Geospatial Analysis and Representation for Data Science 2021

Strava Analytics - Analysis, investigation and mapping of Strava activities

Andrea Debeni

University of Trento

February 18, 2022

1 Introduction

This project aims to implement various techniques seen throughout the course and apply them to a specific kind of data: Strava activities. The general purpose of this approach is to discover interesting results, isolate patterns or explore meaningful visualizations that can help to find useful insights from these activities, and maybe draw a story from them.

At the moment, Strava provides an heatmap of the registered activities, even if only for premium users. Other available analysis are related to the single activity (pace, elevation, etc.), but they are limited in multiple ways: comparisons between activities are difficult, segment-based and usually not free. Moreover, analysis are often presented in a static graph and are not displayable over any kind of maps.

With this project I tried to address these *imperfections*, trying to enrich data from Strava activites with specific analysis and including data from other sources, like OpenStreetMap and Garmin.

The project is divided in four parts:

- An analysis focused on a single long ride
- An ensemble of mapping techniques to identify activity clusters and common routes
- An example of integration of OpenStreetMap data and a Strava hike activity
- A temporal analysis of the activities' performance featuring different parameters and sources

The code for the project is freely available online[1], as well as all the data used in it. The repository comes with a dual solution: a single notebook where all the analysis are performed one after another; four separate notebooks executable stand-alone, one per analysis type. This choice has been made to facilitate testing and integration of potential future improvements.

Given that Strava has a legitimately strict regulation regarding user activities' data, I decided to provide mine activities to carry out the analysis. By doing this, privacy issues of any kind are avoided, while a ground truth (i.e. whether sound results are correct or not) is always available.

[1] https://github.com/Debo790/Geospatial2021

2 Dataset

All data comes in a .gpx format, an XML schema containing data related to position, time and elevation. The dataset features 152 labelled activities and 7 unlabelled (i.e. they are not runs, hikes or rides): these latter are discarded for the subsequent analysis.

Specifically, the dataset is composed by 135 runs, 3 rides and 14 hikes, divided in 3 different regions (this aspect will be further discussed later). In every notebook, activities tracks are first parsed with `gpxpy` [2] and then stored in lists, dataframes, geodataframes and trajectories, since each of these structure which will be used in different phases of the analysis.

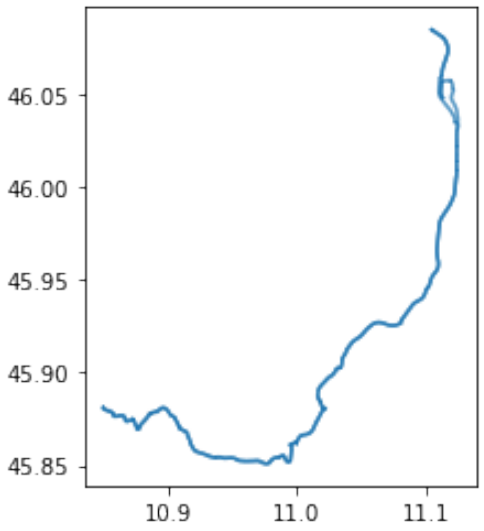

Figure 1: Longest ride trajectory

3 Analysis

In order to make these analysis a little bit more interesting, I decided to focus on peculiar characteristics for each part of it. Every one of them features a small story: this, I believe, add values to every analysis, making them more personal in some way.

3.1 Long Ride Focus

The longest ride of the dataset represents a friends' trip to Garda Lake, more than 100km and 10 hours (not continuous) of riding. Starting and arrival locations were very near, so without knowing the path someone could figure it's been a ring tour. Luckily, Movingpandas Trajectory[3] allow us to easily display the real path, as shown in Figure 1.

Trajectory provides the ability to add the current speed to every point of the trajectory itself. This comes very handy, since it gives the chance to plot the trajectory in a map according to the pointwise speed, thanks to the `hvplot()` function. The result looks something like the one shown in Figure 4.

