Tracking the Invisible: **Early Detection** and Dynamics of **COVID-19 Variants** Through Genomic Sequencing

hQTC JC Talk, Oct 4

Marika D'Avanzo

University of Pavia, University of Naples Federico II, INFN

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

PhD One Health

The pandemic has starkly highlighted the **unpreparedness of human society to confront emerging diseases** and **its struggles in efficiently managing epidemiological waves.**

It is crucial to **cultivate a clear and straightforward understanding** of the global dynamics of epidemic spread.

Rome, Italy – March 2020
Credits: Franco Origlia / Getty Images

Rome, Italy - March 16, 2020
Credits: Stefano Montesi Corbis / Getty Images

Pisa, Italy – March 2020
Credits: Laura Lezza / Getty Images

Politecnico di Milano, Italy – March 2020
Credits: Emanuele Cremaschi / Getty Images

Milano, Italy – April 2020
Credits: Fototeca Gilardi / Getty Images

Codogno, Italy - March 2020
Credits: Miguel Medina /
Getty Images

Motivation

From a scientific perspective, the COVID-19 pandemic leaves behind a **wealth of valuable data**, presenting a unique opportunity to unravel the dynamics of pandemic diffusion.

A **clear and consistent understanding of the multi-wave pattern** is lacking in the scientific literature.

16.556.645

hCoV-19 genome sequence submissions on GISAID

The importance of monitoring variants

SARS-CoV-2 has continuously evolved in variants with different transmissibility, virulence, and immune escape potential.

Tracking and predicting variants is crucial to mitigate outbreaks and optimize vaccination campaigns.

Where are we at?

Compartmental models of the SIR type, complex network models and modern incarnations such as the eRG approach are being employed to characterise epidemiological data.
Including all effects poses a challenge due to the multitude of undetermined parameters, reducing the predictive power of models.

Where do we want to go?

The main objective is to conceive, validate, and establish an integrated system for the early detection of viral infectious diseases, discerning variants and their epidemiological relevance.

Study objectives

1.

Analyse transitions between COVID-19 variants across six European countries

2.

Identify key parameters such as t0 (time from first case to chain detection) and t_react (reaction time window)

3.

Differentiate between stable and non-stable variant chains to predict which ones will become dominant

Data and methodology

DATA SOURCES

Genomic sequences of the SARS-CoV-2 Spike protein from GISAID, for Germany, Italy, Sweden, Denmark, France, and Spain, from January 2020 to January 2024

Data and methodology

DATA SOURCES

Genomic sequences of the SARS-CoV-2 Spike protein from GISAID, for Germany, Italy, Sweden, Denmark, France, and Spain, from January 2020 to January 2024

Why Focus on the **Spike Protein** Instead of the Whole Genome**?**

- Key Role in Virus-Host Interaction
- Mutations with High Impact
- Genomic Efficiency and Focus
- Easier Data Comparison
- Public Health and Vaccine Relevance

Why Choose These **Countries?**

- Well-established genomic surveillance
- Robust sequencing efforts

Data and methodology

DATA SOURCES

Genomic sequences of the SARS-CoV-2 Spike protein from GISAID, for Germany, Italy, Sweden, Denmark, France, and Spain, from January 2020 to January 2024

CLUSTERING ALGORITHM

Unsupervised clustering algorithm (A. de Hoffer et al.) to group into 'variant chains.' Weekly clustering step, threshold of 100 for cluster distances, chains size > 5

Data and methodology

DATA SOURCES

Genomic sequences of the SARS-CoV-2 Spike protein from GISAID, for Germany, Italy, Sweden, Denmark, France, and Spain, from January 2020 to January 2024

