



Changes of behavior towards physical stores in Italy during 2020

Jacobo L. Gil

IronHack

Data Analysis OCT2020

Final project delivered 2020-12-18

Overview

This paper tries to answer the question: “will Italians buy in shops after the COVID-19 pandemic has ended?” through the comparative analysis of patterns of behavior shown in data extracted from open sources.

The sources

The main dataset used for this project was elaborated by Google for the Community Mobility Reports project and can be accessed through their website¹.

As it can be read in the documentation site, these reports show the percentage of variation in the number of visits that certain categories of places receive compared to a baseline. The baseline is formed using the median values of visits for each weekday between 2020-01-03 and 2020-02-06. This means that the percentage of change refers to a specific day of the week (Monday, Tuesday, etc.) and it is not usable for another weekday.

The categories in which the dataset is divided are:

Grocery & pharmacy: mobility trends for places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies.

Parks: mobility trends for places like local parks, national parks, public beaches, marinas, dog parks, plazas, and public gardens.

Transit stations: mobility trends for places like public transport hubs such as subway, bus, and train stations.

Retail & recreation: mobility trends for places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters.

Residential: mobility trends for places of residence.

Workplaces: mobility trends for places of work.²


It is worth noting that the variable ‘residential’, unlike the others, does not measure the number of visitors but the changes in the time spent at home.

Another source that had to be included was Weatherbit³. The API of this site provided the necessary data to contextualize the data series that became the main explanatory variable: the ‘parks’ variable. As it will be shown later, this variable was used to determine the seasonality of the changes in behavior through the patterns of the number of visits to

¹ <https://www.google.com/covid19/mobility/> last accessed: 2020-12-13.

² https://www.google.com/covid19/mobility/data_documentation.html?hl=EN last accessed : 2020-12-20.

³ <https://github.com/weatherbit/weatherbit-python> last accessed : 2020-12-20.



open-air places of leisure. In this direction, a minimal climatological background was needed to make sense of the data. The way this background information was organized during the analysis was by collecting the average temperature from three major Italian cities (Naples, Rome and Milano) for every one of the days that were scrutinized. After that, the average of those three daily temperatures was extracted.

In this manner, we obtained a superficial, but global vision of the weather conditions in the country, since those three cities are big population centres that represent the three main regions of Italy: South (Naples), Center (Rome) and North (Milan). As it has been noted, this analysis is not deep. There are two main reasons for this: first, the limitations that the free Weatherbit account sets prevent the users from running more than 1000 queries a day. This way, including more variables or cities would have limited the number of tests that could have been performed.

Second, because 'seasonality' is a behavioural concept, not a climatological one, thus this kind of data was brought in to explain outliers and to show possible points of disruption, not as an explanatory variable by itself. In this regard, although more depth into the subject of the weather would have been desirable in theory, it was not essential and could have complicated the analysis beyond its functional scope.

Finally, the third open source used for this analysis was Wikipedia⁴. Inasmuch as Wikipedia can be problematic if not handled cautiously, its inclusion was justified in order to have a basic outline of the most remarkable events of the period of interest. Nonetheless, even for this purpose Wikipedia proved to be tricky: first of all, because dealing with unstructured data (as text) is always difficult. This was magnified by the fact that the redaction of the site was neither comprehensive (some important dates were missing) nor well organized.

Second, because of the complexity of the subject itself: on one hand, Italy is a very administratively fragmented country and very different policies took place at the same time in different regions. On the other hand, the pandemic is a complex matter and it is too recent to have a clear picture of what has happened. All this defined the direction that the analysis took, which would be operating at federal level - instead of regional; so that local particularities would not prevent from seeing the broader, general trend - and elaborating the timeline based on behavioural patterns and then projecting the historical milestones on it.