This function generates an HoloView plot, which is a zoomable plot with selectable tiles: in the example I used OSM tiles, which make easy to spot whether there are significant variations in the color line. As a matter of fact, in a pretty long segment from Nago to Torbole the line appeared dark red in a direction and almost white in the other one: the trajectory dataframe shown clearly that the segment was the only part of the route in which I passed 40 km/h. The reason is easily explained: computing the elevation gap between the first and the last point in the segment, I found that in less than a km there was an altitude difference of 85 metres. This explains the high speed in the descent, but also the (ouch!) extremely low one in the ascent on our way home.

[2] https://github.com/tkrajina/gpxpy
[3] https://movingpandas.readthedocs.io/en/master/trajectory.html

3.2 Mapping (almost) like Strava

HoloViews plot are useful even to map multiple activities at once, like in Figure 5. This solutions gives pleasant results when activities are not overlapped, but this is not always the case: for example, it's legitimate to think that runs activities start mostly nearby runner's home. But if the runner registers activities when he's away from his usual residence (if he's a non-resident student, to give a *random* example), how this will affect the plot? An example is represented by the scatter plot in Figure 6, where the starting points for all activities are plotted in a single map.

Trying to find activity clusters with this data is probably pointless, since there will be a giant cluster and various singleton. But reducing the area of interest to the one where the majority of activities lie (Trento, in this case) can produce improvements in identifying clusters: it can even lead to infer personal information, like the area where the user actually lives.

This analysis has been carried out by reducing the area of interest (thanks to Overpass API and `osm2geojson` [4]), Kernel Density Estimation (Figure 7) and finally through a DBscan clustering (results in Figure 2). This result shows how there are three main areas of starting activities, all of them in the south-west area of the city: two of them are on Lungadige, the other one seems to be near San Pio X. And given that I'm not living on the river banks...ok, let's move on.

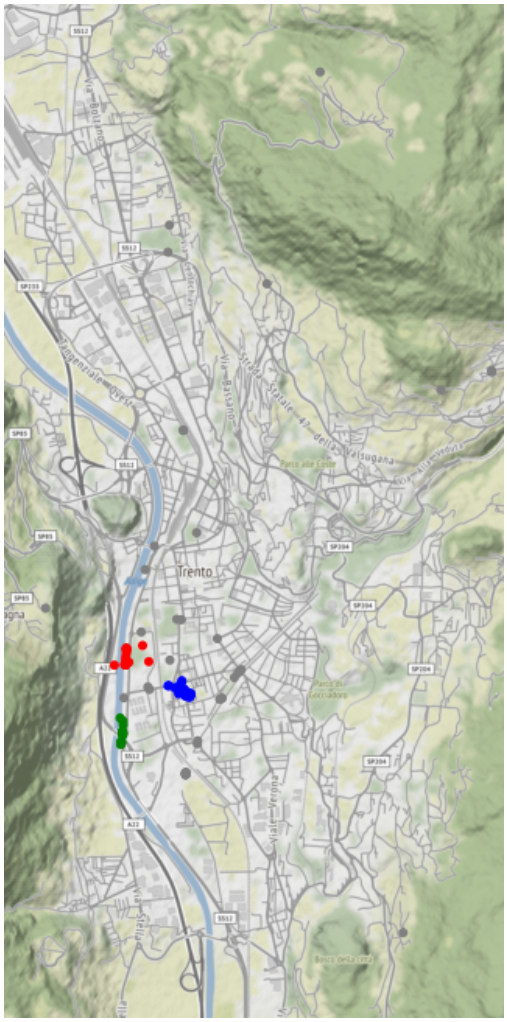

Figure 2: Clusters of activities in Trento

Another important aspect to assess is the common routes: how many times did I take the same path across all the activities made? This is surprisingly easy to achieve, thanks to Folium heatmaps[5]: starting from a set of coordinates, Folium automatically calculates the occurrences for any given coordinate and plot them accordingly to their frequency. By imposing the right set of color gradients[6] on a proper tiles set (e.g. CartoDB dark_matter), it's possible to have a result similar to the one officially provided by Strava (Figure 3).

