Human mobility and COVID-19 epidemic
Life Data Epidemiology

Nguyen Xuan Tung
Feb 2022

Università degli Studi di Padova

Overview

1. Introduction
2. Method
   - Data collection
   - Effective reproduction number R_t calculation
   - Forcasting with ARIMA
3. Results
   - Incidence curve compared with the mobility data
   - R_t values
   - Results of forcasting
4. Conclusion

Summary

With the spread of COVID-19 over last 2 years, a lot of researches are focused mainly on modelling the epidemic. However, earlier research indicates that mobility is also a crucial factor in spread of epidemic. In this project, our work is to investigate how Italian mobility data obtained from Google repository could be related to trend in epidemic spreading namely the prevalence using statistical tools such as ARIMA. Another part of our work includes estimating effective reproduction number R_t.

Introduction:
Idea of the project

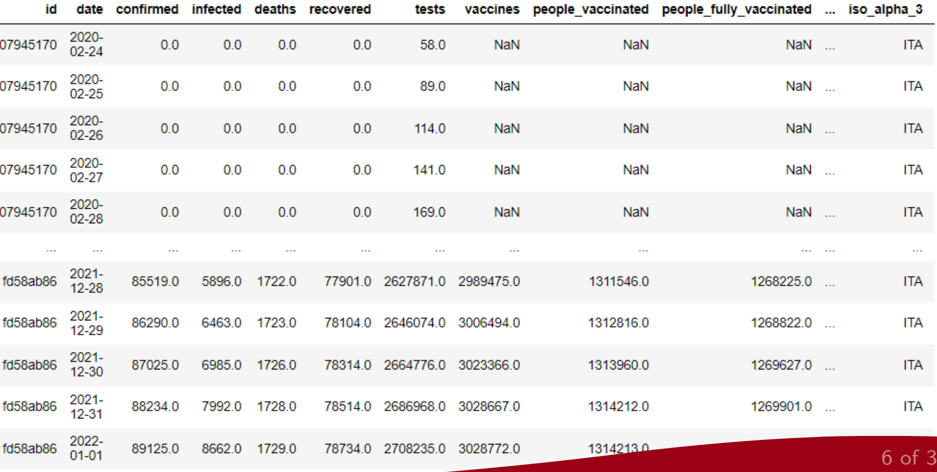
- The project is divided into 3 parts:
- ⟶ Invest the relationship between mobility and and COVID-19 spreading (investigate the effects of restrictions and health advice on the outbreak).
- ⟶ Calculate the reproductive ratio R_t and compare it with the real data
- ⟶ Use the statistical method (ARIMA), we predict the spread of COVID-19 in Italy's regions.

Method: Data collection

- The mobility data is collected in "Google COVID-19 Community Mobility Reports": https://www.google.com/covid19/mobility/
- The incidence data is collected in COVID-19 data hub: https://covid19datahub.io/articles/iso/ITA.html, 2020
- The data for effective reproduction number R_t is collected at: https://ourworldindata.org/

Method: Data analysis

- The analysed data contained the prevalence of the disease for each Italian region. The prevalence ('Infected') is obtained by subtracting the 'Confirmed', 'Death', and 'Recovered' column in the dataset.



id	date	confirmed	infected	deaths	recovered	tests	vaccines	people_vaccinated	people_fully_vaccinated	...	iso_alpha_3
07945170	2020-02-24	0.0	0.0	0.0	0.0	58.0	NaN	NaN	NaN	...	ITA
07945170	2020-02-25	0.0	0.0	0.0	0.0	89.0	NaN	NaN	NaN	...	ITA
07945170	2020-02-26	0.0	0.0	0.0	0.0	114.0	NaN	NaN	NaN	...	ITA
07945170	2020-02-27	0.0	0.0	0.0	0.0	141.0	NaN	NaN	NaN	...	ITA
07945170	2020-02-28	0.0	0.0	0.0	0.0	169.0	NaN	NaN	NaN	...	ITA
...	...	...	...	...	...	...	...	...	...	...	...
fd58ab86	2021-12-28	85519.0	5896.0	1722.0	77901.0	2627871.0	2989475.0	1311546.0	1268225.0	...	ITA
fd58ab86	2021-12-29	86290.0	6463.0	1723.0	78104.0	2646074.0	3006494.0	1312816.0	1268822.0	...	ITA
fd58ab86	2021-12-30	87025.0	6985.0	1726.0	78314.0	2664776.0	3023366.0	1313960.0	1269627.0	...	ITA
fd58ab86	2021-12-31	88234.0	7992.0	1728.0	78514.0	2686968.0	3028667.0	1314212.0	1269901.0	...	ITA
fd58ab86	2022-01-01	89125.0	8662.0	1729.0	78734.0	2708235.0	3028772.0	1314213.0			