CLUSTERING ALGORITHM

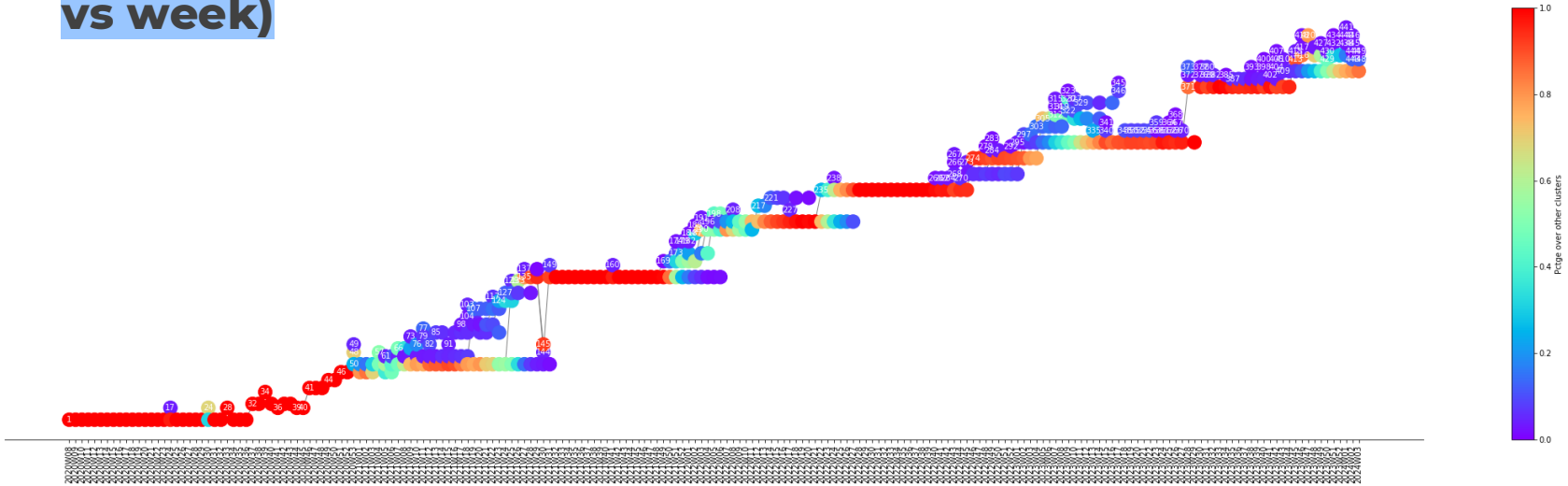
Unsupervised clustering algorithm (A. de Hoffer et al.) to group into 'variant chains.' Weekly clustering step, threshold of 100 for cluster distances, chains size > 5

FITTING MODEL

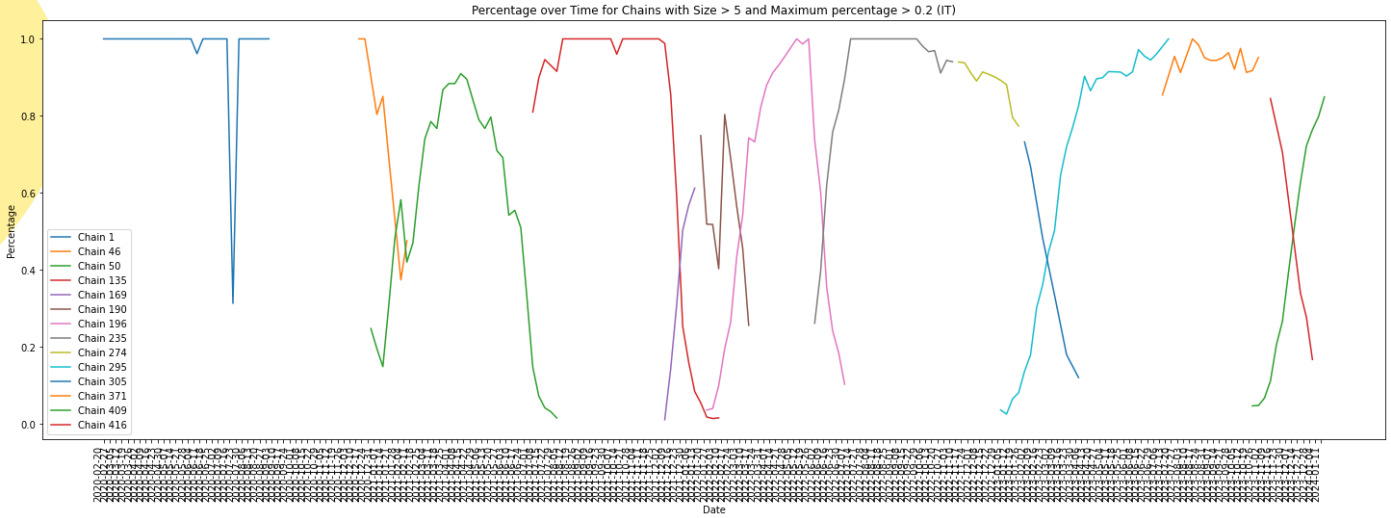
Six-parameter combined sigmoid function to capture the increase and decrease of variant dominance over time

