK which act as the control units. This limitation can be overcome if were to use other major cities, possibly cities that have hosted the Olympics before to ensure that they would be act as suitable controls.

References

1. https://www.visitbritain.org/visitor-economy-facts
2. Data source: data.police.uk
3. UK Police force: https://en.wikipedia.org/wiki/List_of_police_forces_of_the_United_Kingdom
4. https://www.telegraph.co.uk/sport/olympics/news/9472567/Olympics-saw-crime-fall-in-London.html
5. https://www.theguardian.com/sport/2012/mar/09/olympics-security-bill-how-it-soared
6. https://www.bbc.com/news/uk-16195861
7. https://www.visitbritain.org/annual-survey-visits-visitor-attractions-archive

Appendix

Loading the required packages

```
# Load packages
library(splitstackshape)
library(stargazer)
library(plm)
library(lfe)
library(dplyr)
library(ggplot2)
library(gsynth)
library(panelView)
```

Loading the data

The data is sourced from the UK police website: data.police.uk. This is the site for open data about crime and policing in England, Wales and Northern Ireland.

```
# Load data
df = read.csv('collated_data_201201_201306.csv')
```

Basic summrary statistics.

From the below statistics we can see that we have data about the number of different types of crime at a Force, Neighbourhood level for every month in 2012.

```
summary(df)
```

```
##        X             Month                                      Force      
##  Min.   :    1   2012-01: 5289   Metropolitan Police           :11367  
##  1st Qu.:23710   2012-02: 5289   Kent Police                   : 5544  
##  Median :47419   2012-03: 5289   Avon and Somerset Constabulary: 5470  
##  Mean   :47419   2012-04: 5273   North Wales Police            : 5094  
##  3rd Qu.:71128   2013-06: 5268   Sussex Police                 : 4906  
##  Max.   :94837   2012-10: 5266   Lancashire Constabulary       : 4247  
##                  (Other):63163   (Other)                       :58209  
##  Neighbourhood      Burglary          Robbery      Vehicle.crime    
##  A1     :   72   Min.   :  0.000   Min.   : 0.00   Min.   :  0.000  
##  A2     :   72   1st Qu.:  2.000   1st Qu.: 0.00   1st Qu.:  1.000  
##  A3     :   72   Median :  4.000   Median : 0.00   Median :  3.000  
##  C11    :   72   Mean   :  7.396   Mean   : 1.02   Mean   :  6.161  
##  C12    :   72   3rd Qu.:  9.000   3rd Qu.: 1.00   3rd Qu.:  8.000  
##  (Other):70524   Max.   :159.000   Max.   :64.00   Max.   :142.000  
##  NA's   :23953                                                      
##  Violent.crime    Anti.social.behaviour Criminal.damage.and.arson
##  Min.   :  0.00   Min.   :  0.00        Min.   :  0.000          
##  1st Qu.:  2.00   1st Qu.: 10.00        1st Qu.:  2.000          
##  Median :  6.00   Median : 22.00        Median :  5.000          
##  Mean   : 10.42   Mean   : 35.82        Mean   :  8.678          
##  3rd Qu.: 13.00   3rd Qu.: 42.00        3rd Qu.: 11.000          
##  Max.   :201.00   Max.   :748.00        Max.   :200.000          
##                                                                  
##   Shoplifting       Other.theft         Drugs        
##  Min.   :  0.000   Min.   :  0.00   Min.   :  0.000  
##  1st Qu.:  0.000   1st Qu.:  2.00   1st Qu.:  0.000  
##  Median :  1.000   Median :  5.00   Median :  1.000  
##  Mean   :  4.854   Mean   : 10.92   Mean   :  3.193
##  3rd Qu.:  5.000   3rd Qu.: 12.00   3rd Qu.:  4.000
##  Max.   :222.000   Max.   :514.00   Max.   :321.000
##
##  Public.disorder.and.weapons  Other.crime        treatment
##  Min.   : 0.000              Min.   :  0.000   Min.   :0.0000
##  1st Qu.: 0.000              1st Qu.:  0.000   1st Qu.:0.0000
##  Median : 1.000              Median :  1.000   Median :0.0000
##  Mean   : 2.177              Mean   :  2.477   Mean   :0.1206
##  3rd Qu.: 3.000              3rd Qu.:  3.000   3rd Qu.:0.0000
##  Max.   :69.000              Max.   :278.000   Max.   :1.0000
##
```

There are 44 police forces present in the data we have.

Based on our research Metropolitan Police Service and City of London Police are the police forces where the Olympic games of 2012 were held in London.

https://en.wikipedia.org/wiki/List_of_police_forces_of_the_United_Kingdom

```r
length(unique(df$Force))
```

```
## [1] 44
```

Data Transformations

The data used for the analysis comes from UK's police website. It is separated into a variety of types of offenses. These are violence against the person, sexual crimes, robbery, burglary, theft, fraud, damage, drugs and other. Each of these categories provided can be parsed out into further subcategories. However, we combined the various categories into a total crime value as we were interested to understand the impact of the Olympics on total crime.

```r
# Aggregate total crime per month per force
df$vaue = df$Burglary + df$Robbery +df$Vehicle.crime + df$Violent.crime +
          df$Anti.social.behaviour + df$Criminal.damage.and.arson +
          df$Shoplifting+ df$Other.theft + df$Drugs +
          df$Public.disorder.and.weapons + df$Other.crime

df_agg = df %>% group_by(Force, Month, treatment) %>% summarise(value = sum(vaue))

df_agg$date = as.Date(paste( df_agg$Month,"-01",sep=""))

df_agg$t = as.character( df_agg$treatment)
```

