2012 London Olympics' Impact on Income Inequality

Dou Liu, Johnson Wu, Eddie Shim, Aryan Pandhi

July 23, 2020

1 Introduction

London is one of the most prosperous cities of England; however, like most other cities in the world it has socioeconomic divide. The city is divided into 33 boroughs that could be classified into socioeconomic classes based on indicators such as income, wealth, education, and occupation. Hosting the Olympics games benefits a city in numerous ways including increased tourism, improved infrastructure, as well as an increase in the country's global trade and stature. However, these benefits may not apply to every region equally or the same reason. There could be factors such as the location and the socioeconomic class of a region of the city that could dictate how beneficial the games were to that part of the city. Therefore, the aim of the project is to find out how the 2012 London Olympic games affected the different regions of the host city.

For the scope of this project, we measure the real impact of the Olympic games as the change in income income earned, and the research questions that stem from this idea are: **How does the 2012 London Olympics affect the income of people in different boroughs of London? What are the intermediate factors that cause this change in income?**

2 Regional Overviews

2.1 United Kingdom

Since a global event of such magnitude is expected to have an impact on a large region, we begin with a macroscopic view of how the Olympics affect the United Kingdom. As a proxy for economic impact, we looked at international tourism data in the UK.

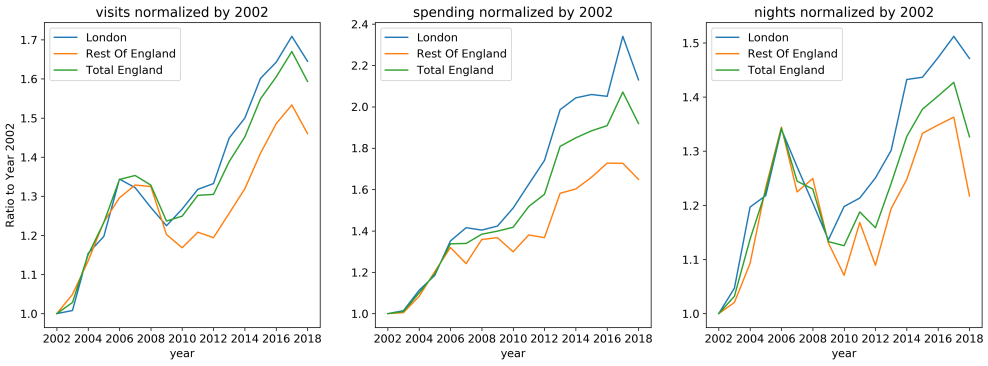


Figure 1: The change in number of visits, spending, and nights of international visitors to England

From Figure 1 shown above, we can interpret the following:

- All regions experienced a dip in 2008-2009 (likely due to the global housing recession). 2012 is a rough inflection point in each graph, where each feature experiences robust, consistent growth from 2012 - 2017.
- London had been moving at the same rate as the rest of England till 2010, then experiences a faster rate in growth than other regions.

2.2 London

London had a greater increase in international visits and spending (post 2010) compared to the rest of England. The Olympic games might have played a certain role in these changes. We now focus on how this increase in spending and other Olympic games factors affected the different boroughs of London. We first look at the distribution of the median income in the different regions of London.

