# Correlation between CoVid-19 and Mobility in Italy

Giovanni Fiordeponti, Antonio Ionta, Silvia Marchiori

***Abstract*—The paper deals with presenting a data visualization regarding the effects of the CoVid-19 pandemic in Italy, during the year 2020.**

**This work focuses the attention towards an hypothetical governmental decision maker, who needs to evaluate the effectiveness of the current anti-contagion measures to plan out the future actions. For accomplishing this task, through a visual environment, a comparison in certain temporal intervals between the variation of people's movement and some indicators of the pandemic situation is performed, with respect to the different Italian Regions.**

**The work should try to give an answer to simple questions, that, at the same time, hides great complexity, like: has the lockdown be effective to reduce the contagion? In the case of an escalation in the severity of the pandemic, is there a direct connection with an increment in people's movement?**

## 1 Introduction

Public health officials often need to analyze simulations of epidemic models to improve preparedness, planning responses, and mitigate the impact of pandemics like *COVID-19*. During the planning and preparatory stages, they need to explore different scenarios and decision measures, and study the impact of these decision measures on controlling the pandemic's spread. In their analysis tasks, they often need to explore outputs in a *spatiotemporal environment* where they can analyze the spread patterns and the evolution of spread across space and time, interactively explore and filter, modify model parameters, and explore different scenarios. [1]

From these evidences the idea of building a visual environment able to fulfill a small part of the above mentioned requirements. The guidelines in the development of the project were the following:

- pandemic severity and mobility variation data are considered, which of restrictive rules application. The visual environment has to present in a coordinated way this source of information;
- the use of analytics, in particular of dimensionality reduction, is the chosen instrument for reveling possible correlation between the presented data, thus highlighting a sort of cause—effect relationship, if it actually exist;
- a form of visualization with space and temporal flexibility is needed;
- *the intended recipient*, i.e. a hypothetical decision-maker government figure, constitutes an audience with intermediate level of knowledge in the field of study: he is informed about the basic mechanisms which rule the pandemic, as his role requires, but, at the same time, he is not a proper expert of medical aspects. Therefore, the visualization can offer something not elementary, but should preserve the sufficient simplicity for not affecting the immediacy of the results interpretation;
- the representation must be a dynamic one. The user should be guided in the system's browsing, having the perception that each interaction step is a part of a longer path toward the discovery of a correlation.

## 2 Data

The data represented in the visual environment are collected from the following two public datasets:

- **COVID-19 Italy Data** [2]
  To inform citizens and make the collected data available, useful for communication and information purposes only, the Italian Department of Civil makes available, under license CC-BY-4.0, information about the situation of COVID-19 pandemic in Italy.