Network graph

(percentage over other clusters vs week)

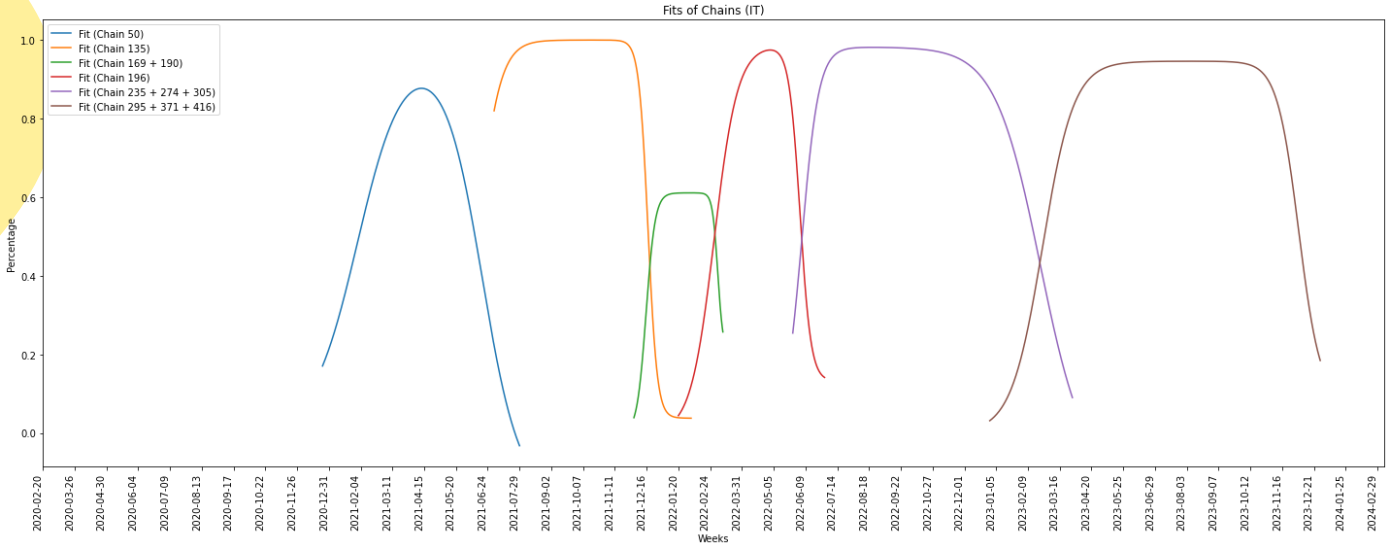


Italy, week 7 (2020W08) to week 211 (2024W03)



Fit procedure

$$\text{combined\_sigmoid}\,(x, a, b, L, a_2, b_2, L_2\,) = L\left(\frac{1}{1+e^{\frac{-(x-b)}{a}}}\right) - L_2\left(\frac{1}{1+e^{\frac{-(x-b_2)}{a_2}}}\right)$$



$$\text{combined\_sigmoid}\,(x, a, b, L, a_2, b_2, L_2\,) = L\left(\frac{1}{1+e^{\frac{-(x-b)}{a}}}\right) - L_2\left(\frac{1}{1+e^{\frac{-(x-b_2)}{a_2}}}\right)$$

a controls the steepness of the transition.
b determines the horizontal position of the transition's midpoint.
L determines the maximum height of the sigmoid.

Analysis of key parameters

The ratio of maximum values of the two sigmoid components

L/L_2

Reaction time window, the interval between detection and reaching 10% of the curve