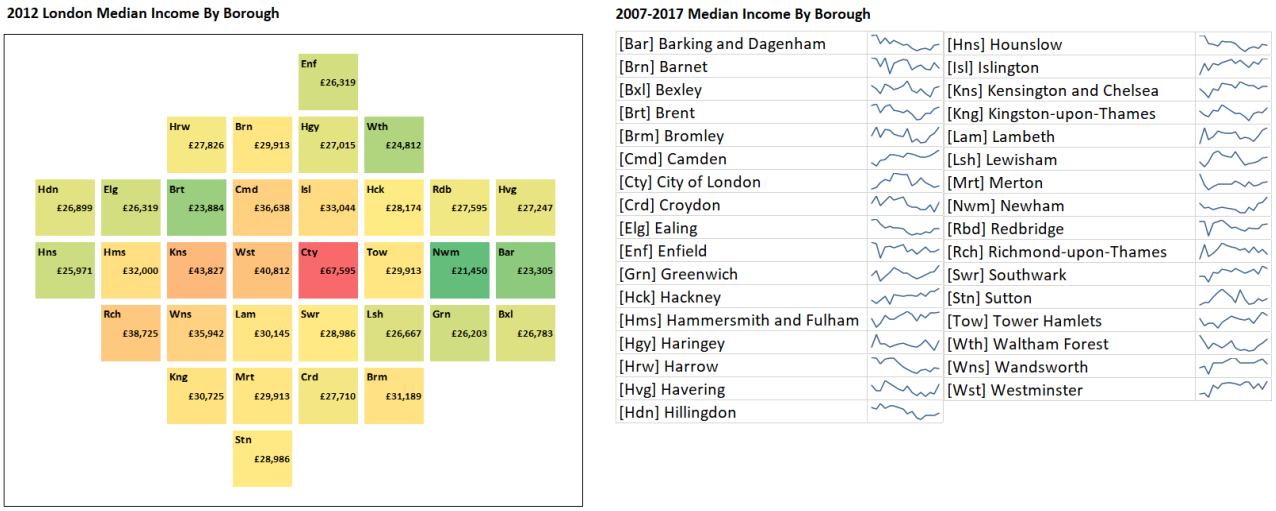


Figure 2: Median income across London's boroughs

From Figure 2 we can observe that in 2012 there was high income inequality in London. The city of London and the boroughs to its west, including Kensington & Chelsea and Westminster, have the highest income, whereas the boroughs in East London, including Newham and Barking & Dagenham, have much lower income.

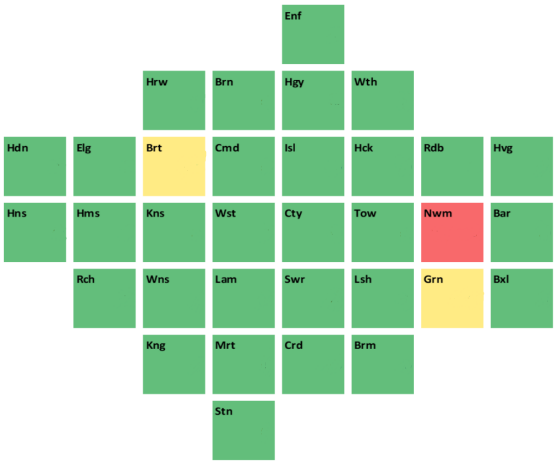


Figure 3: Ticket revenue from where the Olympic events were held

From Figure 3 we can observe that most of the revenue was generated by the events held in Newham. This is because the Olympic facilities were chosen to be built in Newham, including the Queen Elizabeth Olympic Park. Consequently, Newham experienced a sharp growth in median income for several years. In the next sections, we dive deeper into these metrics and support our intuition using statistical models.

3 Data Preprocessing

The following details how we approached data processing and feature engineering required before modeling:

- First we normalized typos in the datasets across employment activity data in London, South West, South East. Furthermore, we imputed missing income data in 2008 by taking the mean of neighboring years for *london-taxpayer-income* data.
- To account for inflation, we calculated real values each year by normalizing the income, earnings and other quantities that are related to the currency.
- We mapped and joined data frames such as the underground stations, the venues for Olympics games, in order to group them into their respective boroughs.
- We discard the borough of City of London for our analysis considering that there are so many missing values for that borough.

There are several external datasets we also use besides the datasets provided by the Data Open. *Jobs-and-Job-Density.csv, business-survival-rates.csv, business-demographics.csv, Qualifications-of-working-age-NVQ.csv.* All of these datasets are from London government website `data.london.gov.uk/dataset`[2]

4 Difference in Difference method for London boroughs

After what we saw for the different trends of income in different boroughs in Figure 2, we tackle the following question– **how can we isolate the economic impact of the Olympic games across boroughs over time?** In order to examine this question, we need to remove external effects in order to obtain the **net effects** of the Olympics games.

We use a **difference in difference (DID) model** to examine the net effects. By defining the Olympics games as the "treatment". The model selects the treatment group which may be affected by the treatment and the control group which are not affected. The DID regression model is defined as following:

$$Y_{it} = \beta_0 + \beta_1 D_i + \sum_{j=t_1}^{n} \beta_{2t} I_{t=j} + \sum_{j=t_1}^{n} \rho_t D_i I_{t=j} + \epsilon_{it} \tag{1}$$

- Y_{it} is the dependent variable (borough's medium income) for borough i at time t
- D_i is the dummy variable showing whether the sample will receive the treatment, $D_i = 1$ for treatment sample i and 0 for control sample. This term $\beta_1 D_i$ are characteristics of individuals that do not change over time.
- I_t is also the dummy variable. It is the indicator function, meaning which period the sample is in. This term shows the time fixed effect. In our work, we divide the whole time range into 4 intervals ($n = 3$ when excluding the pre-treatment interval).
  - $T \leq 2008$: Before the treatment. None of the boroughs are affected by the Olympics, since it is far from beginning.
  - $2009 \leq T \leq 2011$: Preparation for the treatment: We know that before the Olympics games actually happened, there will be possible investments, constructions which may benefits local residents.
  - $2012 \leq T \leq 2014$: During the treatment and short-term effect.
  - $2015 \leq T \leq 2017$: Long-term effect.