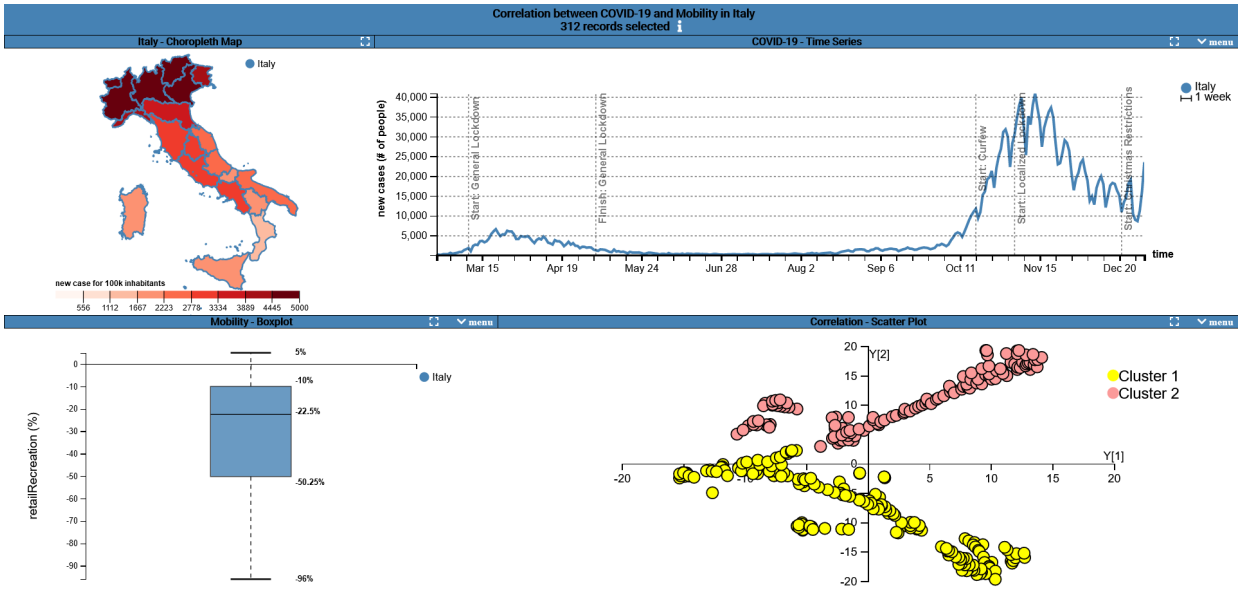


Fig. 1. The Visual Environment

- **Google COVID-19 Community Mobility Reports** [3] These Community Mobility Reports aim to provide insights into what has changed in response to policies aimed at combating COVID-19. The reports chart movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations and residential.

A subset of the data enclosed in these sources were selected and, through an activity of pre-processing, were aggregated into a single CSV file feeding the visualization.

Disregarding the elements referring to spatiotemporal collocation, the attributes reflecting the pandemic's severity, day by day, taken into account are: "Positives", "Death", "Healed", "Hospitalized", "Isolated", "Intensive Care".

The Community Mobility Reports show movement trends by region, across different categories of places. For each category in a region, reports show the changes comparing mobility for the report date to the baseline day, reported as a positive or negative percentage. A baseline day represents a normal value for that day of the week. The baseline day is the median value from the 5-week period Jan 3 – Feb 6, 2020. Places with similar characteristics for purposes of social distancing guidance are grouped: "Groceries-Pharmacy", "Parks", "Residential", "Retail-Recreation" and "Transit".

All the data are referred to Italy as geographical index, and to the period Feb 24 – Dec 31, 2020 as time window.

The overall source of data of the system has these exact dimensions:

- $6864$ tuples;
- $12$ ($15$ considering also id and spatiotemporal indications) attributes;
- $6864 \times 12 = 82368$ as *AS Index*.

Besides the data source described above, two additional sources where added:

- population for each Italian Region, contained within the file *region.js* of the repository and used in the Geographical Map described at Section 3;
- action taken by the Italian Government to address the health emergency, contained within the file *significative_dates.js* from the repository and used in the Time Series described at Section 3.

## 3 Visual Environment

The main result of this project is the production of the visual environment that is shown on Figure 1. Each component, besides a legend which shows the Italian Regions currently selected, has two buttons:

- a *focus button* which gives to the user the opportunity to zoom the involved visualization;
- a *menu button* which shows to the user a panel with different input elements, depending on the view is interested with.

The ecosystem is composed by four interactive visualizations:

- a *geographical map* of Italy showing COVID-19 data about pandemic situation;
- a time series showing COVID-19 data about pandemic situation;
- a *boxplot* showing mobility data;

- a *scatter plot* showing the output of the dimensionality reduction performed on the dataset.

The first two visualizations are, among the other things, of service for filtering the data from the geographical (regional granularity level) and temporal (daily granularity level) point of view, while the last twos can select records independently from geographical and time constraints.

### Choropleth Map

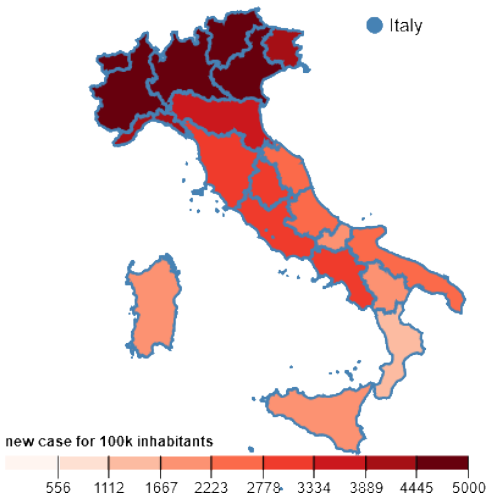


Fig. 2. The Choropleth Map

*Visualization*

Figure 2 acts as a choropleth map for illustrating the spatial spread of the selected COVID-19 parameter (e.g. "New Cases"): the different regions are coloured with respect to the number of cases for every $100k$ inhabitants using a sequential scale of colors from the d3 library itself [4], in particular:

- *d3.schemeReds* has been used for showing the new cases, the number of deaths and the person that has been hospitalized;
- *d3.schemeOranges* has been used for showing the isolated cases;
- *d3.schemeGreens* has been used for showing the healed cases.