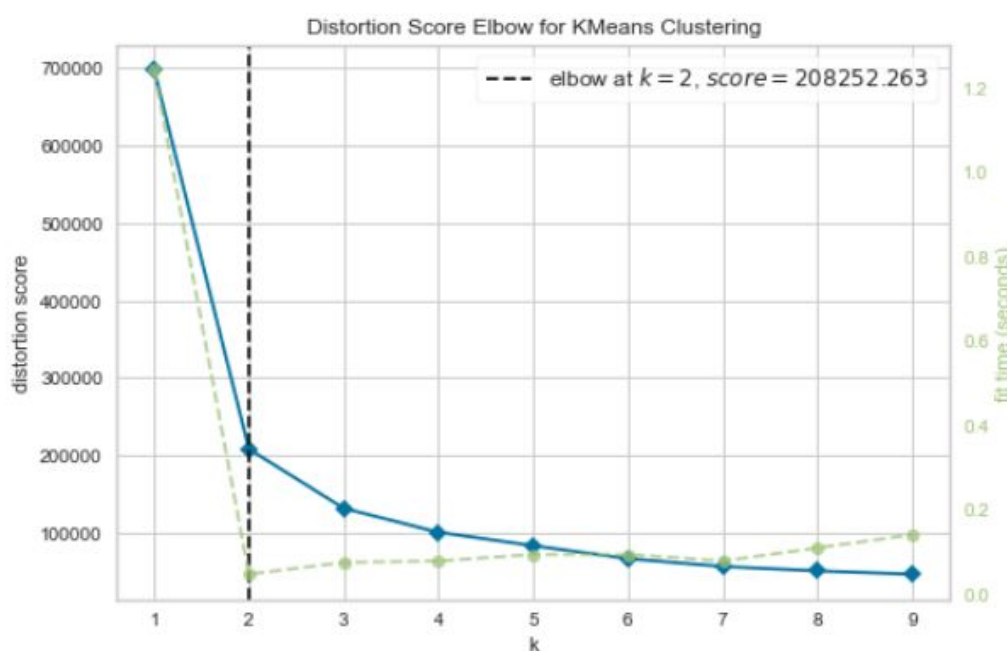
⁴ https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Italy last accessed: 2020-12-18.

Visualization of correlations and clustering

The first relevant step into the analysis was getting a general timeline with the main events since the beginning of the pandemic. Given that Wikipedia had proven - initially - not to be of use, it was decided to divide the timeline into the periods suggested by the k-means clustering algorithm⁵ using the data provided by Google Mobility's dataset.

In this direction, since the variables feeded to the algorithm measure the changes in patterns of behaviour, the clusters - every one of the divisions of our timeline - would be based on visible alterations of behaviour of the Italian population and not necessarily on institutional impositions - although these had a tangible impact on the patterns.

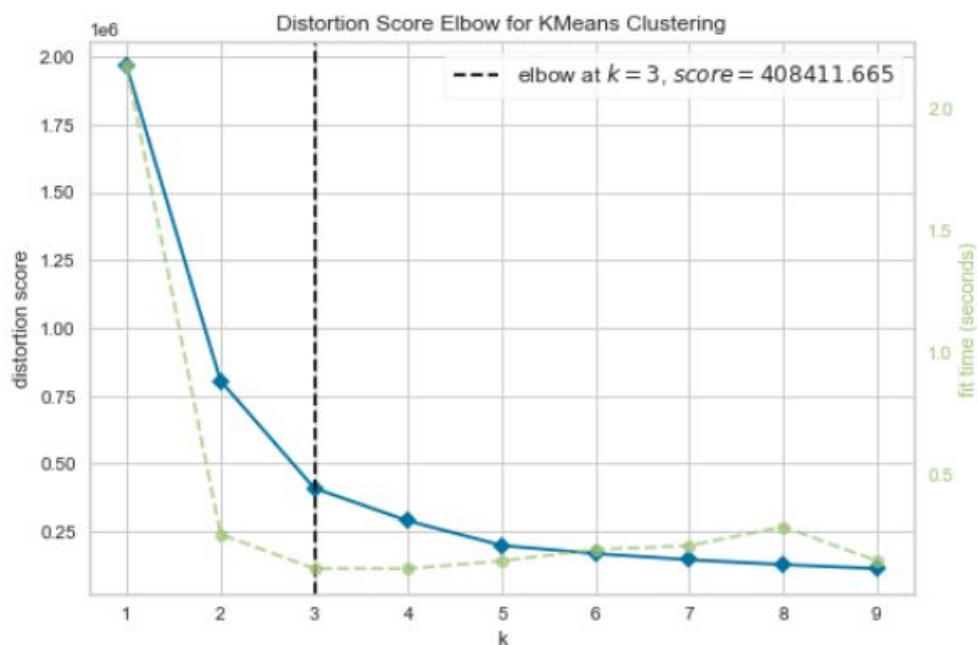
Before performing the clustering, several k-elbow visualizations were run to test what combinations of parameters would be more revealing. Throughout these tests, only two main patterns appeared. The first one, without including the variable 'parks', showed only two meaningful clusters:



What this means is that the days included in the dataset presented characteristics that would allow to classify them in only two relevant groups. In other words, without including the variable 'parks', the algorithm does not see any relevant differences beyond days during a lockdown and days without restrictions.

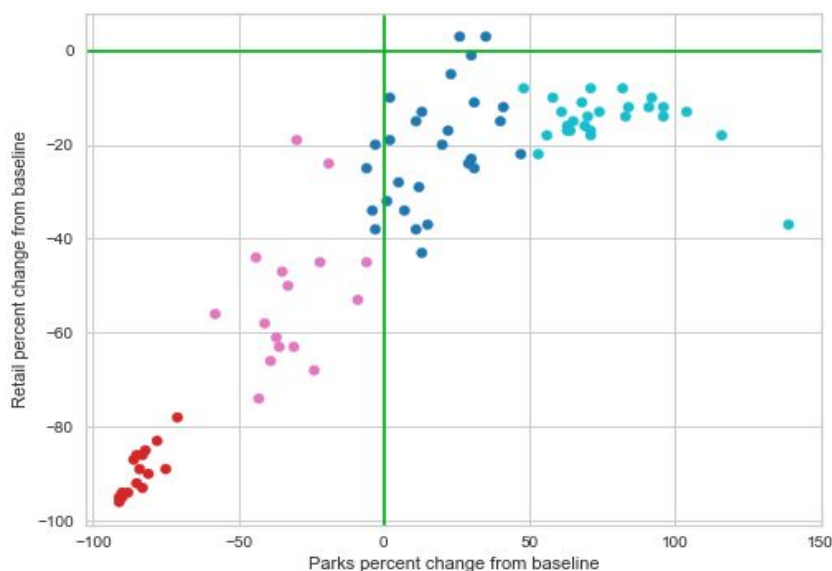
The situation changes when we include the 'parks' variable:

⁵ <https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html> last accessed: 2020-12-21