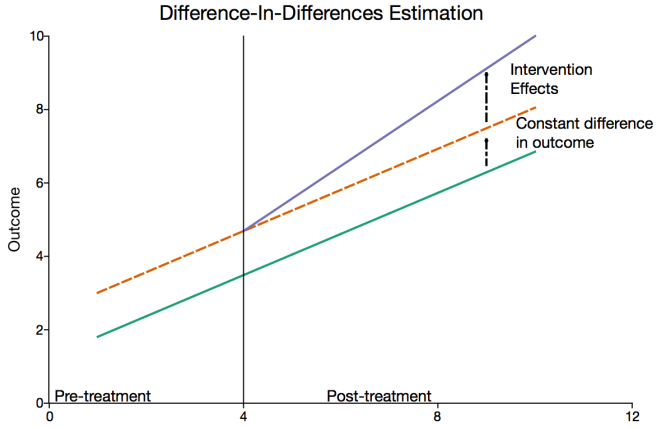
  For the interval before the treatment, the indicator variables are zero. We use that part to validate our assumption for this model, seen below.
- The term $D_i I_{t=j}$ is the term which showing the net effect of the treatment.

the net treatment effect of Olympic on median income of borough i in year t

$$\begin{aligned} E(\delta Y_{it}|D_i = 1) - E(\delta Y_{it}|D_i = 0) &= (E(Y_{it}|D_i = 1, t = t) - E(Y_{it}|D_i = 1, t = 0)) \\ &\quad - (E(Y_{it}|D_i = 0, t = t) - E(Y_{it}|D_i = 0, t = 0)) \\ &= \rho_t \end{aligned} \tag{2}$$

The coefficient ρ_t reflects **the net effect of the Olympic games for the time t compared to the time before the treatment (before 2009)** and is the parameter we want to obtain for this analysis. The magnitude of coefficient is the net effect reflected in the form of **median income change of that borough**.

A simplified illustration of our model (DID with one period) could be shown as Figure 4: Before the treatment, even though the values of the treatment group and control group may be different, they have the same trend and that income difference will remain even after removing for income growth not caused by the Olympics. After the treatment, the difference here is the net effect of the treatment, which is shown as "intervention effects" in the graphic illustration and d in table illustration.



	Before the treatment	After the treatment	Difference
Treatment Group	a+b	a+b+c+d	c+d
Control Group	a	a+c	c
Difference	b	b+d	d (Difference in Difference)

Figure 4: Difference in difference model illustration. Left: graphic illustration. Right: Table illustration

This model relies on parallel trend analysis, meaning the treatment group and the control group should have the same trend before the treatment.

4.1 DID Model Assumptions

- We categorize London's boroughs into two main groups: the control group and the treatment group. And for the boroughs in control group is not or least impacted by the holding of Olympic in London. And the treatment group may experienced difference level of impact on median income due to Olympics due to different factors.
- parallel trend assumption: we assume that for all treatment groups, their corresponding selected control group has a parallel trend in the result of median income with them. It is the critical assumption for DID model to remove the bias that are caused by impact of non-Olympic factors on median income change.[1]
- For selecting control group. Most of boroughs in London are showing spike in underground traffic counts during 2012. (The graph shown in the appendix Figure 11, especially the uppermost brown curve which is Newham, the place where the Olympics Park is located). But there are six boroughs in London that don't have underground traffic growth. According to *uk-visits* data (graph shown in the appendix Figure 10), we know that about 90% of the international spends are from international visitors arriving London by air (rather than driving their own car). **We designate boroughs without underground traffic growth as our control group, under the assumption that they are the boroughs least economically impacted by the Olympic Games.** They are our control group candidates: [Lewisham, Bexley, Kingston upon Thames, Sutton, Croydon, Bromley].
- Different candidates within the control group don't necessarily have the same trend before the treatment because of their possible own development paths.
- We assume the impact of the Olympics begins at 2009, due to preparation required in the years preceding the Olympics.

Based on the assumptions above, **we select the top three candidates as the control group for every treatment sample.** The selection is based on the variance of the difference between their medium income percentage differences compared to Year 2002. Figure 5 is an illustration of the parallel trends before the treatment. As we can

see clearly in the plot, the control group (red-like curves) are more or less parallel to the treatment sample (green curve), while the non-control group differ a lot. That validates our assumption above.[1]