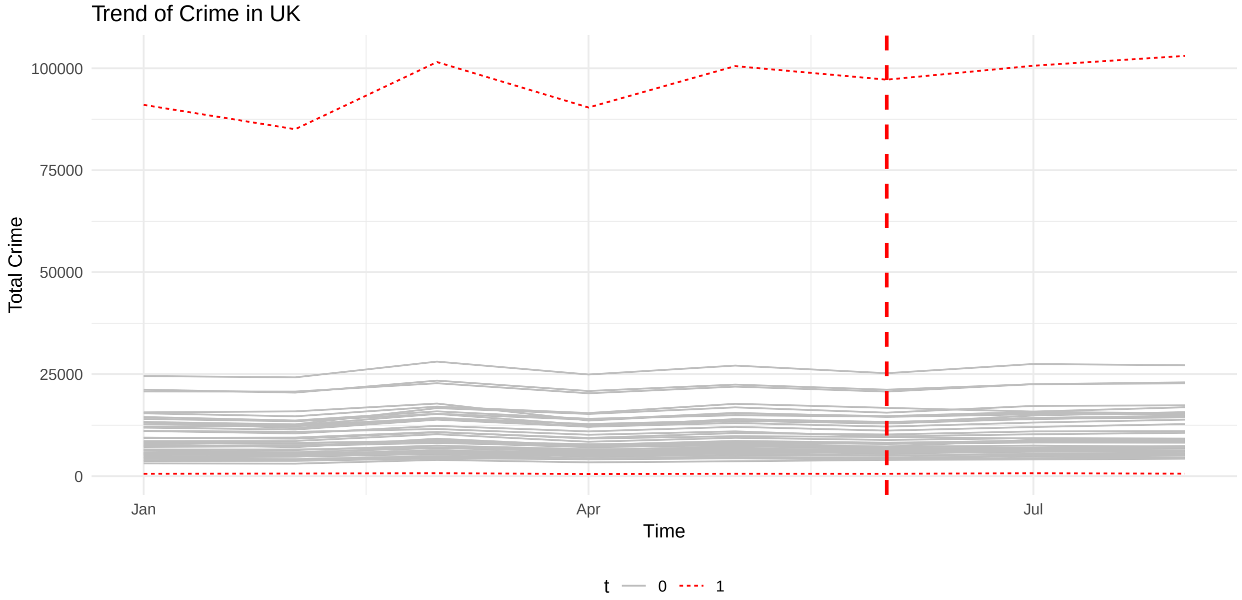
Next we are creating a flag which indicated when the Olympcis events started. The games began in late July 2012 and concluded in mid-August. So for our initial analysis we flag those two months as the dates when Olympics happened.

```r
#Treatment flag
df_flag =  df_agg %>% filter((date <"2012-09-01")) %>%
           filter((date >="2012-01-01")) %>%
           mutate(after = if_else(Month == '2012-07' | Month == '2012-08', 1,0 ))

df_flag =  df_flag %>% mutate(t_flag = if_else(treatment == 1 & after ==1, 1,0 ))
```

Plot of the trend of Crime across Police forces in UK

In the graph the red dotted lines are the two police forces that we are considering to be in our test group as they include locations were the Olympics events occured in UK.

```
ggplot(data = df_flag) +
  aes(x = date, y = value, group = Force,  color = t,linetype=t, fill = t) +
  geom_line() +
  labs(title = "Trend of Crime in UK",
    x = "Time",
    y = "Total Crime") +
  theme_minimal()+
  theme(legend.position="bottom") +
  geom_vline(xintercept=as.numeric(df_flag$date[6]),linetype="dashed",color="red",size=1) +
  scale_color_manual(values = c("1" = "red", "0" = "grey"))
```



The above graph shows that the Metropolitan Police Service Force has one of the highest crime rates in whole of UK. This is worrying because it might then be difficult to find suitable controls among the other Police forces in the US.

Analysis

Difference in difference

To understand if Olympics in London has an impact on crime we cannot do a simple pre post difference because there could be time-invariant confounds within groups ie Group Fixed Effect or inter-temporal confounds that apply to everyone ie Time Fixed Effect or both. Therefore we have to use a technique called Difference in differences. In short, in difference in differences we,

- First we take the after-before difference in each group.
- Next, calculate the difference in those inter-temporal differences.

In our context Difference in differences (DiD) helps determine the causal effect of a treatment (London Olympics) for a location (Greater London) by comparing trends with similar other areas (rest of the areas in the UK excluding the abutting areas of Greater London).

The assumptions for Difference in difference are:

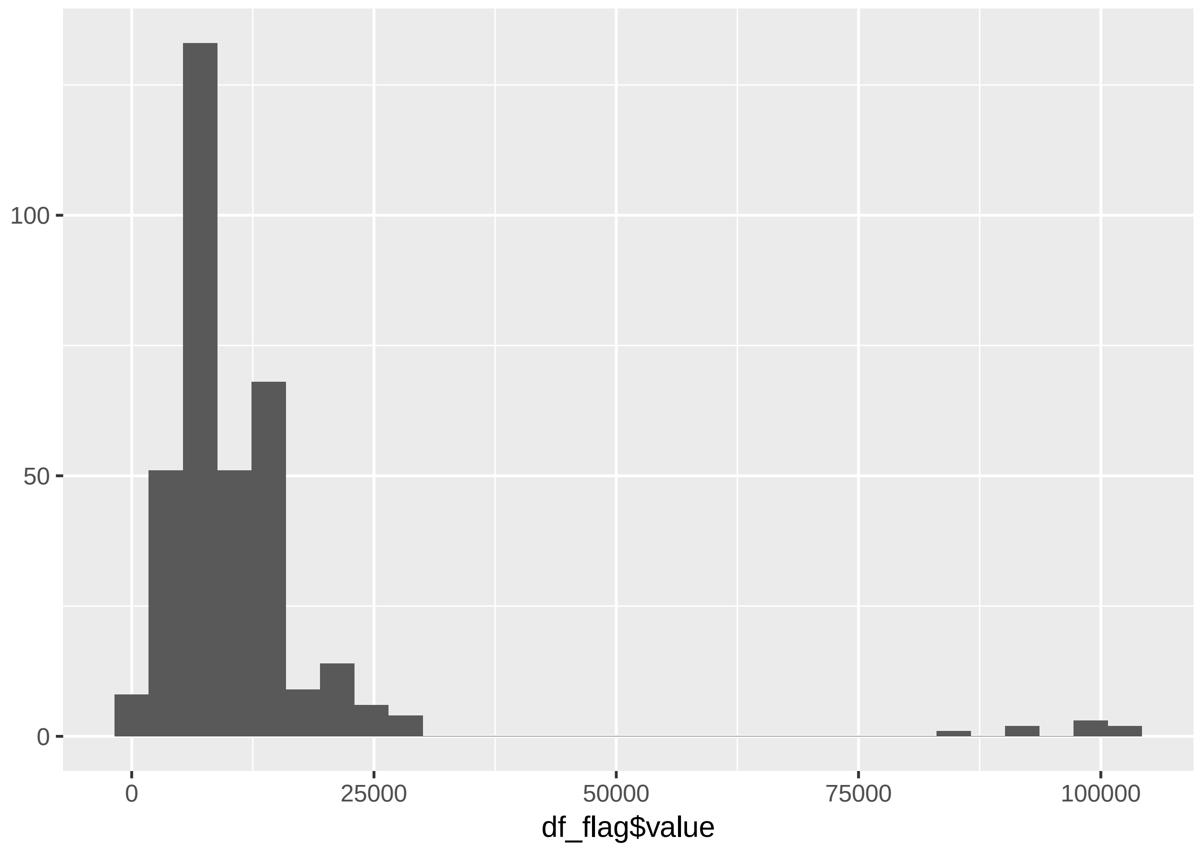
- We assume parallel trends between the two groups on the outcome variable.
- We also assume that treated subjects are not influencing control subjects.

```r
# Difference in difference analysis
did_basic <- lm(data=df_flag,value~treatment*after)
did_log_basic <- lm(data=df_flag,log(value+1)~treatment*after)

stargazer(did_basic,did_log_basic,title="DiD Estimates",
          column.labels=c("Total Crime", "Log(Total Crime)"),type="text")
```

```
##
## DiD Estimates
## =============================================================
##                                     Dependent variable:
##                                ------------------------------
##                                    value      log(value + 1)
##                                 Total Crime  Log(Total Crime)
##                                     (1)            (2)
## -------------------------------------------------------------
## treatment                      37,646.570***      -0.128
##                                 (3,394.681)      (0.214)
##
## after                              731.940         0.079
##                                 (1,447.497)      (0.091)
##
## treatment:after                  3,058.226        -0.006
##                                 (6,789.362)      (0.428)
##
## Constant                       9,807.012***      9.065***
##                                  (723.748)       (0.046)
##
## -------------------------------------------------------------
## Observations                         352            352
## R2                                  0.330          0.004
## Adjusted R2                         0.324         -0.005
## Residual Std. Error (df = 348)  11,489.150        0.725
## F Statistic (df = 3; 348)        57.097***        0.422
## =============================================================
## Note:                             *p<0.1; **p<0.05; ***p<0.01
```

Based on this analysis we see that Olympics caused a 3,058 increase in crime but the standard errors are too large and the treatment estimates are not significant. Lets now look at the distribution of the crime rate to see if we should use log of total crime instead.

```r
qplot(df_flag$value)
```



As expected becuase of the extremely large values of the Metropolitan Police Service crime rate the distribution is skewed which seems to suggest that using the log of crimes would be more approriate. Going back to the previous analysis if we are to consider the result of the log of crime we see that Olympics in London caused a .6% decrease in crime. This result is also not significant.

Now we will rerun the difference in difference analysis adding dummies for police force and month

```
#### Fixed effects
did_fe <- felm(data=df_flag,value~treatment + after + after:treatment|Force)

did_sfe_tfe <- felm(data=df_flag,value~treatment + after + after:treatment|Force+Month)

did_log_sfe_tfe <- felm(data=df_flag,log(value+1)~treatment + after +
                          after:treatment|Force+Month)

stargazer(did_fe,did_sfe_tfe,did_log_sfe_tfe,type="text",
          title="DiD + Time & Subject FEs",
          column.labels = c("Subject FEs","Subject + Time FEs","Log Subject + Time FEs"))
```

```
##
## DiD + Time & Subject FEs
## ==================================================================================
##                                         Dependent variable:
##                     -------------------------------------------------------------
##                                    value                         log(value + 1)
##                         Subject FEs      Subject + Time FEs Log Subject + Time FEs
##                             (1)                 (2)                   (3)
## ----------------------------------------------------------------------------------
## treatment
##
##
## after                    731.940***
##                          (143.043)
##
## treatment:after         3,058.226***        3,058.226***            -0.006
##                          (670.930)           (581.709)             (0.024)
##
## ----------------------------------------------------------------------------------
## Observations                352                 352                  352
## R2                         0.994               0.996                0.997
## Adjusted R2                0.993               0.995                0.997
## Residual Std. Error 1,135.367 (df = 306) 984.384 (df = 300)    0.041 (df = 300)
## ==================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

After adding the dummies we see that the treatment for log of crime is still not significant but the effect on absolute value of crime has become significant and stable between the Force and month fixed effects. Based on this analysis the Olympics in London caused a 2798 unit increase in overall crime. Before we can analyze the result furthur we need to check if the assumptions of difference in difference methods are being met.

The method requires both the treatment and control to have parallel trends so that the change in the slopes of the lines after the treatment happens can be considered to be the effect of the treatment. The control group needs to be a good counterfactual for the treated group.

To evaluate parallel trends we do dynamic difference in differences, the intuition is that we are estimating many difference-in-differences regressions all at once. The omitted month is our “pre” period now, and we estimate diff-in-diff relative to every other period, comparing treatment group with control.

Based on this analysis we should only see significant diff-in-diff (treatment effect) estimates for DiD’s related to post treatment periods, and not for any pre-treatment periods. If we see that the treatment group was different in its differences from the control group in periods leading up to treatment, it implies parallel trends has been violated. We can never guarantee parallel trends has been met. Failure to find significance != proof of equivalence; we never accept the null hypothesis of 0 differences.

Below is the analysis for the same.