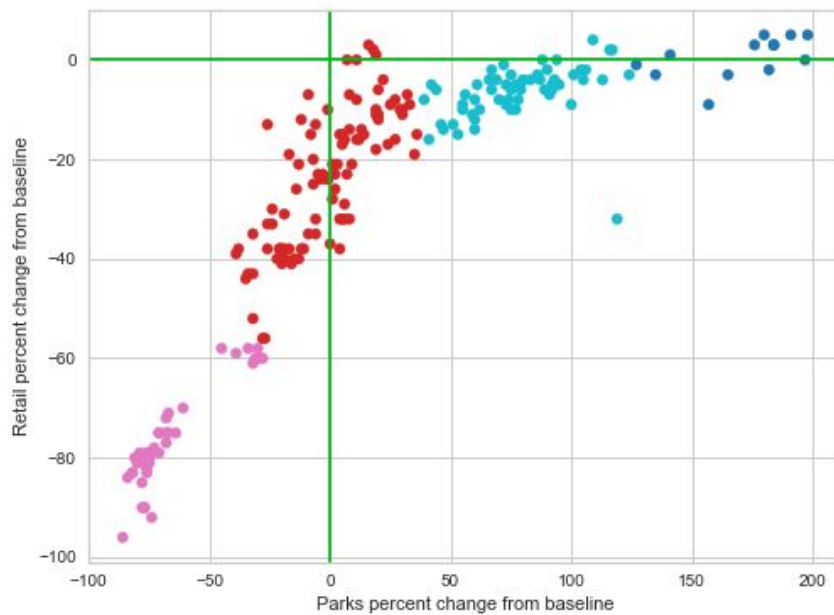


As we can see here, the complexity of the timeline increases. The algorithm suggests a division of the days in three categories that - we can anticipate - are: days in lockdown, days without lockdown in low season and days without lockdown in high season.

The next step is seeing how the 'retail & recreation' variable - the target variable - and the variable 'parks' - the variable that presents the most defining seasonal patterns - relate to each other. To have a more accurate perspective, the data were divided in two graphics, weekends:

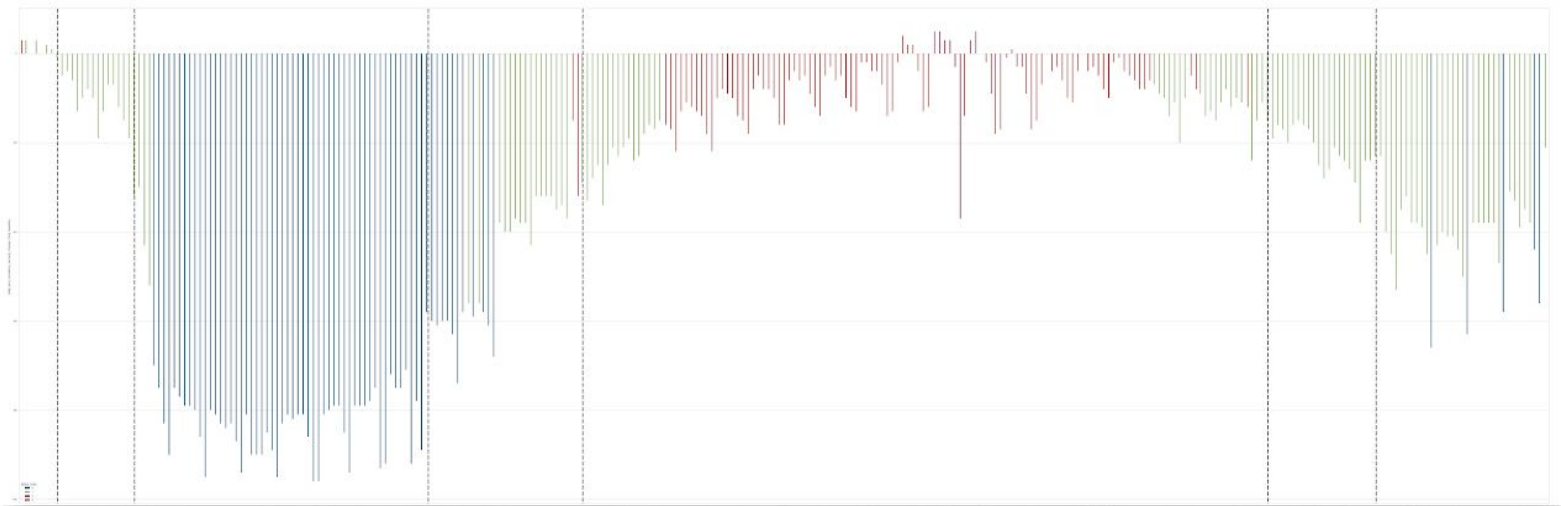



And weekdays:



These two graphics show that - with an adjusted scale - there is a correlation between both variables: in general terms, when the numbers increase in the visits to 'parks', they also do for 'retail & recreation'. It is worth noting that the green lines mark a 0% variation for each variable, so that it becomes evident that, while 'parks' hits the 200% increase over the baseline, 'retail and recreation' struggles to reach the 0% and only surpasses it anecdotally.