[4] https://github.com/aspectumapp/osm2geojson
[5] https://python-visualization.github.io/folium/plugins.html
[6] Interesting one here: https://github.com/remisalmon/Strava-local-heatmap-browser

Figure 3: Strava-like activities heatmap

3.3 Strava activities and OpenStreetMap

As stated before, Strava isn't always perfect to retrieve activities data. For example, during a hike I often need to rest, especially when the ascent is very steep. I'm used to stop the recording when this happens, and then resume it when the climb starts again: Strava automatically merge these two phases, without considering the rest time (and certainly without caring of the stops motivations).

Considering that not every stop has a reason, I tried to reconstruct possible motivations of the stops I had during a hike to the Vigolana shelter, la Madonnina, completed during the last summer. For doing this, I isolated all the stops I had during that hike (Figure 8, only stops long more than 3 minutes), I extracted OpenStreetMap data from Altopiano della Vigolana[7] and selected all the Points Of Interest related to the area of the hike.

Surprisingly, the longest stop has been taken within 10 meters from a mapped viewpoint, the *Polsa* (Figure 9). As a matter of fact, that point is almost equidistant from the start point to the shelter, so it's a perfect point to rest (and to take pictures, obviously). Fun fact: this viewpoint[8] has been mapped last time in 2020 by the user Martin Larcher; at the end of the hike just analyzed, I spent the night in the Madonnina shelter with...Martin Larcher himself, along with his family!

3.4 Temporal analysis - Activities' performance

Another analysis included is not strictly related to the course, but it's a way to integrate the analysis done and to show also how GPX files can be used in several

[7] https://osmit-estratti.wmcloud.org/
[8] https://www.openstreetmap.org/node/910787221

ways. For example, how can an activity performance be measured? Is there any relation over time?

To address these questions, I decided to compare runs length, average pace and VO2Max[9] variations through a time series. The first two data have been extracted starting from Movingpandas Trajectories (as per in 3.1), while VO2Max variations have been provided by the Garmin smartwatch I used to track the activities (and which are not included in Strava by default).

Results are shown in Figure 10, and some information can be taken by the graphs:

- Activities' lengths increased over time, but the average pace didn't generally decrease
- VO2Max had a low point after the 2020 spring lockdown, as predictable
- Autumn 2020 and Winter 2021 were the most effective period according to VO2Max: during those months activities' lenghts started to raise, but average pace basically didn't decrease.

4 Conclusions and future improvements

As mentioned in the introduction, the complete analysis can be found on GitHub to be examined step by step. Obviously, every data exploration presented could be implemented for every other activity stored: the aim of this report is to present all the possibilities explored during the analysis, which are of course not limited to this project. More complex analysis can be done using the same dataset, for example by comparing more than one activities at once (e.g. section 3.1 with multiple rides).

A further improvements can be identified in plotting this kind of data in a 3D visualization. Strava heatmaps provide a similar service, but at the moment there are very few libraries or modules that offer similar results.

[9] VO2Max is a measure of maximum oxygen consumption per minute for any muscle contraction. It's a biological parameter that can be partially improved by trainings.