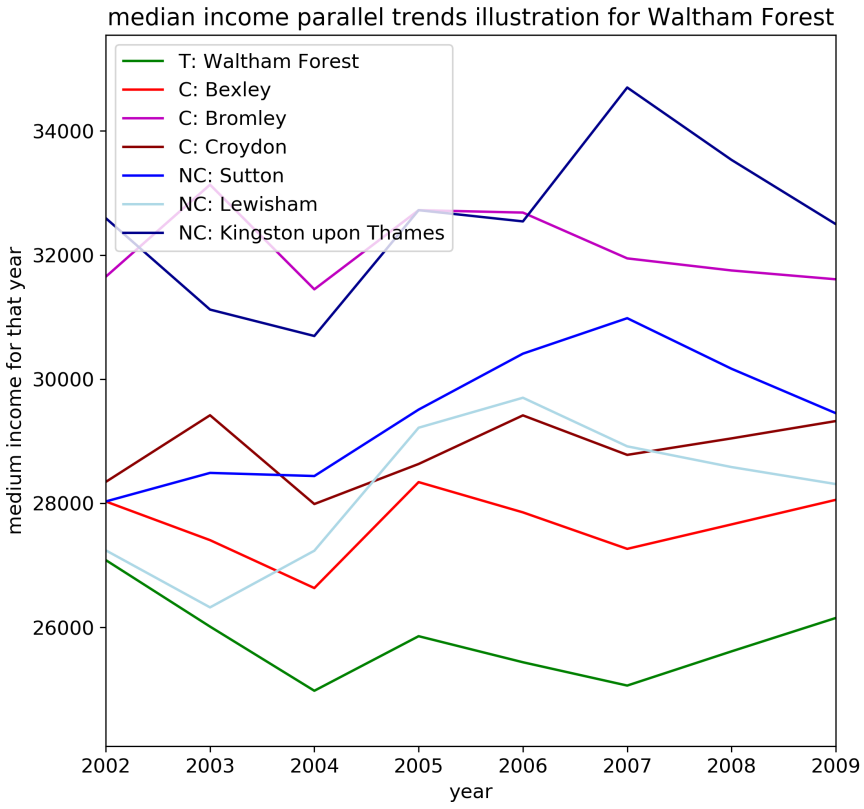


Figure 5: Parallel trends illustration before the treatment. The green curve is the treatment sample "Waltham Forest". The red-like curves starting with 'C' in the legend are the control samples. The blue-like curves starting with 'NC' in the legend are the non-control samples

4.2 DID Model Results

For every treatment sample and its control group with two samples, the DID model on the year interval 2002-2017 arrives at regression on the Equation 2:

$$\mathbf{X}\vec{f} = \mathbf{Y} \tag{3}$$

Here:

$$\vec{f} = [\beta_0, \beta_1, \beta_{2t}, \rho_t]^{\mathrm{T}} \tag{4}$$

where β_1 is the individual fixed effects, β_{2t} are the time fixed effects, and ρ_t is the net effect because of the treatment.

We apply linear regression on these treatments samples and obtained the **net effects for every treatment borough during each period**. They are visualized below.