Now a timeline can be established:





In this graphic the bars show the percentage of variation in 'retail & recreation' (mostly negative, with some exceptions in February and during the summer) and the colors show the category in which the k-means algorithm has classified every day taking into account all 6 variables of Google Mobility. In this regard, more intense cold colors show compound decrease of the variables, while more intense warm colors show an increase of the variables as a whole above the baseline.

This way, the dark blue bars show that in the months of lockdown (March and April, essentially) the decrease of activity was general. Contrarily, in August the dark red bars show that, in spite of 'retail & recreation' remaining under the baseline most of the time, the compound number of visits across all variables was on the rise, with the variable 'parks' being the most influential in this change of classification.

Projected over the graphic are several discontinuous lines that mark some important events that were extracted from Wikipedia. From left to right:

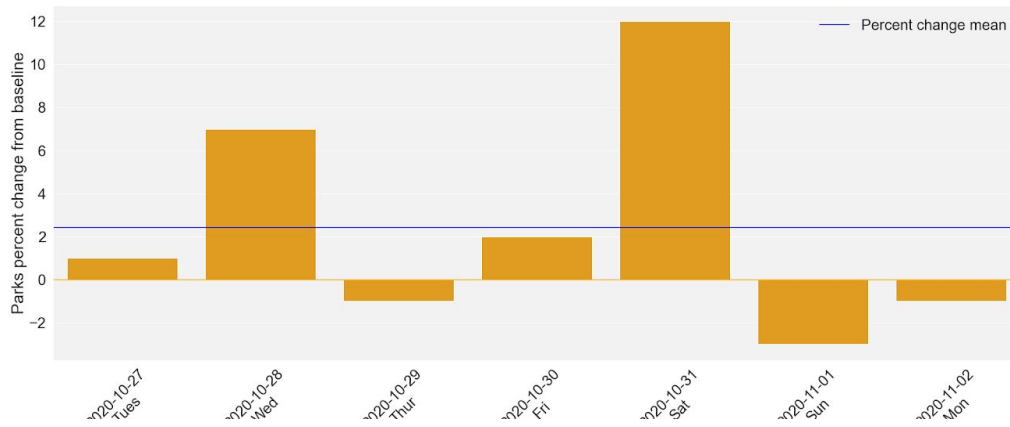
- 22nd. of February: first cities under lockdown in northern Italy.
- 8th. of March: lockdown is imposed over whole regions in northern Italy. 2 days later it will be extended to the rest of the country.
- 4th. of May: some restrictions are lifted. Progressively more degrees of freedom will be granted during the following weeks.
- 3rd. of June: total freedom of movement is reestablished.
- 14th. of October: cases of COVID-19 positives exceeded the peak of the March infections.
- 4th. of November: a second lockdown is declared, dividing the country in regions with different restrictions depending on the severity of their outbreak.

Comparative analysis

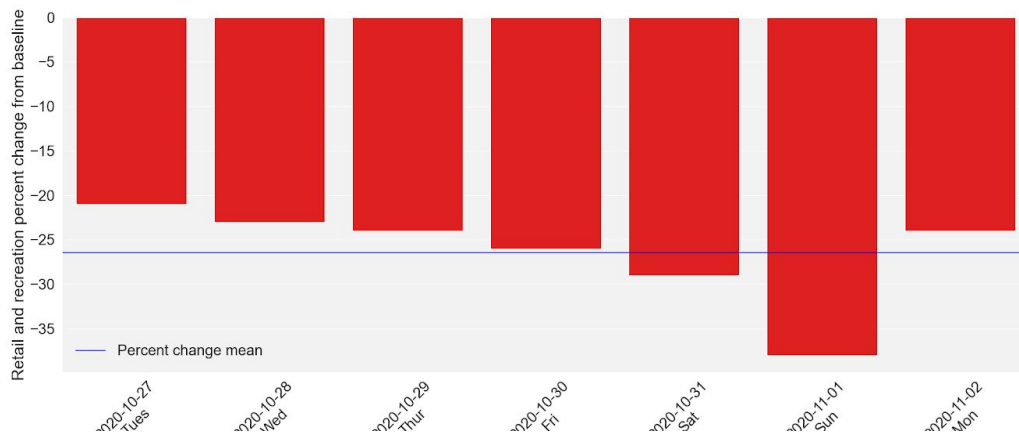
Intuitively, to draw any conclusions around the changes of behaviour of the Italian population towards physical stores it is necessary to find a period which circumstances are comparable to those prior to the start of the pandemic and see how the variable 'retail & store' has evolved.