```r
df_flag$datef <- factor(df_flag$date )
levels(df_flag$datef)
```

```
## [1] "2012-01-01" "2012-02-01" "2012-03-01" "2012-04-01" "2012-05-01"
## [6] "2012-06-01" "2012-07-01" "2012-08-01"
```

```r
df_flag <- within(df_flag, datef <- relevel(datef,ref=6))
levels(df_flag$datef)
```

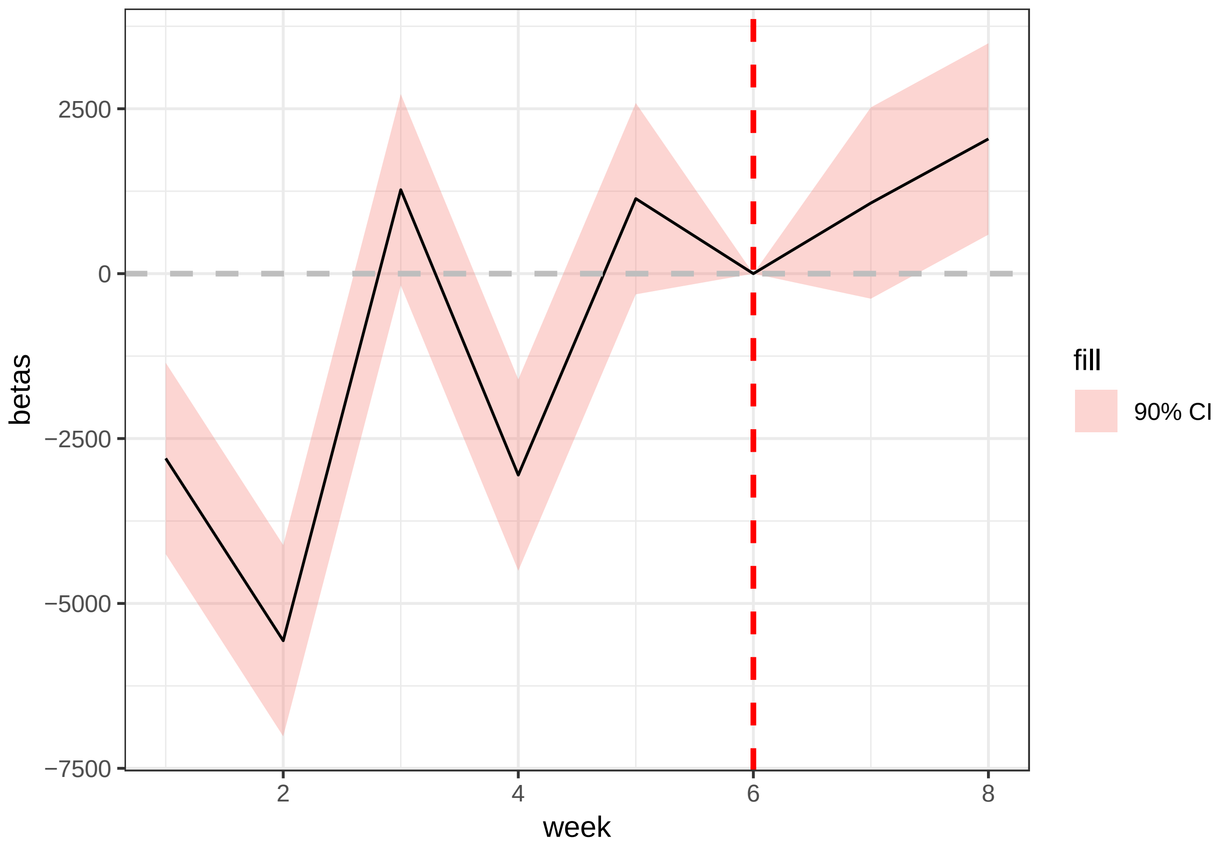
```
## [1] "2012-06-01" "2012-01-01" "2012-02-01" "2012-03-01" "2012-04-01"
## [6] "2012-05-01" "2012-07-01" "2012-08-01"
```

```r
#Now let's run the dynamic model.
did_dyn_sfe_tfe <- felm(data=df_flag,value~treatment*datef|Force+datef)

did_dyn_log_sfe_tfe <- felm(data=df_flag,log(value+1)~treatment*datef|Force+datef)
stargazer(did_dyn_sfe_tfe,did_dyn_log_sfe_tfe,type="text",title="Dynamic DiD + FEs",
          column.labels = c("Dynamic DiD + FEs","Dynamic Log DiD + FEs"))
```