A Appendix - Images

Figure 4: Longest ride colored by pace

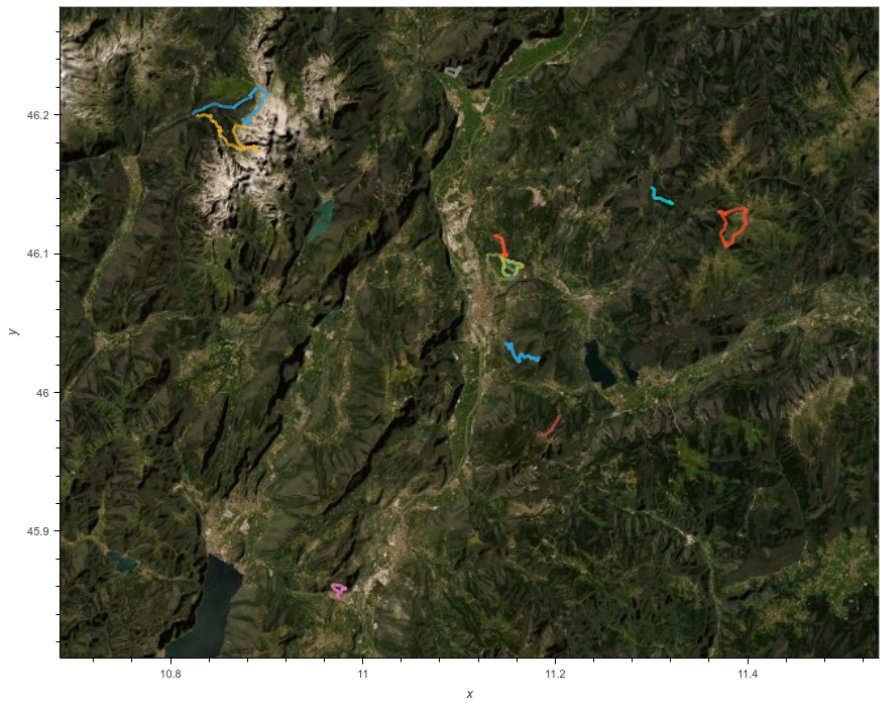

Figure 5: All hikes in the map

