Daylight saving time impact on web usage

Mocanu Alexandru
alexandru.mocanu@epfl.ch

Cosmin Rusu
rusu.cosmin@epfl.ch

Colicchio Alexandre
alexandre.colicchio@epfl.ch

Abstract

The daylight saving time (DST) has been debated more and more recently, bringing into discussion whether it should be abolished or not. While most of the countries do not observe daylight saving time, most of Europe and North America does, making it an important problem, especially for us Europeans. We will analyze the impact that the DST has related to the web usage. In this regard, we will use different datasets to monitor the posting behavior of users around the time of hour changes. To frame the change of general behavior, the use of several platforms is necessary. We will study the data distributions of StackOverflow, Wikipedia and a gym to derive hypotheses and verify them. We would like to study the impact of this time change on the distribution of positions on different scales, and the time taken to find a ”normal” activity.
We focused our study on the US geographical area, having encountered difficulty to obtain localized data for users.

Datasets

We have decided to use three different datasets for our analysis: Stack Overflow, Wikipedia and a gym dataset from Kaggle. We provide a brief description of each of them.

StackOverflow dataset : The dataset was provided during Applied Data Analysis Homework 3. The dataset consists of questions and answers on StackOverflow that span over the period July 2008-September 2018. Among the various fields related to a post, we only made use of the scores.

Wikipedia dataset : Wikipedia consists of vast amounts of data, being an everlasting source of interesting problems for data scientists to explore. We used the Wikipedia API to collect two different datasets. One consists of wider spaced data, but fewer entries (first 100 changes every 15 minutes in the time interval 18th October-17th November 2018), while the other one contains more data, but over a smaller time interval (all changes from the time interval 29th Obtober-11th November). The time window restriction is due to the limit of querying only the activity over the past 30 days through the API. The datasets are saved as two JSON files, the first one of 53MB and the second one of 595MB.

Gym dataset : The dataset consists of 26,000 people counts (about every 10 minutes) over the last year. In addition, we gathered extra info including weather and semester-specific information that might affect how crowded the gym is. The label is the number of people, which I’d like to predict given some subset of the features. The dataset was taken from a Kaggle challenge.

Research questions

We would like to answer the following questions:

- Does the hour change have an immediate effect on the users’ behavior? Section 1
- Does one of the hour changes (spring time or

winter time) have a more drastic impact than the other? Section 2

- Does the hour change affect only digital behavior or also habits in terms of health? Section 3
- How long do users take to find a more conventional pace? Section 4

1 Immediate Effect

We start by checking if there is an impact of the hour change on the users' habits. For that, we will first analyze the StackOverflow dataset and then check if a similar behavior is observed in the Wikipedia dataset as well.

1.1 StackOverflow

The first step taken was visualizing the posts distributions over the hours of a day for each of the years provided. For summer hour changes, we took into account the years 2009-2018 and for the winter hour changes the years 2008-2017. The figures below show samples of the posts distribution before and after hour change, first for summer, second for winter.

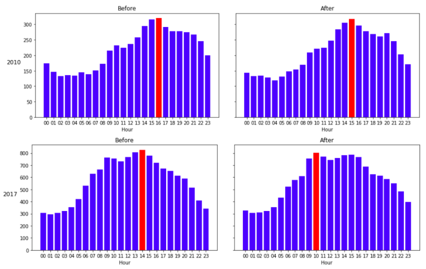

Figure 1: Sample posts distributions

We see that shifts in the distributions do not always occur as we would expect.
We therefore continued with three more numerical analysis techniques:

- Propensity scores: We computed the propensity scores for pairs of before and after hour change data. We used data from before with data from after hour change based on their scores and using these pairs we counted how many hour differences we had for values -12 to 12. For both summer and winter, the 0 hour difference was the most numerous, so we got no pointers for a shift.
- 50% Quantile: Next, we computed the 50% quantile for each of the before and after distributions so that we could find a shift in these values. However, the difference between the quantiles was once again zero in most cases.
- Correlation: In the end, we tried to see if, by shifting the after hour change data appropriately (one hour to the right for summer and one to the left for winter), we could increase the correlation between the before and after hour change distributions. We obtained the following correlation differences:

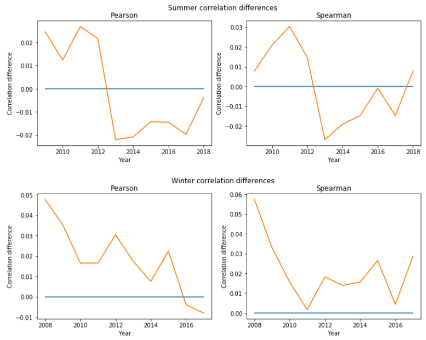

Figure 2: Correlation differences

For the summer hour change, we can see no general improvement, while for the winter hour change, there is generally an increase in the correlation, though it is very small, as the correlation before the shift was already around 0.9.

All in all, the StackOverflow dataset does not show a clear effect of the DST.

1.2 Wikipedia

First, we will look at the distribution of posts in Wikipedia before and after the DST.

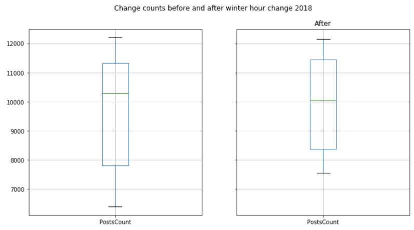

Figure 3: Average posts distribution

We can see that, on average, there are no big differences between the number of changes before and after DST. We will therefore be interested in the changes over each hour of the day. To do this, we take an average over the changes for every hour of the day, the week before and the week after the DST. The figure below shows the distribution of posts before and after hour change on the left and the difference in the number of posts for each hour on the right.

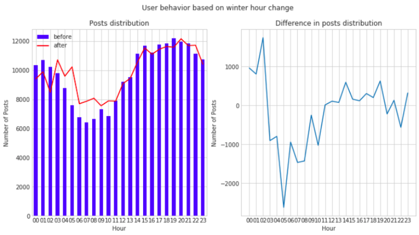

Figure 4: Change in distribution

We can now see that even though the average activity is the same, the habits have changed: the activity in the morning is shifted by more or less one hour and we observe an increase of this activity at the end of the morning. The activity in the afternoon seems to be more or less similar. The plots below show that there is a shift in the distribution for the morning, while in the afternoon the posting behavior remains quite the same.

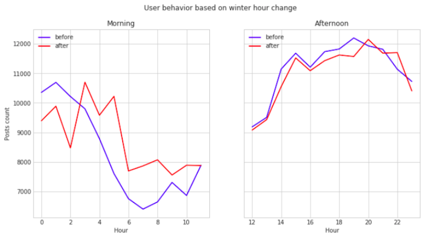

Figure 5: Distribution change for morning and afternoon

After this quick visual study, we focus again on doing some more rigurous, numerical analysis. We will use the correlation to quantify the similarity between before and after DST. For that, we will look at the evolution of these factors by adding an appropriate shift. If the correlation increases after adding this shift, we can say that this indicator is reliable.

As expected, the correlations of the morning are much lower than those of the afternoon. Shifting produces a significant increase in the morning correlations, while for the afternoon it produces a large decrease for the Pearson correlation and a large increase in the Spearman correlation. The correlation for the entire day is almost unaffected by the shift, so overall we can only see a distribution shift for the morning.

Time period	Initial Pearson	Shifted Pearson	Initial Spearman	Shifted Spearman
Morning	0.659	0.738	0.636	0.713
Afternoon	0.944	0.761	0.741	0.902
Overall	0.888	0.896	0.889	0.905

Table 1: Correlations before and after shifting after hour change data

1.3 Overall observation

After analysing both StackOverflow and Wikipedia datasets, we have concluded that the DST does not have a clear impact on the overall posts distribution, but in case of Wikipedia it showed a change in the distribution for the morning. Our first intuition that we may see an evident shift in the distribution may have been undermined by the fact that there may not be such a massive number of users from the United States, as we have expected.