Method: Data analysis

- To perform this analysis both incidence and mobility dataset were filtered, removing the information on days when the mobility reports do not track any data.
- To perform this analysis both R_t and mobility dataset were filtered, removing the information on days when the mobility reports do not track any data.

Method: R_t

- In any epidemic, R_t is the measure known as the effective reproduction number. It's the number of people who become infected per infectious person at time t.
- The most well-known version of this number is the basic reproduction number: R_0 when $t = 0$. However, R_0 is a single measure that does not adapt with changes in behavior and restrictions.
- As a pandemic evolves, increasing restrictions (or potential releasing of restrictions) change R_t.
- $R_t > 1$, the pandemic will spread through the entire population.
- $R_t < 1$, the pandemic will grow to some fixed number less than the population. The lower R_t, the more manageable the situation.

Method: Estimating COVID-19's R_t in Real-Time

- In this project, we used the Bettencourt Ribeiro's Approach to estimate real-time R_t using a Bayesian approach.
- This is Bayes' Theorem as we'll use it:

$$P(R_t|k) = \frac{P(R_t) \cdot \mathcal{L}(R_t|k)}{P(k)}$$

This says that, having seen k new cases, we believe the distribution of R_t is equal to:
- The prior beliefs of the value of $P(R_t)$ without the data.
- times the likelihood of R_t given that we've seen k new cases.
- divided by the probability of seeing this many cases in general.

Method: Estimating COVID-19's R_t in Real-Time

Importantly, $P(k)$ is a constant, so the numerator is proportional to the posterior. Since all probability distributions sum to 1.0, we can ignore $P(k)$ and normalize our posterior to sum to 1.0:

$$P(R_t|k) \propto P(R_t) \cdot \mathcal{L}(R_t|k)$$

This is for a single day. To make it iterative: every day that passes, we use yesterday's conclusion (ie. posterior) $P(R_{t-1}|k_{t-1})$ to be today's prior $P(R_t)$ so on day two:

$$P(R_2|k) \propto P(R_0) \cdot \mathcal{L}(R_2|k_2) \cdot \mathcal{L}(R_1|k_1)$$

And more generally:

$$\mathrm{P}(\mathrm{R}_t|k_t) \propto P(R_0) \cdot \prod_{t=0}^{T} \mathcal{L}(R_t|k_t)$$

With a uniform prior $P(R_0)$, this reduces to:

$$\mathrm{P}(\mathrm{R}_t|k_t) \propto \prod_{t=0}^{T} \mathcal{L}\left(R_t|k_t\right)$$

Method: Estimating COVID-19's R_t in Real-Time

We have to choose a Likelihood Function $\mathcal{L}(R_t|k_t)$
A likelihood function will say how likely a value of R_t is given an observed number of new cases k.
Given an average arrival rate of λ new cases per day, the probability of seeing k new cases is distributed according to the Poisson distribution:

$$P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

The Poisson distribution says that if you think you're going to have λ cases per day. In the likelihood we fix k in place while varying λ.

Method: Estimating COVID-19's R_t in Real-Time

We have $\mathcal{L}\left(\lambda_t|k_t\right)$ which is parameterized by λ but we were looking for $\mathcal{L}\left(R_t|k_t\right)$ which is parameterized by R_t. We need to know the relationship between λ and R_t.