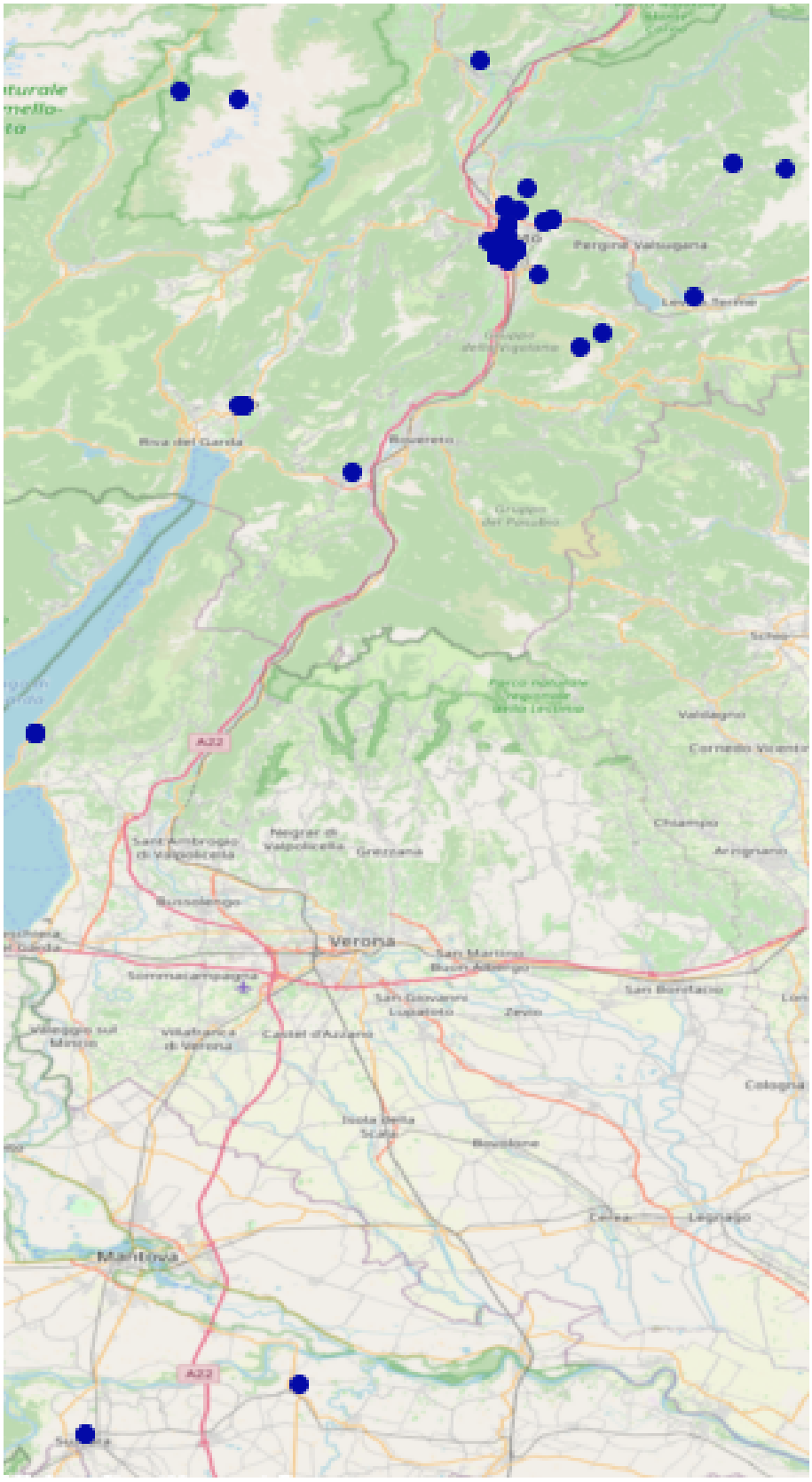

Figure 6: Activities starting points scatter plot

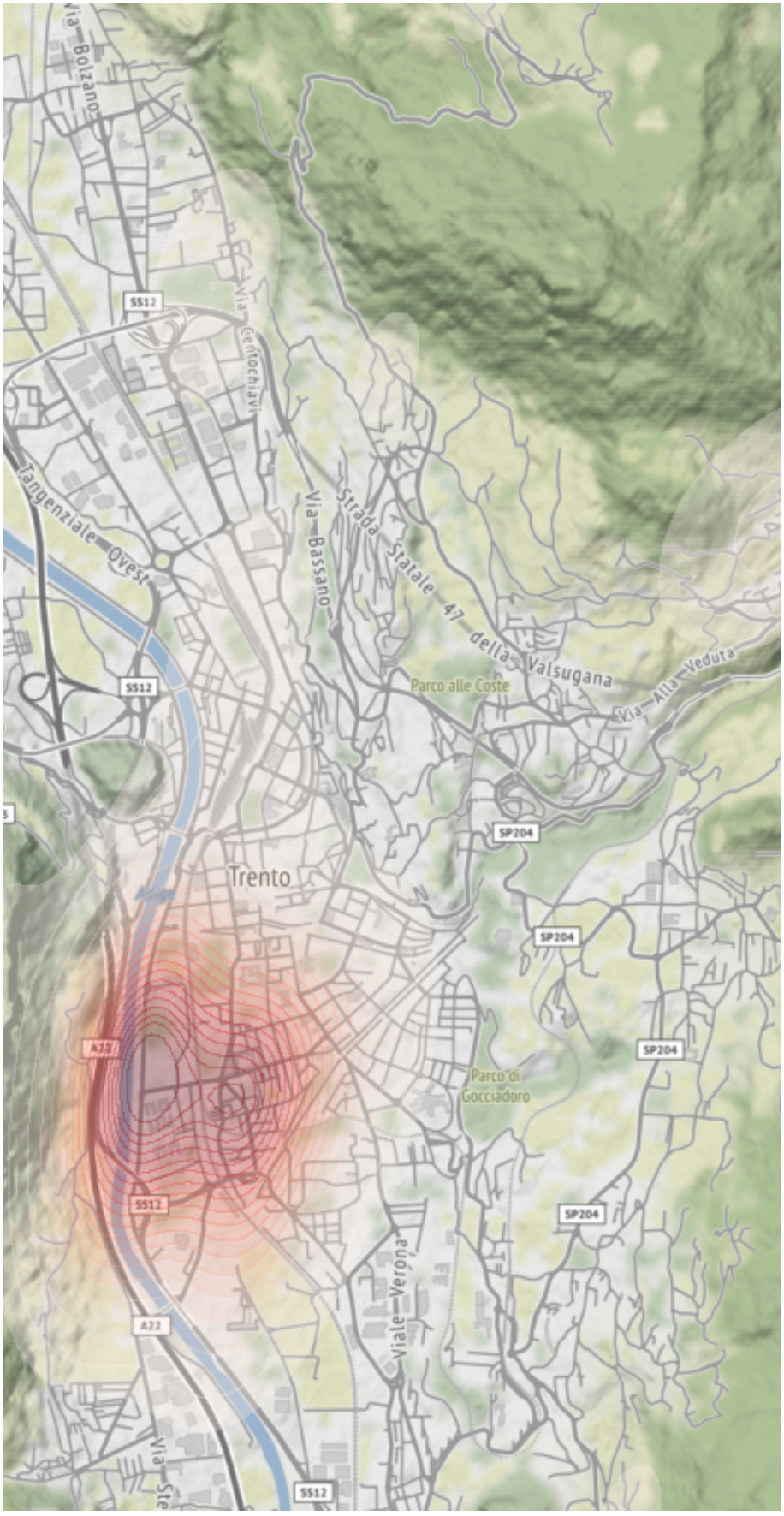

Figure 7: Kernel Density Estimation plot