t_react

t_10%_to_90%

Time in weeks for the curve to rise from 10% to 90%

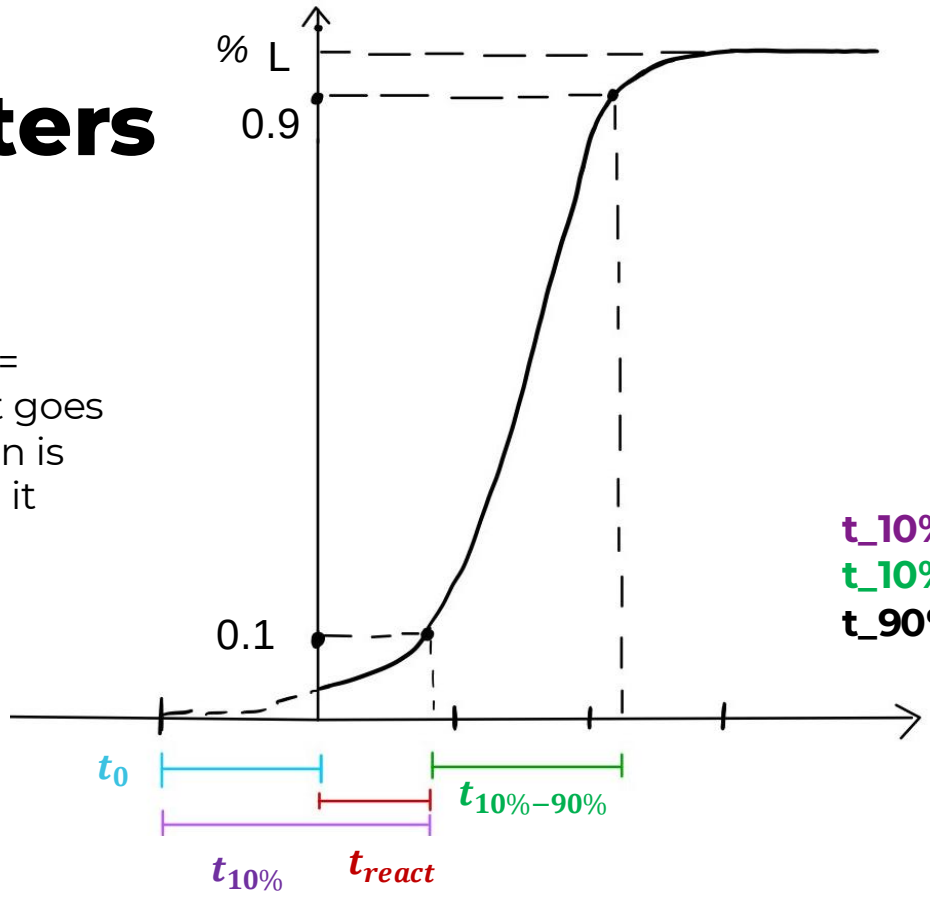
t_0

Time from the first case to the chain detection

Derived parameters

t0 = b - 6.91 * a = t_first_observation

t_react = t_10% - t0 = interval of time that goes from when the chain is detected and when it reaches 10%.