```
##
```

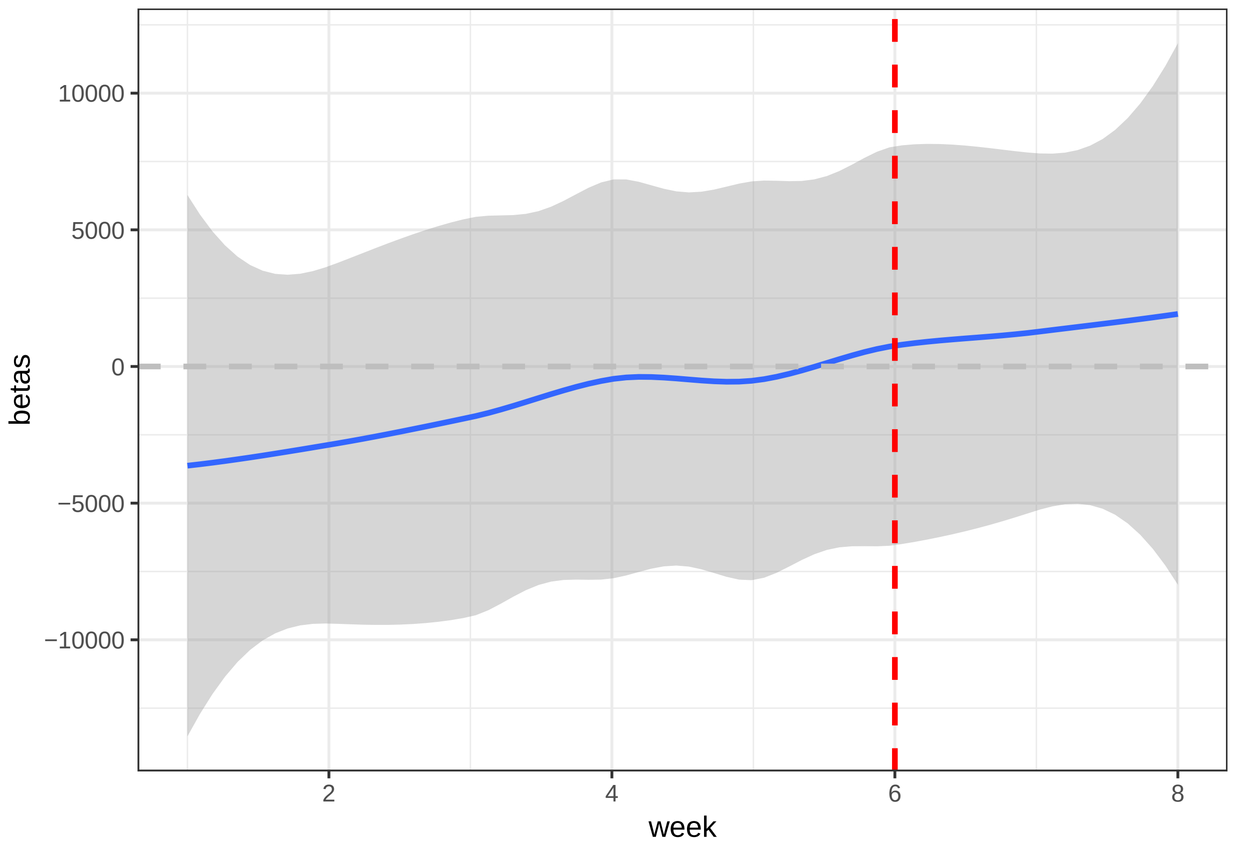
```
## Dynamic DiD + FEs
## ======================================================================
##                                           Dependent variable:
##                                 ---------------------------------------
##                                       value          log(value + 1)
##                                 Dynamic DiD + FEs Dynamic Log DiD + FEs
##                                        (1)                 (2)
## ----------------------------------------------------------------------
## treatment
##
##
## datef2012-01-01
##
##
## datef2012-02-01
##
##
## datef2012-03-01
##
##
## datef2012-04-01
##
##
## datef2012-05-01
##
##
## datef2012-07-01
##
##
## datef2012-08-01
##
##
## treatment:datef2012-01-01        -2,800.881***           0.004
##                                    (881.735)            (0.042)
##
## treatment:datef2012-02-01        -5,564.786***           0.024
##                                    (881.735)            (0.042)
##
## treatment:datef2012-03-01          1,271.833             0.020
##                                    (881.735)            (0.042)
##
## treatment:datef2012-04-01        -3,053.571***          -0.032
##                                    (881.735)            (0.042)
##
## treatment:datef2012-05-01          1,136.976            -0.025
##                                    (881.735)            (0.042)
##
## treatment:datef2012-07-01          1,070.619             0.029
##                                    (881.735)            (0.042)
##
## treatment:datef2012-08-01         2,042.357**           -0.044
##                                    (881.735)            (0.042)
##
## ----------------------------------------------------------------------
## Observations                               352                 352
## R2                                        0.997               0.997
## Adjusted R2                               0.996               0.997
## Residual Std. Error (df = 294)           861.462              0.041
## ====================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

```
# Let's plot the coefficients and confidence intervals now...
# First, we pull out the coefficients and standard errors from the output object
# of the regression.
coefs_ses <- cbind(did_dyn_sfe_tfe$coefficients,did_dyn_sfe_tfe$se)[]
View(coefs_ses)

# We just want to keep the dynamic treatment estimates...
coefs_ses <- coefs_ses[c(9:nrow(coefs_ses)),]
View(coefs_ses)

# We don't get estimates for the reference period so let's plug in 0's for it.
coefs_ses <- data.frame(rbind(coefs_ses[c(1:5),],c(0,0),coefs_ses[c(6:nrow(coefs_ses)),]))
row.names(coefs_ses)[6] <- "treatment:datef2012-06-01"
#View(coefs_ses)
#row.names(coefs_ses)

# Now let's make it a dataframe, and construct our confidence interval.
# This is 90% confidence interval.
coefs_ses <- data.frame(coefs_ses)
names(coefs_ses) <- c("betas","ses")
coefs_ses$ub_90 <- coefs_ses$betas+1.645*coefs_ses$ses
coefs_ses$lb_90 <- coefs_ses$betas-1.645*coefs_ses$ses

# Let's create a time period indicator for the graph that's easy to read.
coefs_ses$week <- seq(1:nrow(coefs_ses))

# Let's connect the estimates with a line and include a ribbon for the CIs.
plot <- ggplot(coefs_ses, aes(week,betas)) +
        geom_ribbon(aes(ymin=lb_90,ymax=ub_90,fill="90% CI"),alpha=0.3)

plot <- plot + geom_line() +
        geom_hline(yintercept=0,linetype="dashed",color="gray",size=1) +
        geom_vline(xintercept=6,linetype="dashed",color="red",size=1) + theme_bw()
plot
```



From the above graph we see that the crime rate trends did not match well during the pre period. The normalized slopes for both treatment and control group should equate to 0 but in our graph they are varying widely in the pre period.

```
# Fitting a smoothed line (lowess curve) to the coefficients, though this is less
# "rigorous" and more of a heuristic.

plot <- ggplot(coefs_ses, aes(week,betas))
plot <- plot + geom_smooth() +
        geom_hline(yintercept=0,linetype="dashed",color="gray",size=1) +
        geom_vline(xintercept=6,linetype="dashed",color="red",size=1) + theme_bw()
plot
```



Paralled trend assumptions is very important for difference in differences because the control group needs to be a good counterfactual for the treated group.As our data during the pre period has violated the paralled trends assumptions we are now trying another method called synthetic control.