Due to different mass testing capacities of the Italian Government during the different months, different ranges are used based on the maximum date range chosen by the user:

- until October the color scale goes up to $1000$ daily cases;
- from October the color goes up to $5000$ daily cases.

*Interaction*

The user can click on each region on the map for filtering the data focusing on a specific geographical area, by choosing up to three different regions. The feedback the user receives for the selection is that the region's borders becoming colored (with the specific color chosen for identifying that region data in the rest of the visual environment) and the legend is updated accordingly. The color's palette used to identify each of the 22 different italian regions has been chosen using the *colorbrewer web utility* [5].

### Time Series

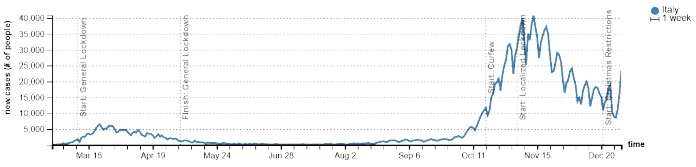


Fig. 3. The Time Series

*Visualization*

Figure 3 shows the trend in time of the pandemic, plotting the data of a particular COVID-19 related parameter, chosen by a select drop-down list. On the x axis the time, scanned by the months indication; on the y axis the units of the corresponding parameter, as raw number of people. Vertical lines corresponding to different actions taken by the Italian Government have been added to give to the final user a general overview of the measures taken by the government to stop the spread of the pandemic.

*Interaction*

Through brushing or zooming the user can temporally restrict the time range to a particular period, with a modification which is coordinated in the entire visual environment. Each path related to a different region can be focused to see in detail the number of cases during a particular day. The user, also, can insert manually the range of dates using two *HTML5 date inputs*. Finally, the time range can also be manipulated by selecting only particular days of the weeks, using a list of check boxes where each of them represents a particular day of the week.

### Boxplot

*Visualization*

Figure 4 represents the mobility data belonging to a certain category with indications of their statistical distribution. The category of mobility which will be represented is chosen by a select drop-down list. The vertical axis encodes the percentage reflecting the variation of the movement with respect to a "normal" day, while labels encoding the different quartiles are encoded near the boxplot itself.

*Interaction*

Through brushing the user can filter the data focusing on a particular range of variation in mobility: if multiple regions are selected, the user can brush each of them

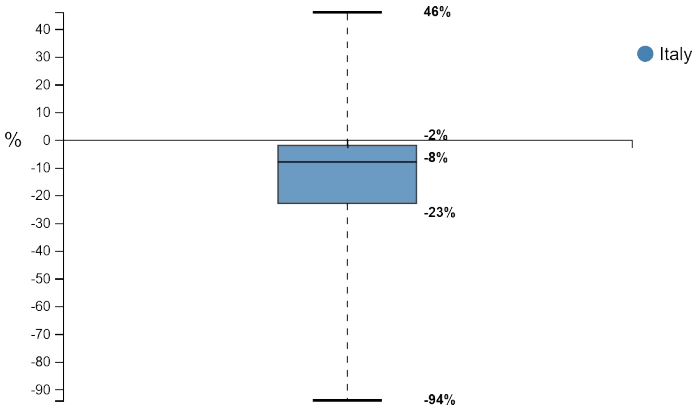


Fig. 4. The Boxplot

independently, giving him the opportunity to compare different variations of mobility. After the brushing, the modification is coordinated with the entire visual environment and the selected records appear as dots within the filtered boxplot region.