2 Spring or Winter ?

For determining if there is a difference between the impacts that spring and winter hour changes have on the posts, we use the StackOverflow dataset. We saw that there is no real impact on the distribution for either spring or winter, so we used the scores for this analysis instead. We computed the total scores for two days before and two days after the hour change for each year. We then take the difference between the after and before hour change scores and obtain the following figures:

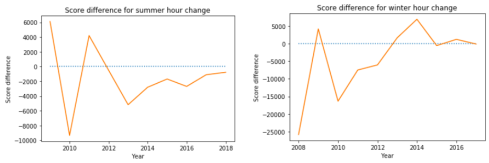

Figure 6: Score differences

The score differences do not show any clear sign of being impacted by the hour change either, as they fluctuate for different years. However, the general trend is for the scores to be higher before the hour change rather than after. The mean score difference for the spring is -1376, with a variance of 4123, while for winter there is a mean of -4235, with a standard deviation of 9576. Theerefore, the winter hour change tends to have a greater impact on the scores both in terms of mean and standard deviation of the difference between after and before hour change scores.

3 Health habits

While going to EPFL and UNIL's sport center, one of us noticed that right after the daylight saving time from November this year, more people were getting to the gym early. Our team member went to the gym at 7am exactly when the program starts, in order to avoid waiting to use some equipment. Before DST, there were barely one or two people before 8am, and right after, a lot more came (that number almost tripled).

This gave us the idea to also research how people's health habits are affected by the DST change. Indeed, the well being of a person is very likely to influence how one expresses themselves on the Internet. We quickly came across a dataset on Kaggle (**?**) that contains information about the number of people that were in the UC Berkeley's gym, every 10 minutes starting with 14th August 2015 all the way through 18th March 2017. This data is collected from a single timezone, which makes our analysis simpler, but it suffers from biasness from the fact that the data is being collected from a prestigious university's campus gym, and so probably the vast majority of people going to that gym are students. We kept this in mind while analyzing our data.

3.1 Dataset Description

The dataset consists of 26000 people counts (about every 10$minutes$) over the last year. In addition, it includes extra info including weather and semester-specific information that might affect how crowded it is. Concretely, each row contains the following information:

- $date$ (string; datetime of data)
- $timestamp$ (int; number of seconds since beginning of day)
- day_of_week (int; 0 [monday] - 6 [sunday])
- $is\_weekend$ (int; 0 or 1) [boolean, if 1, it's either saturday or sunday, otherwise 0]
- $is\_holiday$ (int; 0 or 1) [boolean, if 1 it's a federal holiday, 0 otherwise]
- $temperature$ (float; degrees fahrenheit)
- $is\_start\_of\_semester$ (int; 0 or 1) [boolean, if 1 it's the beginning of a school semester, 0 otherwise]
- $month$ (int; 1 [jan] - 12 [dec])
- $hour$ (int; 0 - 23)

Daylight Saving Time changes do not necessarily occur on the same date every year. The changes contained in this dataset are shown in Table 2.

Date & Time	Time Change	Abbreviation	Offset After
2015, 1 Nov, 02:00	-1 hour (DST end)	PST ⇒ PDT	UTC-8h
2016, 13 Mar, 02:00	+1 hour (DST start)	PST ⇒ PDT	UTC-7h
2016, 6 Nov, 02:00	-1 hour (DST end)	PST ⇒ PDT	UTC-8h
2017, 12 Mar, 02:00	+1 hour (DST start)	PST ⇒ PDT	UTC-7h

Table 2: DST changes in Gym Crowdness dataset

3.2 Analysis

We started plotting the aggregated sum one month before and after the change, grouped by day of the week. The results can be seen in the figure below.

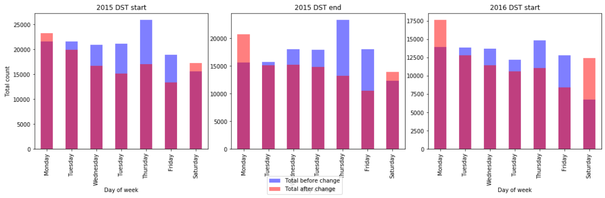

Figure 7: Distribution change

We can see the results are very consistent. The distribution skews to the left, perhaps because people are motivated to come to the gym as early as possible (the peak is switched from Thursday to Monday). We can think of this as the *New Year's Resolution* effect, when people are very motivated to start going to the gym and eating healthy right after the New Year's Eve, but they lose momentum in the following months, perhaps returning to normal, as soon as the vibe is gone.