Synthetic Control

Synthetic control method is generally used when pre-treatment Trends are Violated, ie our set of controls is not useful for DiD, so we cannot just trust the DiD estimate of the treatment effect.

The synthetic control method is a statistical method used to evaluate the effect of an intervention in comparative case studies. It involves the construction of a weighted combination of groups used as controls, to which the treatment group is compared. This comparison is used to estimate what would have happened to the treatment group if it had not received the treatment. Unlike difference in differences approaches, this method can account for the effects of confounders changing over time, by weighting the control group to better match the treatment group before the intervention. Another advantage of the synthetic control method is that it allows us to systematically select comparison groups.

To be able to perform Synthetic control we need the below conditions to be met

- Need Repeated Observations of Treated and Control, in the Presence and Absence of Treatment.

- No Interference,we must not have any interference between treated and control units, else the counterfactual will be wrong.

To ensure that the algoithem can construct an adequate control we need to give it more pre period data (The R package Gsynth actually requires atleast 8 periods of data in the pre period). Since our previous data was only from 2012 we downloaded more historical data from 2011-06 - 2011-12 seperately. So now our entire data is from 2011-06 - 2012-10. All other information remains the same.

Load Data

```
myfiles2 = read.csv('myfiles1.csv')
myfiles21 = read.csv('myfilesold.csv')
```

Data Transformations

Similar to difference in differences the data used for the analysis comes from UK's police website. Data recorded about monthly criminal activity at a police force level from 2011 through 2013 for Great Britain. It is separated into a variety of types of offenses. These are violence against the person, sexual crimes, robbery, burglary, theft, fraud, damage, drugs and other. Each of these categories provided can be parsed out into further subcategories. However, we combined the various categories into a total crime value as we were interested to understand the impact of the Olympics on total crime.

```
df = myfiles2 %>%
     mutate(treatment = if_else(Force == 'City of London Police'|
                                Force == 'Metropolitan Police', 1,0))
df1 = myfiles21 %>%
      mutate(treatment = if_else(Force == 'City of London Police'|
                                   Force == 'Metropolitan Police',1,0))

df1$vaue = df1$Burglary + df1$Robbery +df1$Vehicle.crime + df1$Violent.crime +
          df1$Anti.social.behaviour+ df1$Other.crime

df1_agg = df1 %>% group_by (Force, Month, treatment) %>% summarise(value = sum(vaue))

df$vaue =  df$Burglary + df$Robbery +df$Vehicle.crime + df$Violent.crime +
            df$Anti.social.behaviour + df$Criminal.damage.and.arson + df$Shoplifting+
            df$Other.theft + df$Drugs + df$Public.disorder.and.weapons + df$Other.crime

df_agg = df %>% group_by (Force, Month, treatment) %>% summarise(value = sum(vaue))
df_agg = rbind(df1_agg, df_agg )
df_agg$date = as.Date(paste( df_agg$Month,"-01",sep=""))
df_agg$t = as.character( df_agg$treatment)
```

Next we are creating a flag which indicated when the Olympcis events started. The games began in late July 2012 and concluded in mid-August. Although the games were held over a short period, it is fair to assume that the time period over which effects that the games might have had could be observable over a longer period. Hence we are trying to see if crime increased during the months of July to October 2012.

```
df_flag =  df_agg %>% filter((date>"2011-05-01")) %>%
            filter((date <"2012-11-01")) %>%
            mutate(after = if_else(Month == '2012-07' | Month == '2012-08'|
                                  Month == '2012-09'| Month == '2012-10',1,0))
```

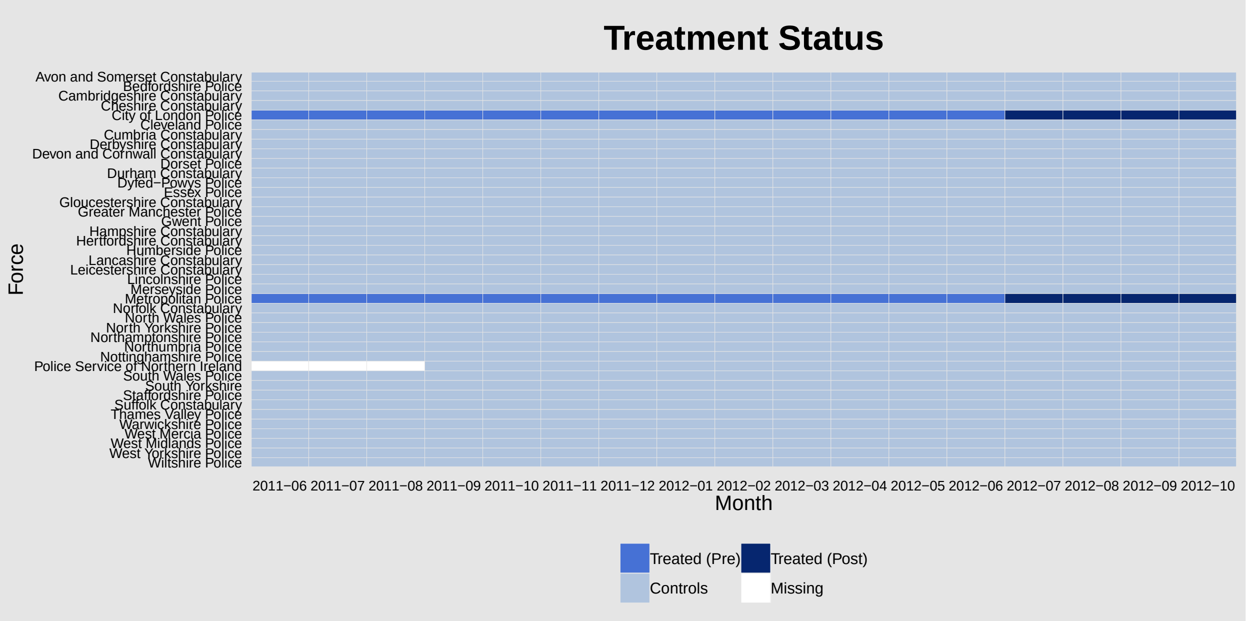
To avoid spillover impact the areas surrouding the test force districts were removed (Condition two as mentioned above).The same reason we don't want interference or spillovers between experiment groups. We require that the potential outcome observation of one unit should be unaffected by the particular assignment of treatments to other units. Interference amounts to violating this assumption. Hence we removed the Force areas geographically closest to the location of Olmpics from the analysis.

```
df_flag =  df_flag %>% mutate(t_flag = if_else(treatment == 1 &  after ==1, 1,0 )) %>%
          filter((Force != "Surrey Police" )) %>%
          filter((Force != "Sussex Police" )) %>%
          filter((Force != "Kent Police" ))
```

```
df_flag$logvalue =  log(df_flag$value+1)
```