The key insight to making this work is to realize there's a connection between R_t and λ: $\lambda = k_{t-1}e^{\gamma(R_t-1)}$
where γ is the reciprocal of the serial interval.

Since we know every new case count on the previous day, we can now reformulate the likelihood function as a Poisson parameterized by fixing k and varying R_t.

$$\lambda = k_{t-1}e^{\gamma(R_t-1)}$$

$$\mathcal{L}\left(R_t|k\right) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Method: Estimating COVID-19's R_t in Real-Time

We calculate the function to obtain the highest density intervals for R_t

And finally we make the function to calculate the Posteriors.

To calculate the posteriors we follow these steps:

1. Calculate λ - the expected arrival rate for every day's poisson process
2. Calculate our initial prior because our first day does not have a previous day from which to take the posterior.
3. Calculate each day's likelihood distribution over all possible values of R_t.

ARIMA

- ARIMA (autoregressive integrated moving average) is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.
- An ARIMA model can be understood by outlining each of its components as follows:
  $\longrightarrow$ Autoregression (AR)
  $\longrightarrow$ Integrated (I)
  $\longrightarrow$ Moving average (MA)

Method: ARIMA parameters

- The standard notations for ARIMA model are: p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used.
- The parameters can be defined as:
  $\longrightarrow$ p: the number of lag observations in the model
  $\longrightarrow$ d: the number of times that the raw observations are differenced
  $\longrightarrow$ q: the size of the moving average window

Method: ARIMA calculator

- Autoregressive models in ARIMA, we forecast the variable of interest using a linear combination of past values of the variable. The term autoregression indicates that it is a regression of the variable against itself.
- Thus, an autoregressive model of order p can be written as

$$y_t = c + \phi_1 y_t + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t$$

- Where: ϵ is white noise, and y_t is the lag value of model

Method: ARIMA calculator

- In terms of y, the general forecasting equation is:

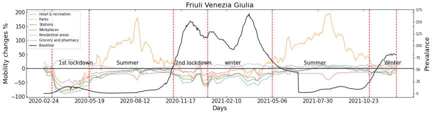
$$y_t = c + \epsilon_t + \phi_1 \epsilon_{t-1} + ... + \phi_q \epsilon_{t-q}$$

- Where: ϵ is white noise. Each value of y_t can be thought of as a weighted moving average of the past few forecast errors.

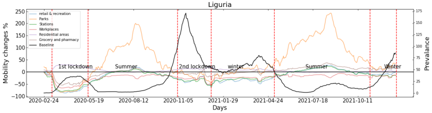
Method: ARIMA calculator

- Point forecasts can be calculated using the following three steps:
- 1 Expand the ARIMA equation so that y_t is on the left hand side and all other terms are on the right.
- 2 Rewrite the equation by replacing t with T + h.
- 3 On the right hand side of the equation, replace future observations with their forecasts, future errors with zero, and past errors with the corresponding residuals.

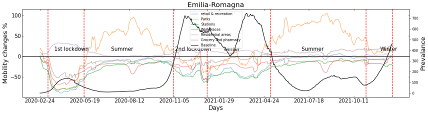
Results: Behaviour of Incidence curve compared with the mobility data