The first place to look would be the week (to have at least a whole cycle that includes, consecutively, both weekdays and the weekend) that presents the minimal variation possible of the 'parks' variable. The week with that characteristic is that between the 27th. of November and the 2nd. of December.

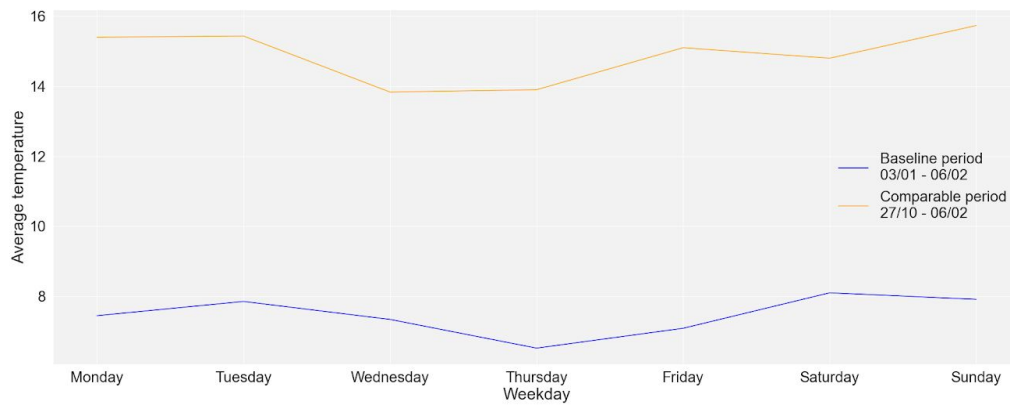
This, week, in spite of its outliers, presents a 2.5% average variation over the baseline:



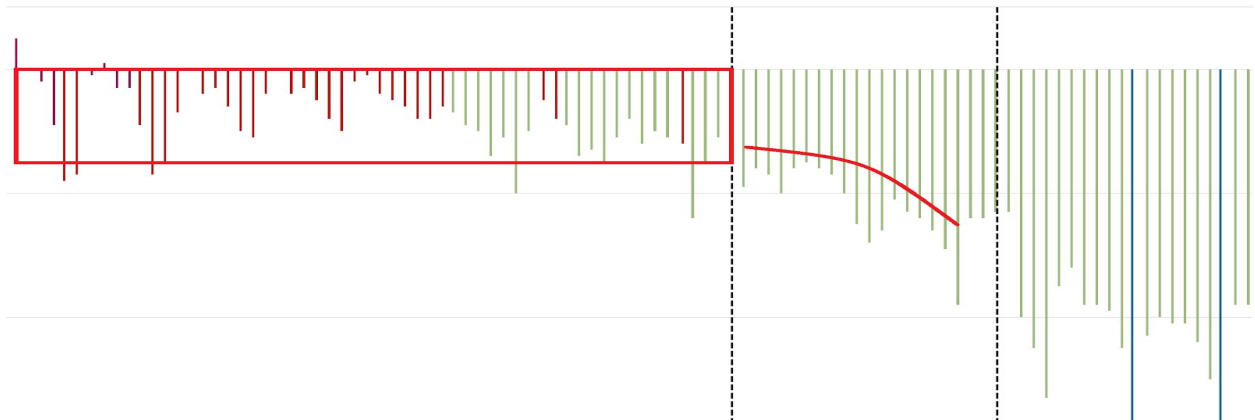
The results concerning the variable 'parks & recreation' are, however, quite different with an average variation of -25% :