Before we conduct any statistical analysis, it is helpful to visualize the data structure and/or spot missing values (if there are any). We can easily do so with the help of panelView. The following figure shows that: (1) there are 2 treated units and 42 control units; (2) the treated units start to be treated in period 2012-07; and (3) there are missing values for one Police force.

```
panelView(value~t_flag, data=data.frame(df_flag),index=c("Force","Month"),
          na.rm=TRUE,outcome.type="continuous",treatment=TRUE)
```



Now lets perform synthetic control. Using pre-period data from other states, we are building a model that assigns fixed weights to each control police force, and arrives at a weighted average that closely resembles crime rate in London before the Olympics.

We will then be able to use the resulting model to synthesize what crime in London would have looked like in post period, too (absent treatment).

Now we run the gsynth algorithm. The first variable on the right-hand-side of the formula is a binary treatment indicator. The rest of the right-hand-side variables serve as controls. The index option specifies the unit and time indicators.

```
out <- gsynth(formula=(logvalue)~t_flag  , data = df_flag,
              index = c("Force","Month"), force = "two-way",inference="nonparametric",
              CV = TRUE, se = TRUE, r = c(0,10),nboots=60, min.T0=10, parallel=TRUE,
              cores=8,na.rm=FALSE)
```

```
## Parallel computing ...
## Cross-validating ...
##  r = 0; sigma2 = 0.00246; IC = -6.00823; MSPE = 0.00545*
##  r = 1; sigma2 = 0.00142; IC = -6.01487; MSPE = 0.00546
##  r = 2; sigma2 = 0.00103; IC = -5.81926; MSPE = 0.00614
##  r = 3; sigma2 = 0.00085; IC = -5.50795; MSPE = 0.00734
##  r = 4; sigma2 = 0.00076; IC = -5.13588; MSPE = 0.00631
##  r = 5; sigma2 = 0.00068; IC = -4.78590; MSPE = 0.00947
##  r = 6; sigma2 = 0.00058; IC = -4.49592; MSPE = 0.01223
##  r = 7; sigma2 = 0.00049; IC = -4.25056; MSPE = 0.00406*
##  r = 8; sigma2 = 0.00041; IC = -4.02189; MSPE = 0.00604
##  r = 9; sigma2 = 0.00036; IC = -3.77452; MSPE = 0.01478
##  r = 10; sigma2 = 0.00031; IC = -3.54340; MSPE = 0.00877
##
##  r* = 7
##
##
Bootstrapping ...
##
```

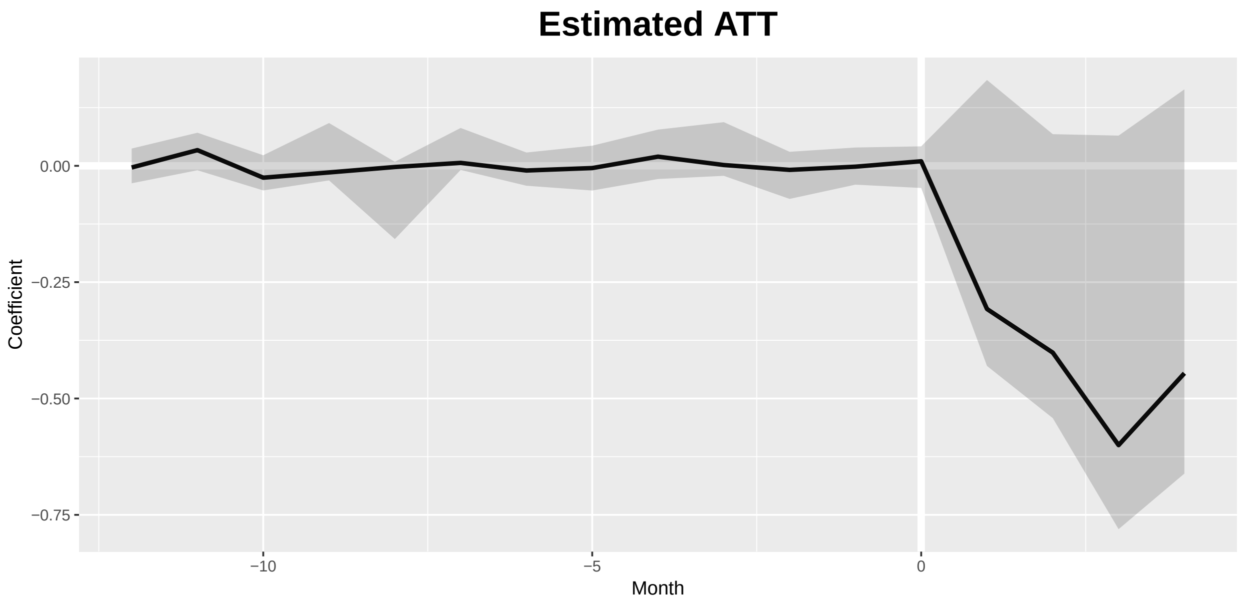
In the output sigma2 stands for the estimated variance of the error term; IC represents the Bayesian Information Criterion; and MPSE is the Mean Squared Prediction Error. The cross-validation procedure selects an r that minimizes the MSPE.

```
out
```

```
## Call:
## gsynth.formula(formula = (logvalue) ~ t_flag, data = df_flag,
##     na.rm = FALSE, index = c("Force", "Month"), force = "two-way",
##     r = c(0, 10), CV = TRUE, se = TRUE, nboots = 60, inference = "nonparametric",
##     parallel = TRUE, cores = 8, min.T0 = 10)
##
## Average Treatment Effect on the Treated:
##  ATT.avg   S.E. CI.lower CI.upper p.value
##  -0.4388 0.1694   -0.606  0.06735  0.2667
##
##    ~ by Period (including Pre-treatment Periods):
##          ATT    S.E.  CI.lower CI.upper p.value n.Treated
## 1  -0.003580 0.01918 -0.037531 0.037132  0.7667         0
## 2   0.033742 0.02197 -0.009549 0.071128  0.3000         0
## 3  -0.025494 0.02127 -0.052860 0.022717  0.5000         0
## 4  -0.014223 0.03444 -0.031355 0.091911  0.8000         0
## 5  -0.002537 0.04984 -0.157394 0.008655  0.2667         0
## 6   0.006342 0.02424 -0.008813 0.081308  0.2000         0
## 7  -0.010063 0.01802 -0.042796 0.028571  0.6333         0
## 8  -0.004828 0.02490 -0.053051 0.043215  0.9333         0
## 9   0.019690 0.03198 -0.028354 0.077831  0.7667         0
## 10  0.001728 0.02848 -0.021503 0.093806  0.4667         0
## 11 -0.008736 0.02644 -0.071089 0.030146  0.4000         0
## 12 -0.001847 0.02271 -0.040492 0.039346  0.8667         0
## 13  0.009806 0.02222 -0.047596 0.041840  0.8333         0
## 14 -0.307666 0.14329 -0.430029 0.184395  0.3667         2
## 15 -0.401391 0.15581 -0.542154 0.068033  0.3333         2
## 16 -0.600188 0.22009 -0.780751 0.064927  0.1333         2
## 17 -0.445776 0.19212 -0.661327 0.164284  0.3333         2
```