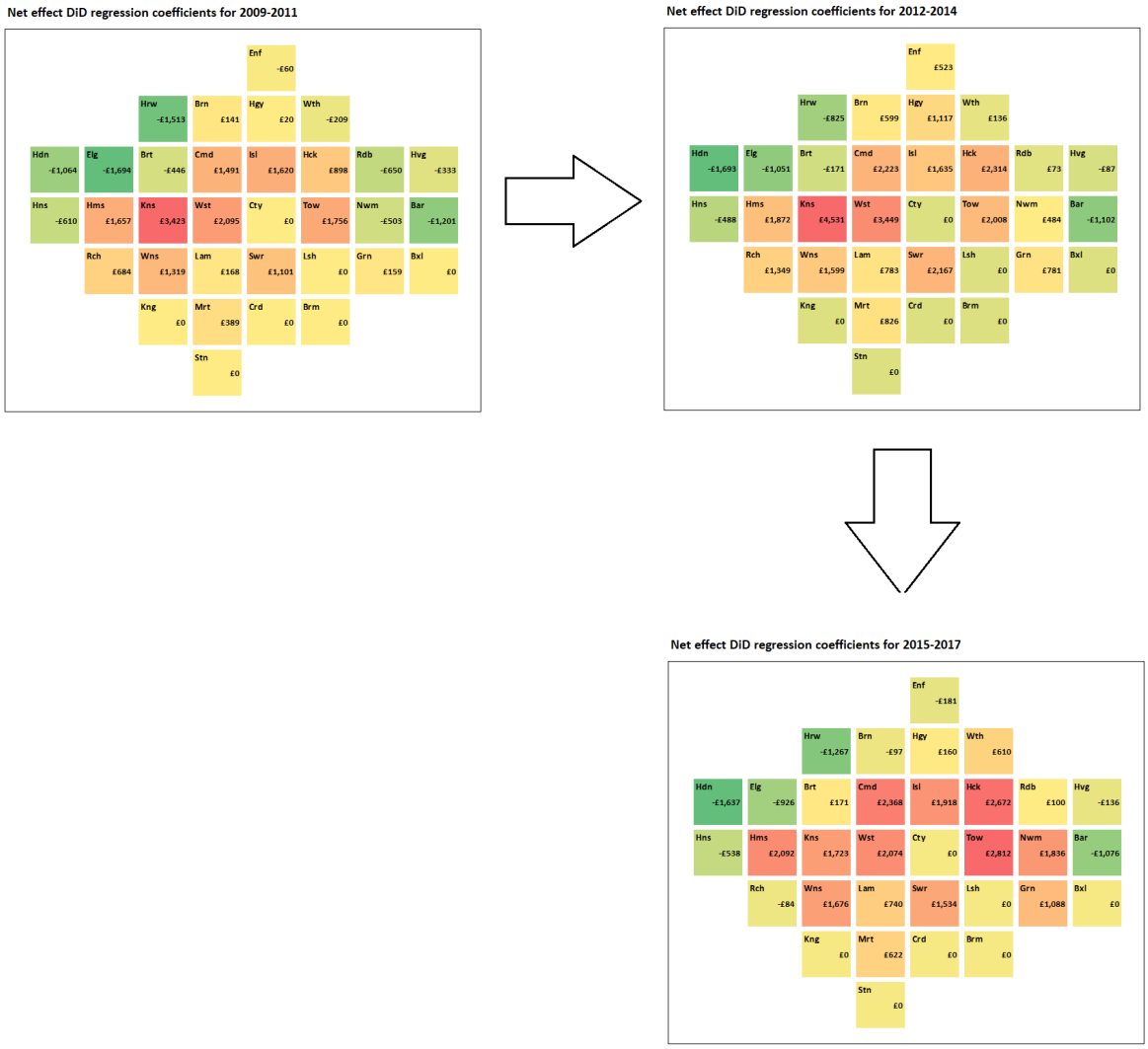


Figure 6: Difference in difference result for time interval 2009-2011, 2012-2014, 2015-2017

Things that need to notice and that we can read from the net effect DID regression results

- We can see there are several boroughs that with zero coefficient. That is based on our designation that they are the control group. For them, the net effect is zero.
- By comparing results for different periods, we can see that:
  - For **poor areas** like Newham (Nwm), where the Olympics held or Hackney(Hck) and Tower Hamlets(Tow), where is close to the host place of Olympic games, we can see, the net effect for them is significantly large. And also notice that the benefits for them are more significant in long-term (the last period). It is possibly because of more investments or constructions for them.
  - For **rich areas** like Westminster (Wst), Kensington and Chelsea (Kns) and Camden (Cmd), the net effect is also significantly large for them, tending to be throughout the time or mainly short term. It is possibly because of more visitors are willing to travel the rich areas of a big city.
  - For **remote regions** like Hillingdon (Hdn) and Ealing (Elg), the net effect is negative. They are neither the center of the city/rich areas that visitors will go nor close to the host place Newham. Oppositely, it is possible that more economic activity that was there could move to rich places/host places.

5 Using Random Forest to rank economic features which most significantly contributed to income change due to Olympic effect

With the plot, we see the different trends for different boroughs because of the Olympic games. But that is not the end of the story, because the effect of Olympics games doesn't come from a simple statement that London will hold a global event. There should be something which caused by the Olympics games that further results in the net effects shown above. **We pose the follow up question: what features that contributed the most significant impact towards income change caused by the Olympics?**

In order to answer this question, we use several existing datasets and external data to include possible features that may contribute to the net effects. The datasets and extracted features we use are: (The meanings are in detail in the appendix.)

- earnings-by-borough, since earnings are another way to interpret income, instead of raw data, we extract the features: female-to-male earnings ratio to reflect the possible sexual trend and parttime-to-fulltime earnings ratio to reflect the possible job category that may show the net effect
- economical-activity: the population within the working age and the participating may be the causes for the net effect.
- Data of jobs from London government websites, including number of jobs, job density
- Information of enterprises from London government, including the number of new enterprises, its survival chance after one year, etc.
- Percent of population with different kinds of qualifications/degrees (the meaning for each category listed in the appendix). This can show the job qualifications and industrial distribution.
- Olympic revenue: From Figure 3, we know that extremely most Olympics game events happened in Borough Newham, we then create a column that is 1 for Newham and 0 for otherwise as the factor directly due to the Olympic games event.

5.1 Feature Selection and Engineering

Using the net effects coefficients as the responses, we start with these 29 possible features. Many of them are highly correlated (The pairwise correlation is shown in the appendix Figure 12, and we decided to drop redundant features. We select them based on pairwise correlation matrix in an iterative approach. For the remaining 16 features (The pairwise correlation is shown in Figure 7, we first conducted linear regression for individual indicators to test their corresponding predictabilities for the response variable. Some of them showed significant t-statistics but overall we get poor R squares, which leads us to suggest there may exist some indicators with high non-linear relationship with our response variable, which shows a need to use non-linear model. We decide to use random forest to capture the relative importance for the features.