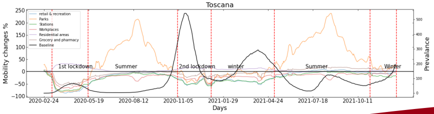
Friuli Venezia Giulia



Liguria

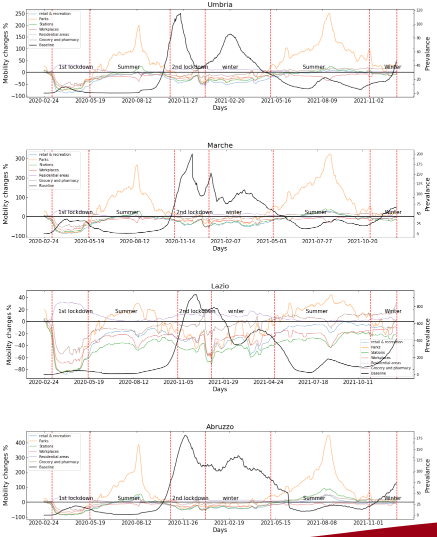


Emilia-Romagna



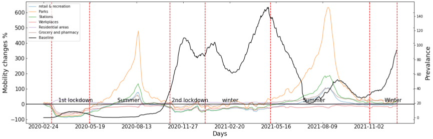
Toscana

Results: Behaviour of Incidence curve compared with the mobility data

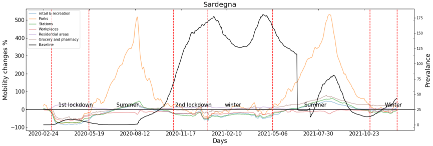


Results: Behaviour of Incidence curve compared with the mobility data

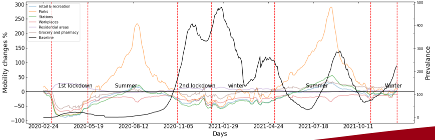
Calabria



Sardegna



Sicilia



Results: R_t

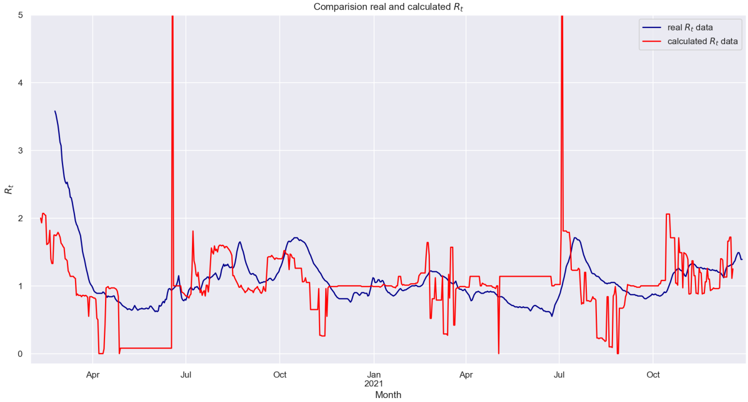


Figure: Comparison of real and calculated R_t

R_t value

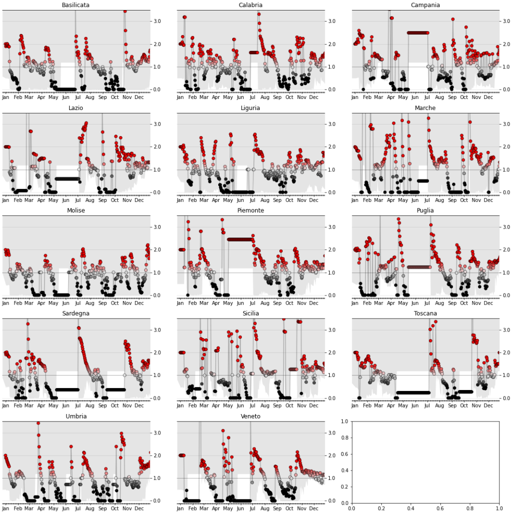


Figure: Reproduction number for Italy's regions

Results: R_t

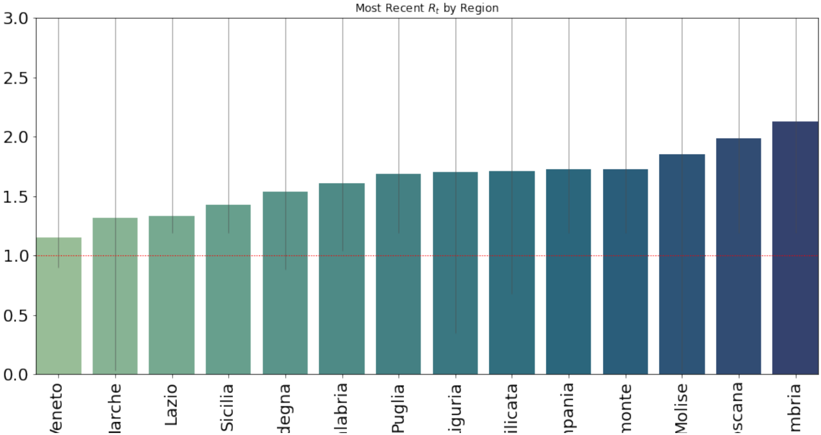


Figure: Most recent R_t by region

Results: ARIMA Forecasting

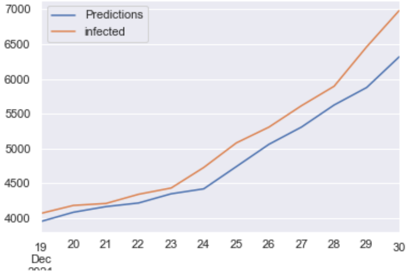


Figure: ARIMA model for prediction

The left figure represents the predictions of ARIMA Model against the test set, while the right figure performs the prediction of ARIMA Model against the data set

Results: ARIMA Forecasting

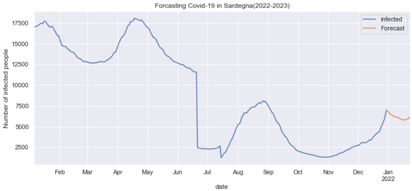


Figure: ARIMA model for prediction for Italy's regions

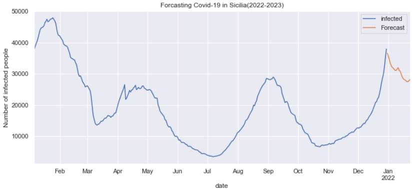


Figure: ARIMA model for prediction for Italy's regions

Results: ARIMA Forecasting

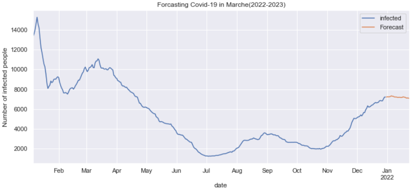


Figure: ARIMA model for prediction for Italy's regions

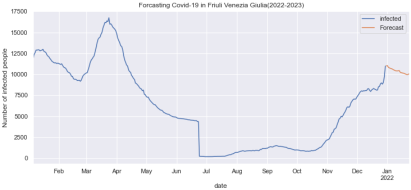


Figure: ARIMA model for prediction for Italy's regions

Results: ARIMA Forecasting

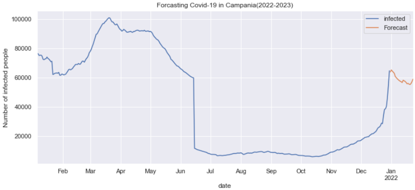


Figure: ARIMA model for prediction for Italy's regions

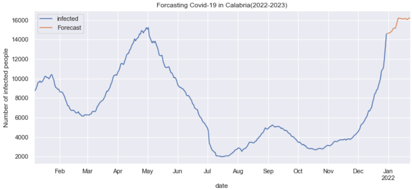