Now lets use the estimates from the above analysis and visuale the average treatment effect on the treatment units.

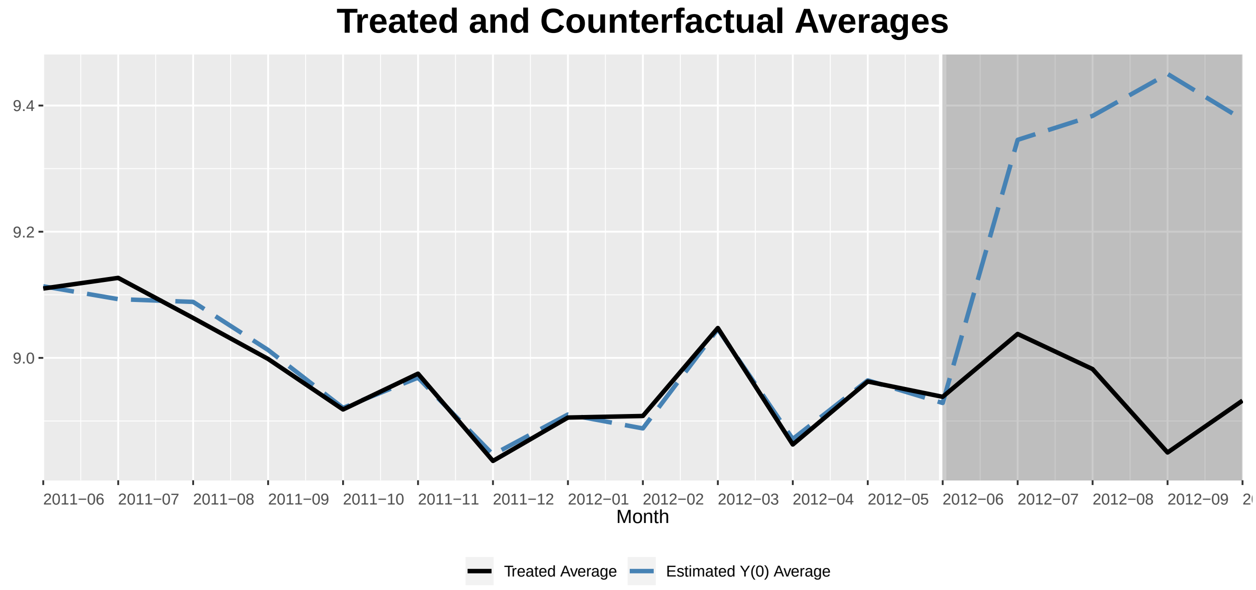
```
plot(out)
```



From the graph above, the crime rate for the treatment units is shown to be decreasing but it has a very large confidence interval. Therefore it is difficult to conclude with statistical significane whether London Olympics caused a reduction in crime rates in Greater London.

When we look at the treated and counterfactual averages we see that the result below shows that there is actually a reduction in the crime rates in London after the Olympics. The black line denotes the true crime rates in Greater London while the dashed line represents the predicted crime rates in London.

```r
plot(out,type="counterfactual",xlim=c(-24,44))
```



Conclusion

Under the assumptions and limitations of our analysis we see that crime decreases but we are unable to find statistically significant causal result on the effect of London Olympics on crime. Further analysis listed in the sections below might help bolster the claim that crime went down during the London Olympics.

To bolster the analysis we recommend doing either one or both of the following analyses

Strategy 1:
The first approach we could do is to collect more information about London and other control cities in the UK. As mentioned in the above sections if we are able to gather more data in terms of the monthly police force size or other economic indicators our synthetic control might be able to give us more confident results and we might be able to bolster our existing analysis enough to give get a conclusive answer.

Strategy 2: Analyze London's crime rate in comparison to similarly populous cities across the world for better control. In the current method one of the drawbacks is that the crime rate of the London which is the test unit is very different from the crime rate in the other parts of the UK which act as the control units. This limitation can be overcome if were to use other major cities, possibly cities that have hosted the Olympics before to ensure that they would be act as suitable controls.