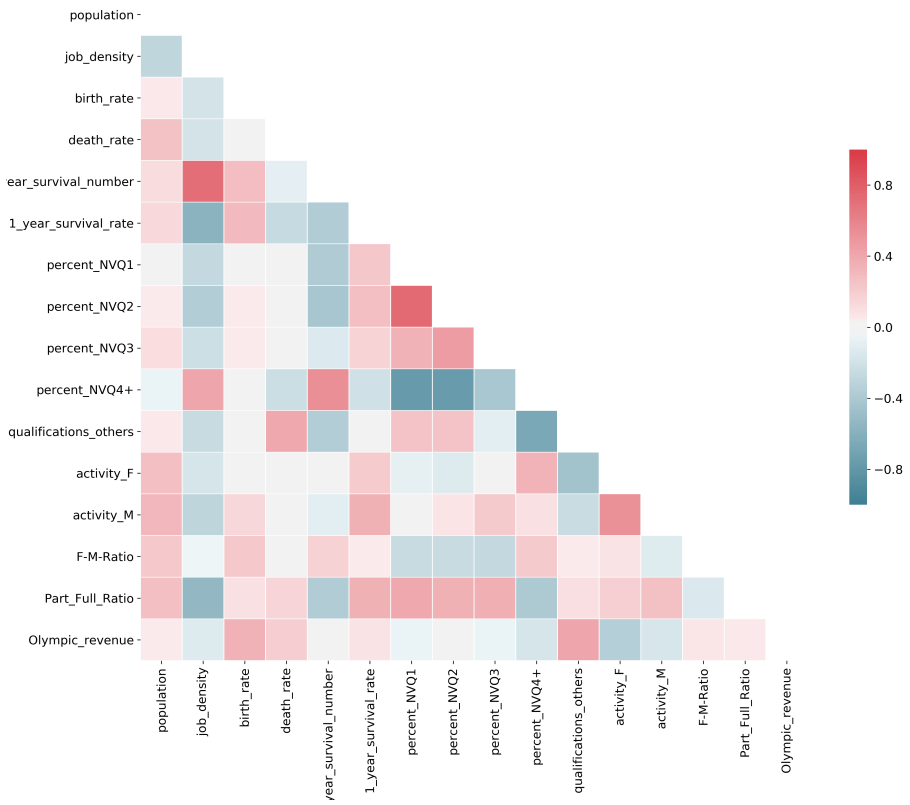


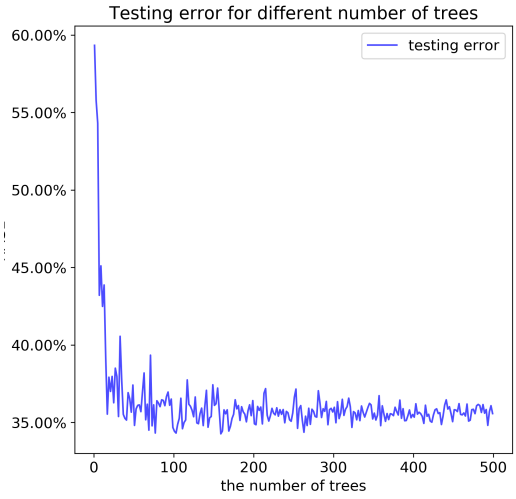
Figure 7: Pairwise correlation between different features after our selection.

The features are listed below: population, job-density, birth-rate,death-rate, 1-year-survival-number, percent-NVQ1, percent-NVQ3, percent-NVQ4+, qualifications-others, activity-F, activity-M,F-M-Ratio, Part-Full-Ratio, Olympic-revenue. The meaning for each feature is listed in the appendix. Please refer to appendix section 12 for full feature descriptions.

We split 1/5 of data as testing data, and use the others to train the model.

5.2 Model Calibration and Results

With the features we select, we go through different hyper-parameters for our model. The RMSE will converge after $n\_estimators$ riches 100, shown in the left panel of Figure 8. After adjusting the max_depth, the best parameters for our model is around $n\_estimator = 60, max\_depth = 4$.



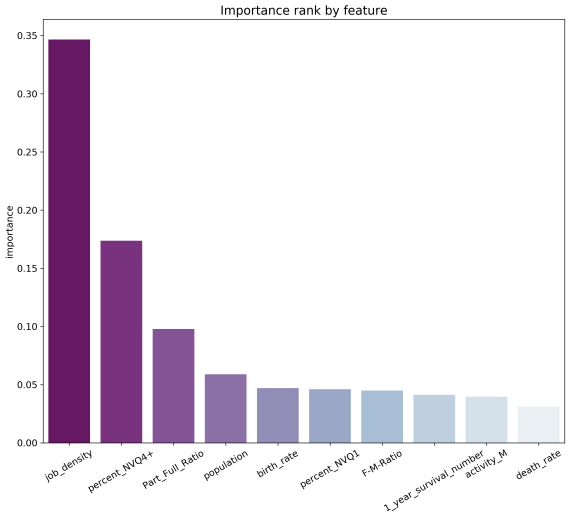


Figure 8: Left: Overall RMSE test error for different number of trees. Right: Feature ranked by importance that contributes the net effect of Olympic games on incomes