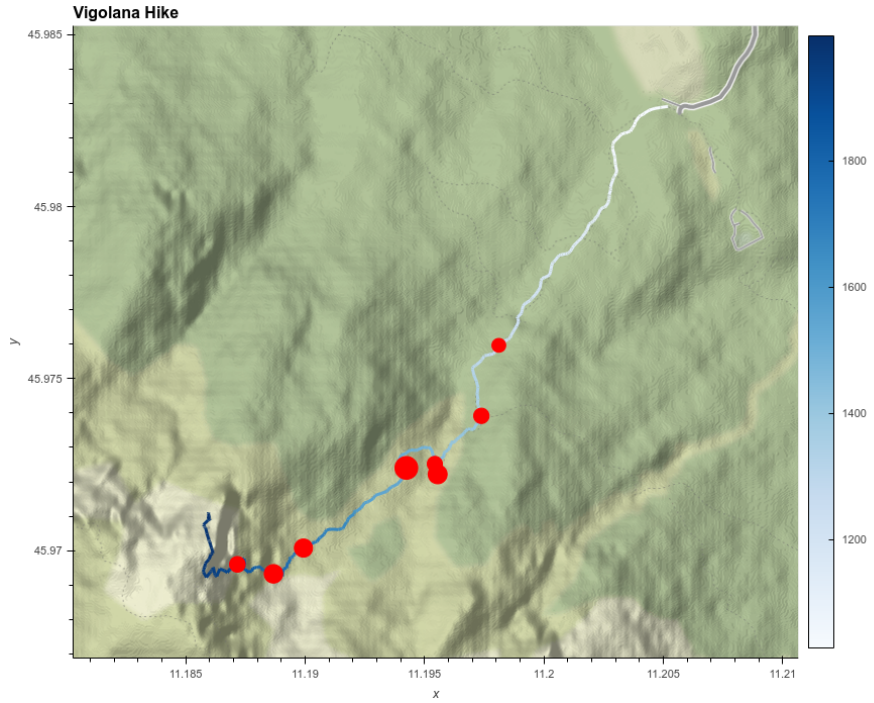

Figure 8: All the stop points during the Vigolana hike

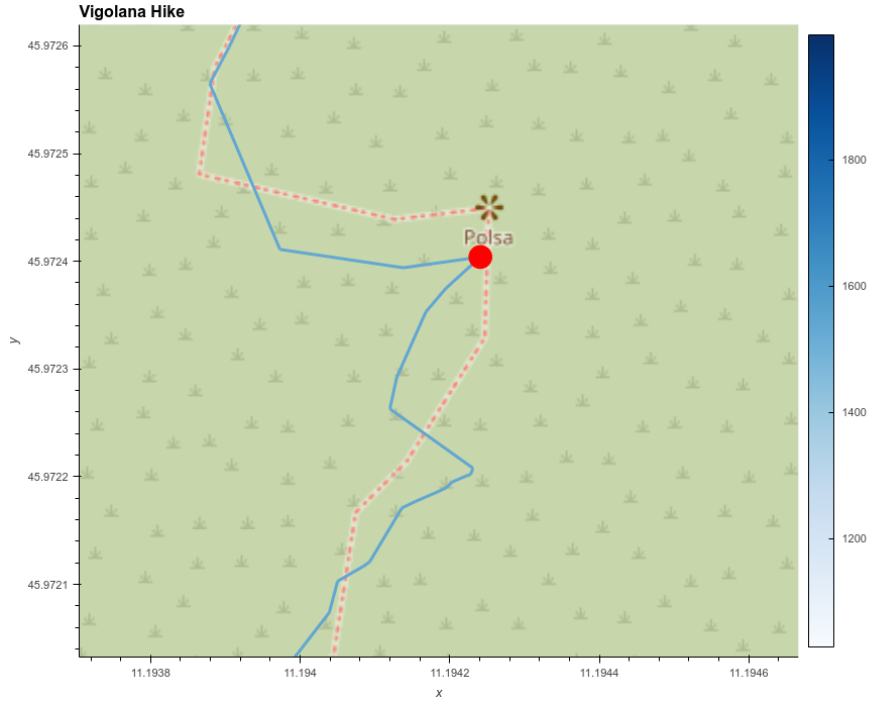