### Scatter Plot

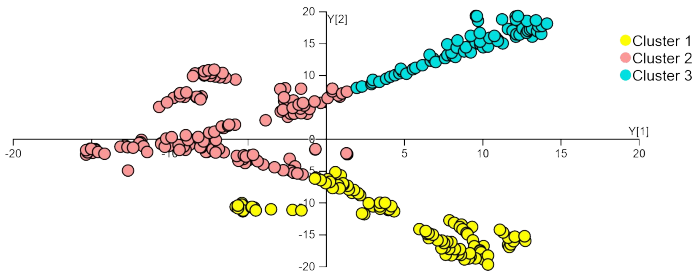


Fig. 5. The Scatter Plot

*Visualization*

Figure 5 plots the result of the dimensionality reduction and clustering executed on demand on the subset of data currently selected. This procedure will be described later during Section 4. Cluster colors have been chosen to avoid overlapping with respect to region colors and scale colors used within the choroplet map.

*Interaction*

The user can trigger the computation of multidimensional projection specifying also the number of clusters that the K-Means algorithm is going to use as main parameter: if only one cluster is selected, the points will be represented with the color corresponding to their related region. The scatter plot supports brushing for highlighting or zooming on a specific portion of the graphic. Finally, by focusing the records, a tooltip will be displayed showing for each of them the corresponding information, in particular:

- the current parameters for COVID-19 and mobility;
- the date (in format *YYYY-mm-dd*, plus the day of the week);
- the region where the record belongs.

## 4 Analytics

The reasoning part of the project is based on the execution of dimensionality reduction paired with a consecutive step of clustering on the currently selected portion of the dataset, using, respectively, the algorithms t-SNE and K-Means.

This process is performed on demand, exploiting a back-end service deployed on an Application server and continuously listening for requests. The back-end endorses the RESTful paradigm and it is developed in Python, with the support of Flask framework, and the use of sklearn library for executing the dimensionality reduction and clustering algorithms.

*t-distributed Stochastic Neighbor Embedding* (*t-SNE*) was chosen as algorithm for performing dimensionality reduction as it results, after some experimental attempts with other multidimensional projection techniques, the best approach for separating data into clusters, which is the ultimate objective of the analytics. In fact, the empirical evidence, is confirmed by the theory behind t-SNE, which is funded on plotting data according to the optimization of a cost function, focusing on similarities associated to distance, an approach which ends in amplifying separation between samples.

After the dimensionality reduction step, the clustering one is actuated, with the simple aim of assigning a label to each single data point. K-means is adopted as algorithm, relying on the fact that is probably the most well-known and fast clustering algorithm, easy to comprehend and to implement. The principal disadvantage of K-means is that it requires to select in advance how many groups/classes there are, which is not a trivial task. This criticality is mitigated in the system thanks to the possibility of choosing on demand the number of groups on which to execute the clustering, thus having the possibility of tuning the process with an immediate visual feedback. Ideally, the user can choose the number of clusters which reflects visually the best surfaces of separation among data case by case.

## 5 Insights

This section gives some intuitions about the capacity of the system to answer to the questions it is intended for. Is it possible to perceive a correlation between pandemic's aggressiveness and mobility given common spatiotemporal coordinates?

A semantical trip for reaching the objective is proposed as example. Fixed the temporal interval, the entry point is the geographical map: taking advantage of choropleth, the user can identify an area characterized by a not strong COVID-19 manifestation, as confirmed by the time series report. The dataset can be restricted by zoom or brushing filtering to verify, with the help of the dedicated boxplot, if the samples in consideration are effectively related to a great decrease of mobility.

This instinctive analysis can be corroborated or, on the contrary, weakened by the outcome of the dimensionality reduction plus clustering process: if selected data points are mainly grouped in the same cluster, the correlation is confirmed.

The intended user, a decision maker with intermediate knowledge of the field of study, can extract the following information from the data:

- **The differences between North and South Italy**: during the first phase of the pandemic, i.e. the first quarter of the year, the virus was widespread mainly in the north regions of the country. By comparing regions such as *Lombardy* and *Calabria*, the user can see with *the scatterplot* shown in Figure 6 how the two regions are *loosely correlated*, by appearing in two separated clusters, without the needs of K-Means.