From the Figure 8 right panel result, we can see, ranking by the importance, the most four important intermediate features that result in the net effects of Olympics games and the social reason that why they stand out are shown below,

- **job-density**: This number is the number of jobs in that borough divided by the working-age(male and female: 16-64) population. This aligns with our intuition that a higher job density in that area is the leading factor for the net effect.
- **percentages-NVQ4+**(college degree or higher): This indicates that the percentage of higher degree holders contribute significantly towards income growth.
- **Part-Full-ratio**: This feature describes
- **population**: This feature describes the number of people in the working population. Growth in population consequently contributes to income growth.

6 Conclusion

In this project, we address the answers to two key questions:

1. How does the 2012 London Olympics affect the income of people in different boroughs of London?
2. What are the most significant economic features that drive the change in income caused by the Olympics?

Using the DID model, we isolated the impact of the Olympics on boroughs' median incomes. We showed that relatively poor regions such as Newham, Hackney, and Tower Hamlets all experienced the strongest income growth due to the Olympics, characterised as long-term, sustained growth. Rich boroughs also experienced income growth, but more distinct around 2012 with growth tapering off in the long run likely due to a short-term increase in tourism. Using the Random Forest model, we showed that job density and education were the two largest driving factors in explaining income growth due to the Olympics.

With these conclusions, we can propose the following actionable advice for various stakeholders:

- A host city should look to center its Olympic events around a poorer region, as the Olympics can help drive long-term sustained growth and ameliorate income inequality.
- The economic impact of the Olympics not only occurs during the year of its event, but is primarily weighted on the proceeding years, for at least 5 years. Host cities should continue to address increasing tourism, businesses, and infrastructure growth even after the Olympics in order to capture this growth. Metrics to specifically focus on in order to ensure growth is captured are job density and education levels.

- Host cities should be aware of negative externalities that consequently come with rising income in poorer regions. They should be aware of possible issues such as gentrification and look to actively address these issues.

References

[1] Bertrand, Marianne; Duflo, Esther; Mullainathan, Sendhil (2004). "How Much Should We Trust Differences-In-Differences Estimates?"

[2] https://data.london.gov.uk/dataset/

A Appendix

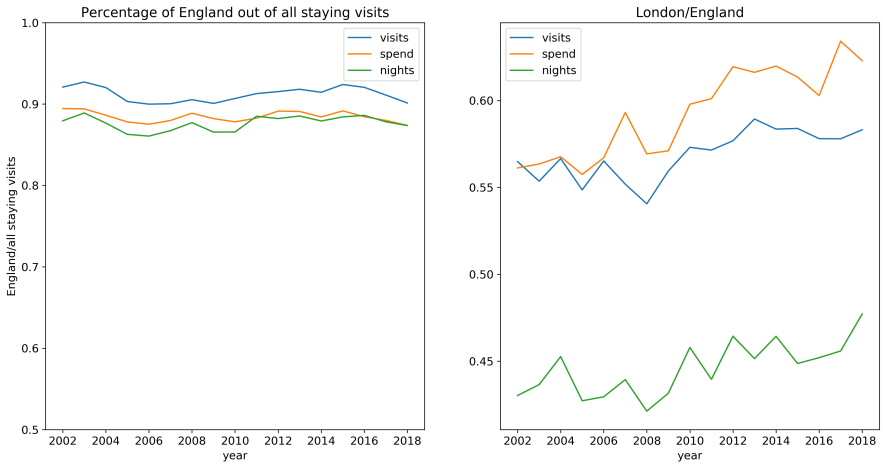


Figure 9: Left: percentage of England visits/spending/nights outside of the total staying visits for international visitors. Right:percentage of London visits/spending/nights outside of the total England for international visitors.

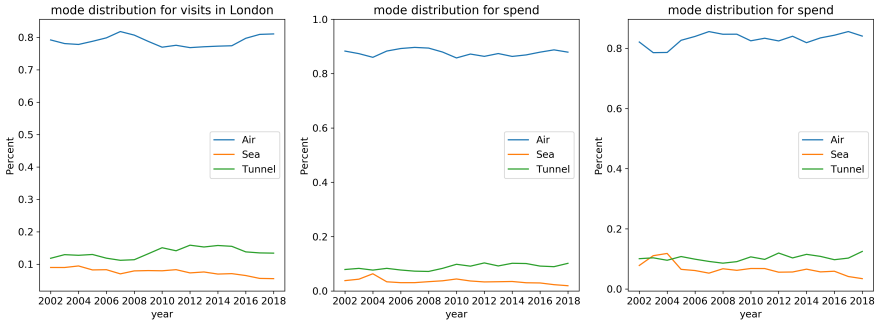


Figure 10: The mode distribution comparison between by air, by sea and driving for the record of visits, the spending, and the nights of international visitor to London during 2002-2018