Two fact are worth taking into account: first, that temperatures were, in general, higher during this week than during the baseline period:



Second, that this week occurs after the 14th. of October. This is significant, because days with outstanding numbers of new cases tend to be publicized, thus provoking changes of behaviour. The change becomes apparent in the ascending curve between the 14th. of October and the establishment of the second lockdown in November:



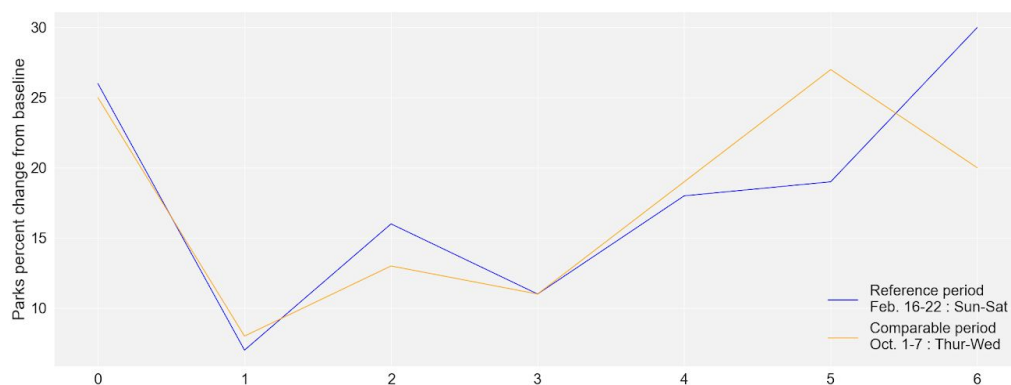
This partially invalidates this period as a reference. There is, however, an alternative: the week between the 16th. and the 22nd. of February are registered in the dataset and contain information of a period prior to the declaration of any lockdown, and, thus, the behaviour registered during this week could be considered 'normal'.

Indeed, it has been clustered as non-lockdown period in low season - as it should be January, after the high season of Christmas - and presents 'retail & recreation' values above 0% - which are, otherwise only found during the high season of this year's summer, but should be expectable under normal circumstances:



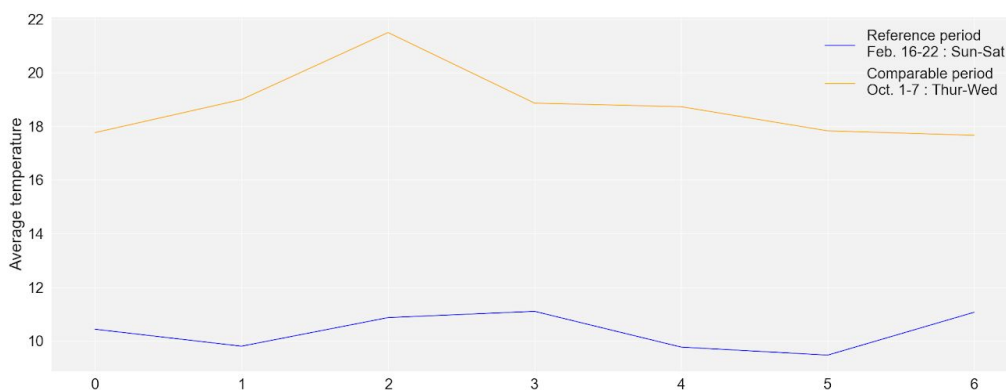
This way, if a week with a similar behaviour with regard to the variable 'parks' could be found in the inter-lockdown period, it could be compared how the variable 'retail & recreation' evolved in each case.