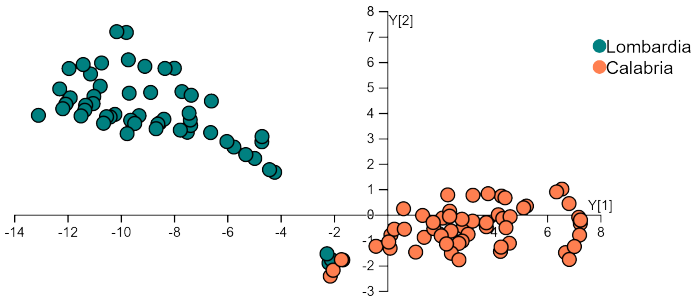


Fig. 6. Correlation between Lombardy and Calabria during lockdown

By *brushing* the *different quartiles* of *the boxplot*, the user can see in detail how different health situations tend to give the same outcome in people's mobility.

- **The critical role of weekends**: since the beginning of the pandemic, specialist are concerned by the social gatherings occurring at retail stores and recreational activities, such as restaurant. By considering Italy in its entirely and the whole year, the user can see only the behaviour of the pandemic and mobility during the weekends, i.e. Saturday and Sunday, as shown in Figure 7.

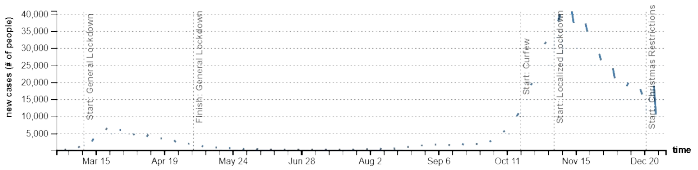


Fig. 7. Time Series with weekends selection

By applying *the dimensionality reduction with three clusters*, the algorithm can correctly separates three different situations:

1) the first is related to the period between October and December, when the government imposed more strict rules by adding nightlife curfew. The mobility goes from the median ($-25\%$) to the minimum ($-93\%$), which corresponds to a decreasing trend in *COVID-19 cases*;
2) the second is related to the period between June and October, when the government imposed soft rules. The mobility goes from the lower quartile ($-52\%$) to the maximum ($-3\%$), which corresponds to an increasing trend in *COVID-19 cases*, in particular starting from August, when summer vacations lead to a major infection rate;
3) the third is related to the first period of the pandemic between Febraury and May, when the government imposed a full lockdown on the whole population. The mobility goes from the lower quartile ($-52\%$) to the minimum ($-93\%$), in which the records are placed in descending order with respect to time, i.e. the more severe is the pandemic in the first stage of the lockdown, the lower is the mobility.

- **The criticality of summer holidays**: even if the first semester of the year was characterized by a strict and severe lockdown, during the second semester of the year, which corresponds to summer season, all the main restrictions on movements and gatherings were lifted by the government.
At this point in time one of the main concerns raised by authorities was the risk of creating outbreaks of the pandemic in the regions characterized by summer tourism. It is possible to focus on this aspect in the visualization by considering only the summer period and focusing on two regions, like Sardinia and Calabria: those two regions appear to be the ones with the larger positive variations in retail and recreations movements ($+45\%$ and $+105\%$, respectively), which ultimately lead to a severe outbreak of the pandemic in Sardinia after a series of night club events. [6]
That stated, one governmental figure could be interested in knowing how a summer season without restrictions has affected the last semester of the year, looking for common behaviors and isolated patterns.
An attempt to perform this type of study can be realized with the system: temporally restrict to the last part of the year (enclosing holiday time); then execute the analytics step (K=2). Filter the selection on the boxplot, focusing on the data points from the maximum to the median first, and from the median to the minimum after. Now observe the results on the scatter plot: we can notice that the two selections correspond almost exactly to the two groups of points, which, at the same time, refer to analogous trends in the pandemic's strength parameters for the regions. We can see this in Figure 8.
Therefore, it is possible to conclude that compa-

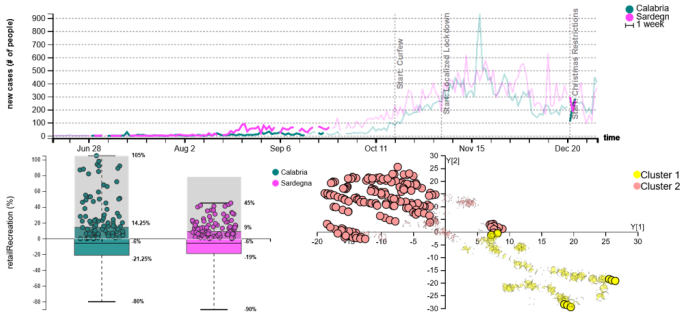


Fig. 8. The effect of summer holidays on the pandamic

rable variations in population movement are correlated to almost the same effect on CoVid-19 cases' fashion.