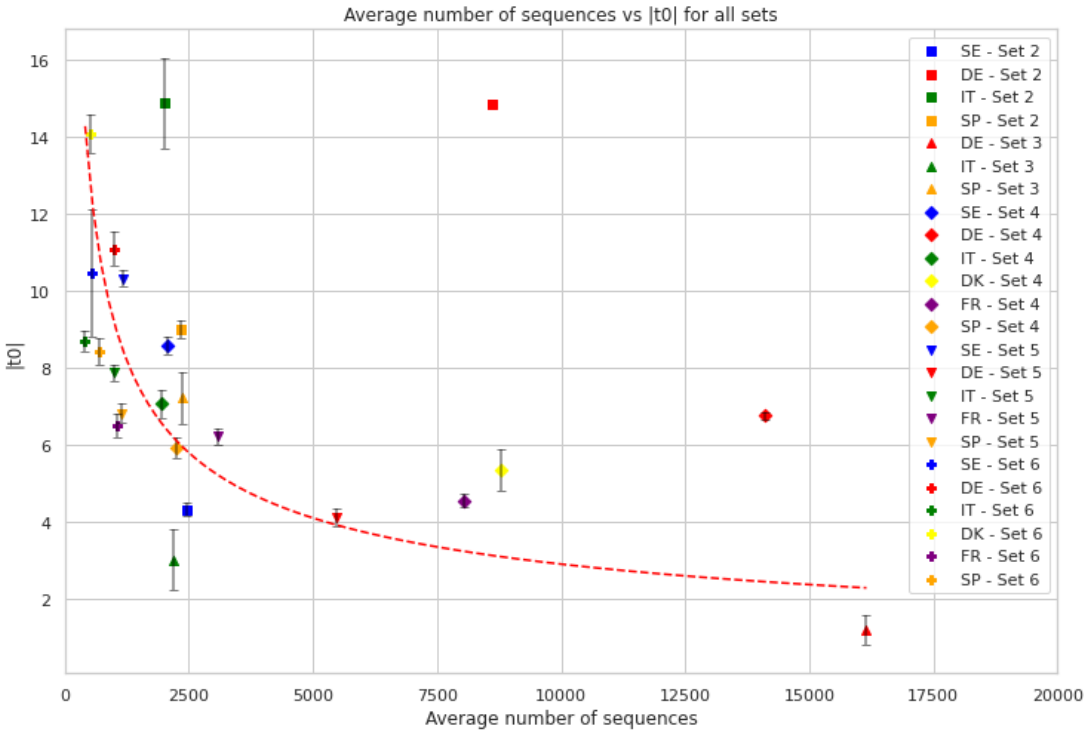
t_10% = b – 2.197 * a – t0

t_10%-90% = 4.39 * a

t_90% = b + 2.197 * a – t0

Calibration curve for t0

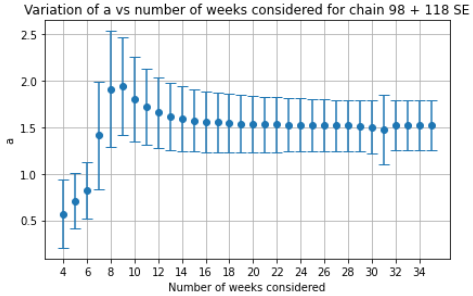
If we want t0 to be approximately 4 weeks, we need 5000 sequences per week.



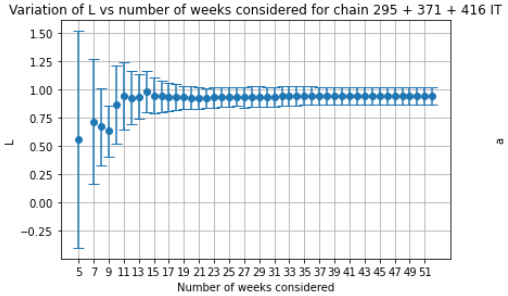
$t_0 \cong \frac{A}{\sqrt{n\_seq}}$ and consequently, $n_{seq} \cong \left(\frac{A}{\widehat{t_0}}\right)^2$, $A = 290 \pm 29$

Predicting Stable and Unstable Chains

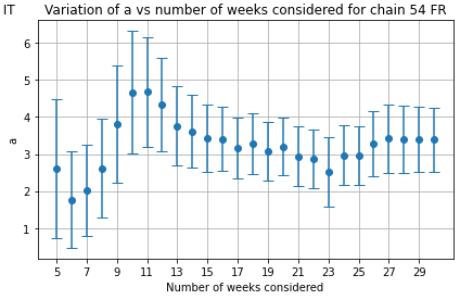
Can we differentiate analysing the parameters a, b and L derived after just 3 or 4 weeks of data?