We further plotted heat maps **??** of aggregate sum from each day of the week and each hour. It can be seen that the number of people coming early in the morning (6am) increases after the DST, which makes sense and we were expecting that, since people can sleep the same amount of time, but actually show up one hour earlier at the gym as a result of DST saving time. However, that should not be the case when one sleeps less than usual.

A very interesting pattern that we noticed is that on every change, the highest traffic at the gym before the change was always on Thursdays (white squares) afternoon, between 14-20, and this always shifts to Mondays on the same hour interval.

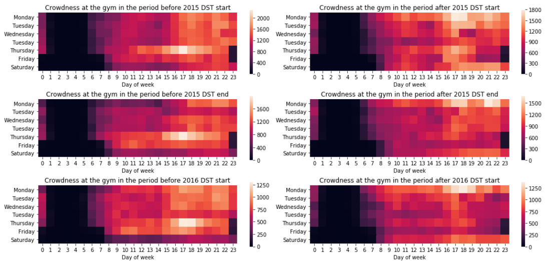

Figure 8: Crowdedness heatmap

In Figure 9 we see the correlation between the number of people and the other features that we have. As seen, these change with each DST change, but we have to be careful to understand how the bias of the data affects this. For example, after DST, the $is\_start\_of\_semester$ variable is very correlated with the number of people at the gym.

3.3 Key takeaways

After our careful analysis, we can conclude that DST is beneficial for the people that are practicing weightlifting because it acts as an impulse to go to the gym earlier. In such cases, the hourly change acts as a refresher, generating momentum.

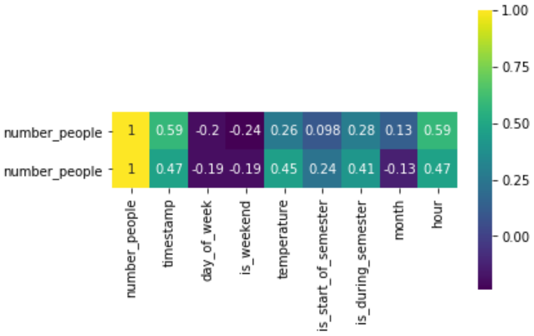

Figure 9: Correlation before and after hour change

4 Find these habits

In this part, we will try to answer the question ”How long do users take to find a more conventional pace?”. For this we will compare the results of each day of the following week after the DST for the Wikipedia dataset with the average distribution for days of the week before the DST, which should correspond to ”normal behavior” of the users. We will exclude from our analysis Saturday and Sunday, as they are non-working days, the behavior of the users being too different from a normal day of the week.

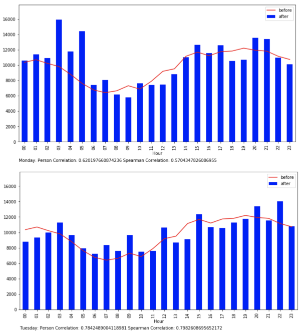

Figure 10: Monday and Tuesday distributions

As seen above, the correlation increases between Monday and Tuesday. So we will check for the rest of the week if we can say that most of the shift operates on Monday and there is an improvement thereafter.

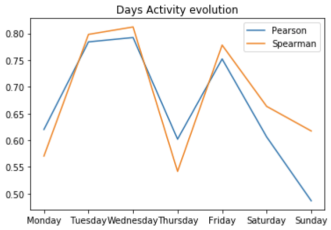

Figure 11: Daily correlation

As explained above, if we exclude Saturday and Sunday, we can observe that the correlation seems to reach a high level from Tuesday. Unfortunately, there is a sharp fall on Thursday, which does not seem to have any causality with the DST. We should extend this analysis over the same periods over several different years to validate our hypothesis, and especially to affirm that there is no causality with the fall of Thursday.

5 Conclusion

All in all, if considering data from only one timezone (Berkeley gym dataset), we see that there is an impact of the DST on the people’s behavior. StackOverflow and Wikipedia datasets did not lead to the same conclusion, showing only some visual cues that something might have happened, while there was no significant numerical evidence towards this assumption. This leads us to the conclusion that such an analysis should be done on datasets regarding only one timezone, as a mixture of timezones tends to balance the effects that the DST may have on some of them.