## 6 Related Works

How can be this work located with respect to existing literature about similar topics? To answer this question a succinct analysis of related scientific papers is going to be executed.

For the unique purpose of disambiguation, the references to the present work, i.e. the overall visual environment plus the analytic process, we'll be denoted with the initials *PW*. [7] proposes a huge panoramic of COVID-19 crisis visualizations, which offers numerous causes for reflection. First of all, it is important to highlight a fundamental point of discontinuity between the visualizations taken into account in [7] and *PW*. The firsts refer to something created to communicate information to the public, with the key concept that Information exposure in a public health crisis can impact people's attitudes towards and responses to the crisis and risks, and ultimately the trajectory of a pandemic. Therefore, the intended consumers are common people, with a non-deep knowledge of the analytics tools an no responsibility about political measures. This clearly differs from the intended user of *PW*, already discussed.

According to [7] COVID-19 crisis visualizations communicate a wide range of information about the pandemic. These messages can be classified as reflecting six purposes: informing of the severity; forecasting trends and influences; explaining the nature of the crisis; guiding risk mitigation; communicating risk, vulnerability, and equity; gauging the multifaceted impacts of the crisis. Note that these message categories are not mutually exclusive. It is possible to insert PW in the categories (1) and (4). Let's examine the visualizations in these groups.

Temporal visualizations focused on depicting the trajectory of the pandemic over time. The most basic solution was to visualize daily and cumulative numbers over a time period mostly using bar charts, line charts, and area charts. The wide adoption of linear scale charts may be because they are easier to understand for the general public. Geospatial visualizations displayed variables of interest over geographical maps to demonstrate which regions were impacted and compare how the impact differs by regions. Hence, for informing the severity of pandemic PW uses the most common solutions explored in the corpus of [7], which can be defined as the state of art for the matter, thus centering the objective of providing a non-ambiguous, almost familiar, visual environment concerning COVID-19 data.

While choropleth maps were the most frequently used geospatial visualization in the corpus of [7], they do present a challenge in that color-coding raw data may mislead viewers. For example, if a highly populated zone has the same case number as a less-populated zone, the color will be the same on the map. A reader may interpret the map as conveying that the intensity of cases is similar in the two zones, while in actuality, the intensity in the less-populated zone is higher. Normalizing data may help remove this type of bias, and, indeed, this advice is followed in PW.

Conceptual flattening-the-curve charts have become prevalent, particularly with the addition of a horizontal line marking "healthcare system capacity". They appear to emphasize personal responsibility in minimizing the unprecedented strain on the health system. However, these type of visuals have become controversial amidst the pandemic as they simplify complex pandemic situations (e.g., implying that it is good enough to keep the patient counts below the healthcare system's capacity) and they may also be misinterpreted by the public. Guiding risk mitigation is faced from another point of view PW, trying not to trivialize the situation which is, on the contrary, depicted with the rigor of an analytics process.

[7] emphasizes the importance of reporting on the source (and even the source of source) and recency of data. Crises such as the COVID-19 pandemic introduce particular threats to the production of trustworthy visual information. Having a trustworthy data source is crucial for creating visualizations, especially in times of crisis where misinformation and disinformation are rampant. [7] highlights also the importance of reporting the recency of data, that is, the date of data collection and visualization production. And, the discussion of how COVID-19 visualizations have changed over time, highlights how visualizations produced at one point in time may not be sufficient for depicting the state of the crisis at a later time. Indications followed by PW.

After an overview on a plethora of COVID-19 related visual environments, concreate examples of works similar to PW for objectives and datasets are now going to be examined.

[1] proposes a Visual Analytics solution which shares the target with *PW*: provide an environment to facilitate COVID-19 modeling, exploration and visualization for offering support to decision-maker user. What *PW* and [1]'s system have in common are also the visual

components adopted for describing the virus trend: the visual analytics environment consists of multiple linked views comprising the geospatial map view and time series visualizations.