Figure: ARIMA model for prediction for Italy's regions

Conclusion

- The results show that there is a relationship between mobility and the COVID-19 pandemic.
- The analysis results indicate that various types of mobility have an econometric causality on the COVID indicators.
- Some types of mobility did not seem to have an influence on the number of infected people.

Conclusion

Results show a strong correlation between mobility and R_t. Specifically, retail/recreational activity, such as eating in restaurants, office work activities, and public transit usage all are associated with increases in transmission of the virus. Shopping at grocery stores and pharmacies has a smaller association, while affects associated with parks are minor; staying at home reduces transmission. Implications for reducing the spread of the coronavirus are that the reductions in mobility have been effective, but also need to be maintained for longer periods.

Conclusion

- We used the ARIMA model to predict infected cases for all regions in Italy. With careful considerations for the model's parameters, both manual and automated, we have found a best-fit model for our training data set
- The prediction of ARIMA model against test set is accurate.
- The accuracy of the ARIMA model prediction is appropriate and satisfactory.

Reference

https://www.medrxiv.org/content/10.1101/2020.05.06.20093039v3.f

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.02

https://www.sciencedirect.com/science/article/pii/S22113797210061

https://arxiv.org/ftp/arxiv/papers/2006/2006.01754.pdf

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.00

Thank You