Figure 9: Stop points nearby the Polsa

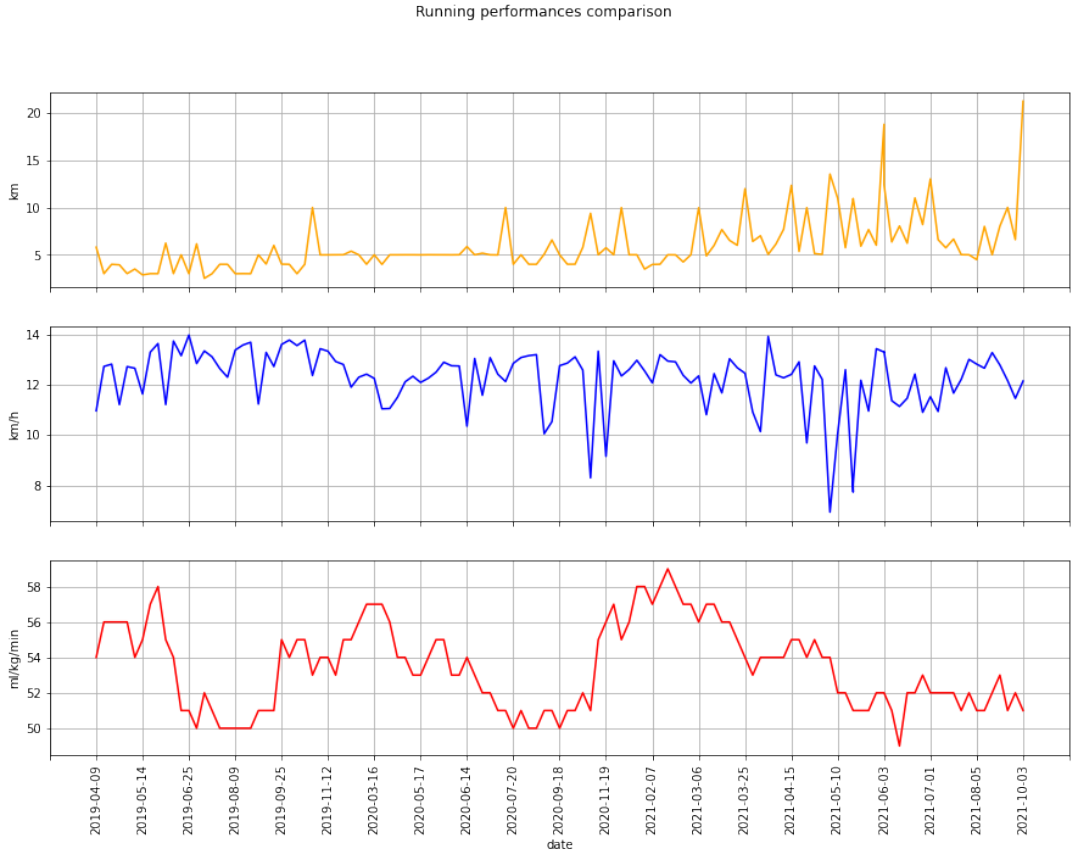

Figure 10: Length, average pace and VO2Max variations over time