The main distinction between the two projects regards the use of the data in the analytic part. [1] develops a sort of predicting model, which exploits pandemic data plus user inputs for actuating a simulation inherent to the effect that hypothetic restrictive measures might have in a certain context (with also a prevision about the time the effect needs to spread). It is a complex model, affected by numerous variables, presenting a challenging purpose. On the contrary, PW has a more restricted focus, identifying the eventual correlation between pandemic effects and mobility of population. It is carried on with a form of reasoning involving less parameters, but enhanced by the used of standard analytics methods (like dimensionality reduction plus clustering). While [1] looks ahead, to something that might happen, PW looks back to what happened, trying to give it an explanation.

The goodness of the choice of including mobility as a parameter at stake in facing the problem of offering a support in the management of the public health crisis is reflected also in one of the planned extensions of [1]: to incorporate additional datasets like connectivity information (road networks and google maps traffic datasets) to adjust the spatial dynamics.

Furthermore, there exist another work that deals with exactly the same *Google Mobility dataset* of *PW* in the context of COVID-19 crisis: [8] measures people's movement before and during the COVID-19 with a visualization in the form of *scrollytelling* in tandem with tile grid maps. It lacks in performing analytics on the dataset, while it offers an interesting and innovative approach of presenting results engaging users, which might be though as a possible extension of PW.

## 7 Conclusions and Future Works

This project tried to offer a visual environment of support for verifying the effectiveness of the main anti-COVID measure, social distancing, in Italy; an environment suitable for reasoning from the perspective of a user involved in governmental decisions. It took advantage of simple views, coordinated each other, for guiding the user through a semantic tour toward the discover of a correlation between represented data in a precise spatiotemporal window. One of its points of strength is the possibility of executing analytics (dimensionality reduction plus clustering) on demand, which allows user to dynamically tune its investigation.

Related works offer interesting cues for future refinements of the system. We are now going to report only a small subset of them.

The corpus analyzed in [7], among the other interesting examples, includes instances of an emerging solution, *the Growth Chart*, which plots the total number of the confirmed cases as the x-axis and the weekly confirmed cases as the y-axis. Applying the log scale to both axes and using the total case number as the x-axis (rather than time) helps reveal trends and patterns (e.g., demonstrating if cases are exponentially growing in a region or if a region is on the path to containing the virus). This can be a candidate visualization to be added in the ecosystem of the present work.

[8], as well, suggests an engaging form of story-telling, which might be exploited as complementary visual environment for exposing the results of a certain analysis conducted through the system here described.

Last, but not least, [1] is an example of how further the Visual Analytics can be integrated with the public health official context to be of help for preparing and exercising response plans in pandemic outbreak scenarios. This evidence hints the possibility of exploring deeper the correlation between pandemics trends and other kinds of data in a future work.

## References

[1] S. Afzal, S. Ghani, H. C. Jenkins-Smith, D. S. Ebert, M. Hadwiger, and I. Hoteit, "A visual analytics based decision making environment for covid-19 modeling and visualization," in *2020 IEEE Visualization Conference (VIS)*, 2020, pp. 86–90.

[2] P. Civile, "Dati covid-19 italia," https://github.com/pcm-dpc/COVID-19, 2020, [CC-BY 4.0].

[3] Google, "COVID-19 Community Mobility Reports," https://www.google.com/covid19/mobility/, 2020, [CC-BY 4.0].

[4] Data Driven Documents, "D3 scale chromatic references," https://github.com/d3/d3-scale-chromatic.

[5] M. H. Cynthia Brewer and T. P. S. University, "Colorbrewer, color advice for cartography," https://colorbrewer2.org/.

[6] ANSA, "Sardegna da covid free a sorvegliata speciale," https://www.ansa.it/sardegna/notizie/2020/08/20/solinas-volevo-i-tamponi-per-i-turisti_7cf43d37-e223-4a4c-9df2-bb398846c36e.html.

[7] Y. Zhang, Y. Sun, L. Padilla, S. Barua, E. Bertini, and A. G. Parker, *Mapping the Landscape of COVID-19 Crisis Visualizations*, ser. CHI Conference on Human Factors in Computing Systems (CHI '21). New York, NY, USA: CHI, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445381

[8] V. P., *Tile Narrative: Scrollytelling with Grid Maps*, Visualization for Communication (VisComm), Std., 2020. [Online]. Available: https://doi.org/10.31219/osf.io/xr64m