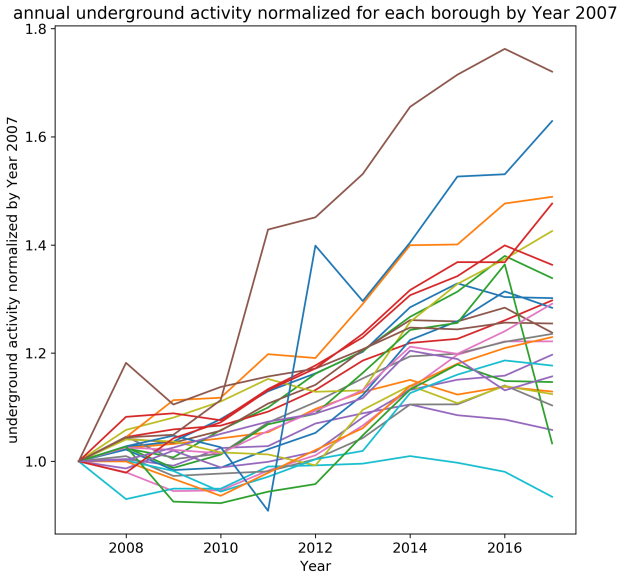


Figure 11: Underground activity (the annual entries and exits records, different stations in the same borough will sum up) for different boroughs during 2007-2017. Different curves are different boroughs, including the topmost one, Newham.

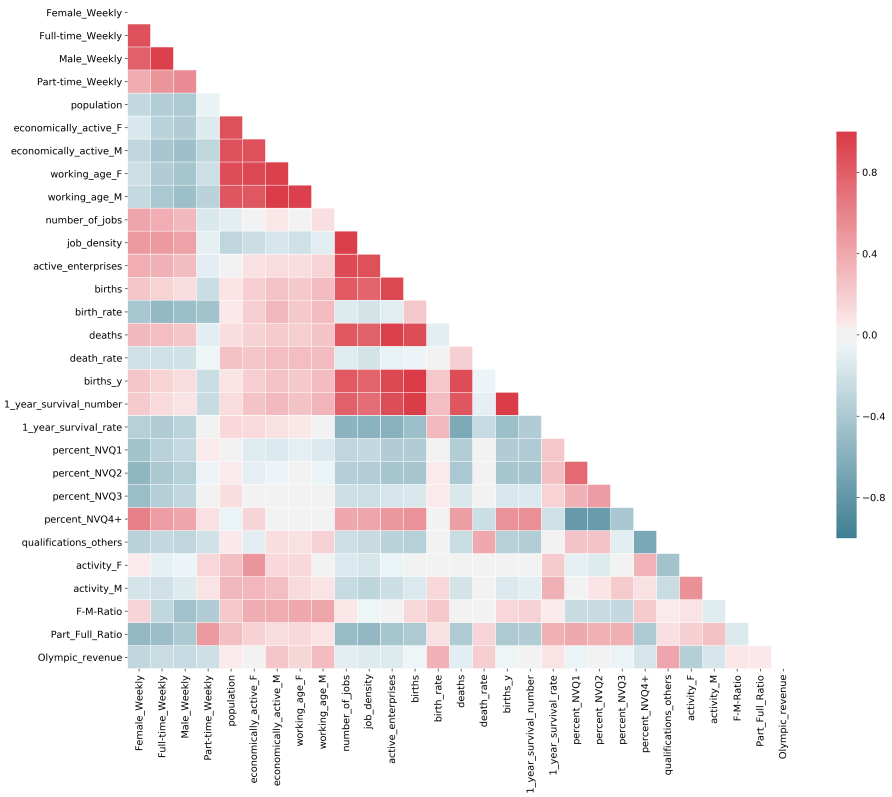


Figure 12: Pairwise correlation between different features before selection.

B About the meaning for each feature:

- job-density: the number of jobs in that borough divided by its working-age(male and female: 16-64) population
- birth-rate: the number of new companies divided by the current number of companies
- death-rate: the number of companies which broke up divided by the current number of companies.
- 1-year-survival-number: the number of enterprises that survive the first year
- population: the taxpayer population
- activity-F: female population with jobs divided by the working-age female population
- activity-M: male population with jobs divided by the working-age female population
- F-M-Ratio: median earnings of female divided by that of male
- Part-Full-Ratio: median earnings of part-time job divided by that of full-time
- percent-NVQ4+: the percent of people with higher degrees.
- percent-NVQ3: the percent of people with higher school certificates

- percent-NVQ1: the percent of people of entry level.
- qualifications-others: the percent of people without NVQ1, NVQ2, NVQ3, NVQ4+
- Olympic-revenue: whether the borough is holding the Olympic games