Fortunately, there is a certain alignment between this week and that between the 1st. And the 7th. of October:



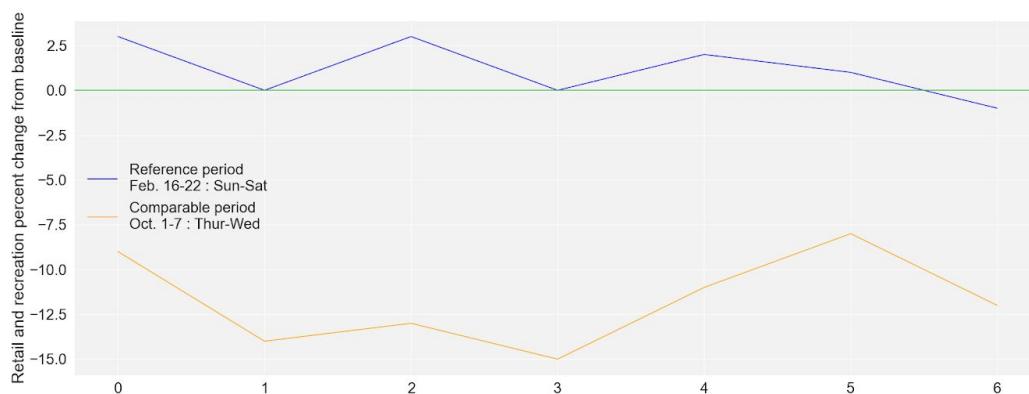
There is, however, a complication: the pattern is similar, however, the days do not correspond between both weeks: while the reference week runs from Sunday to Saturday, the week of inquiry runs from Thursday to Wednesday.

To this it has to be added the sheer difference in temperatures and in their evolution:



Still, there are reasons to use this comparison as a reference: the first one is very prosaic, it being the absence of a better alternative. The second main reason is that, regardless of the disposition, we have established a linear correlation between the variables 'parks' and 'retail & recreation', thus, it should be expected that even unaligned common patterns in the first variable should show reciprocity in the other one if the circumstances were similar. Finally, this comparison is useful when not used on its own, but in combination with the previous one.

The comparison between the evolution of the variable 'retail & recreation' can be seen in the following graphic:



As it can be noted, both lines evolve somewhat parallelly (with the one of October running sustainably below the baseline) in patterns relatable to those in the variable 'parks'.

Recap and conclusions

In the last graphic it can be appreciated how comparable patterns in the 'parks' variable meet a parallel response in the 'retail & recreation' variable lying, however, around a 12% below in the week of inquiry. Nonetheless, it would not be cautious dragging any conclusions from this alone, for two main reasons: it is never a good idea relying on one single observation and, in this particular case, for all the reasons exposed above, specially, because of the lack of alignment between the days of the week of the two compared periods.

What can be done is comparing the results of the two inquiries, departing from the fact that both present behaviours in the 'parks' variable that are reasonably similar to those of the baseline - or, arguably, to the more general idea of 'normalcy'.

On their side, the results in the 'retail & recreation' variable can not be more diverging: whereas the first week of October reacts proportionally to the variable 'parks' and stays around the 10% under the baseline; in the last week of the same month both variables dissociate, with the average value of -25% for the variable 'parks & recreation'. As it has been shown, one fact that explains this divergence is the peak in COVID-19 cases in the middle of the month.

What can be concluded is that the Italian population tends to go back to their pre-pandemic habits when they perceive that the risk is lower. In this direction - following the trend shown by the current available data - the Italians should display a preference to buy in physical stores comparable to that previous to the outbreak of COVID-19 once this health-crisis ends.

However, two matters have to be taken into account: first, this is a tendency. What this means is that the data show serious alterations in behaviour in absolute terms. Of our two weeks of inquiry, the most positive ran around a 12% below the baseline and the total reasons for that cannot be accounted for: probably, it is in part explained because part of the population that decided to still act cautiously. In part it might also be because of the worsening of the Italian economy. Also the seasonality might have altered the parameters of comparison, given that the temperatures in October were higher than those of February, influencing a disparity in behaviour, encouraging people to go to the parks, instead of going to the stores and restaurants.

Summarizing: it has been argued that it is reasonable to think that the Italians will be willing to go back to their old habits, but the extent to which this will happen goes beyond what can be analyzed with the data used for this report.

Finally, this report analyzes the changes of behaviour that have taken place so far. Subsequent lockdowns and confinements can further alter those patterns in the future.