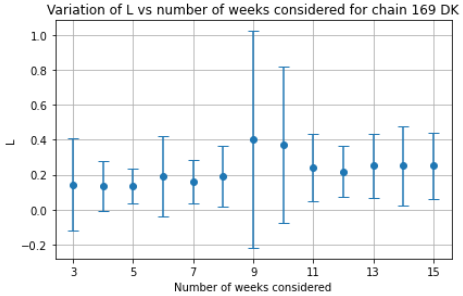
Relevant chain



Relevant chain

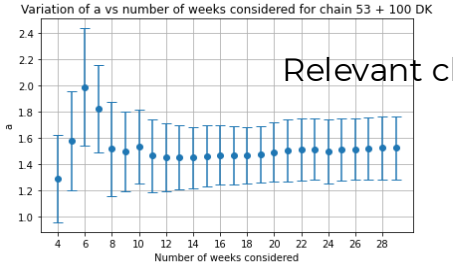


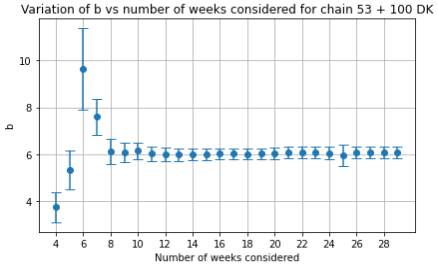
Non relevant chain

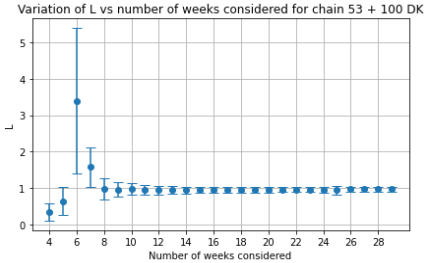


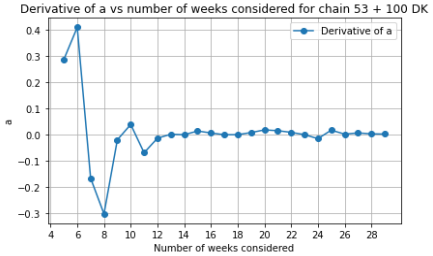
Non relevant chain

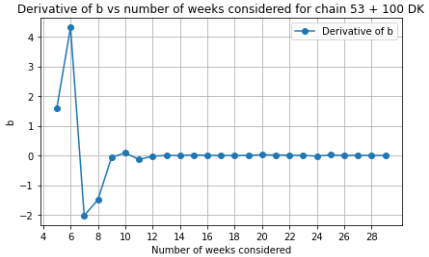
Relevant chain

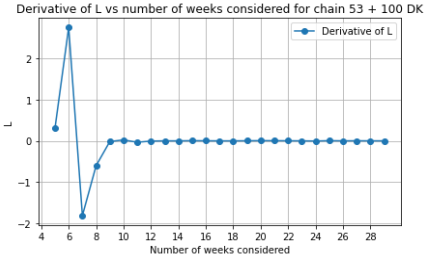




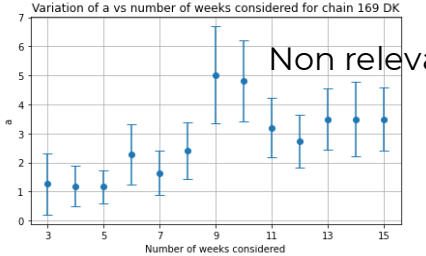


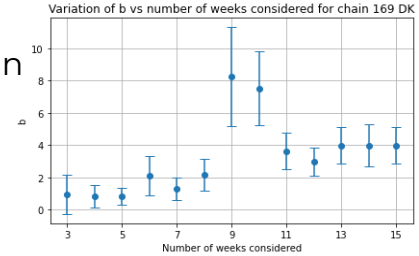


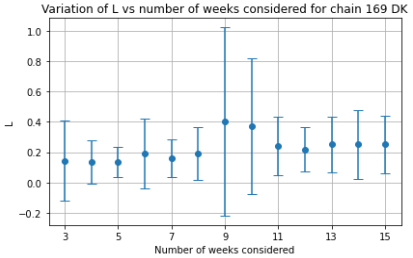


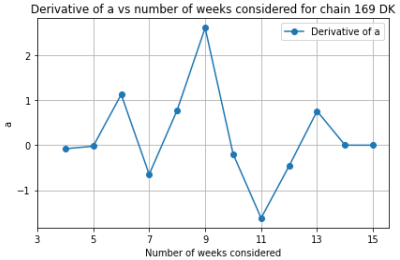


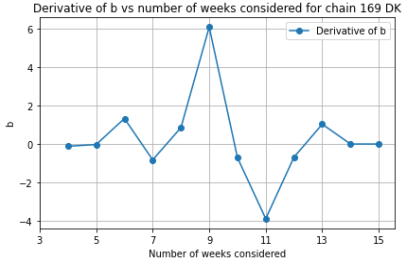
Non relevant chain

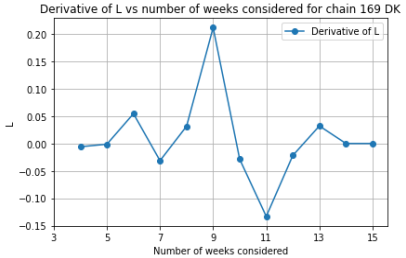










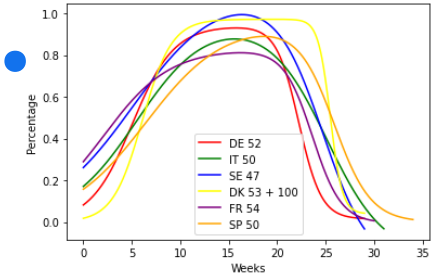


Main **Results**

Variant Transitions Across Countries

Differences are not significant when transitioning from one nation to another

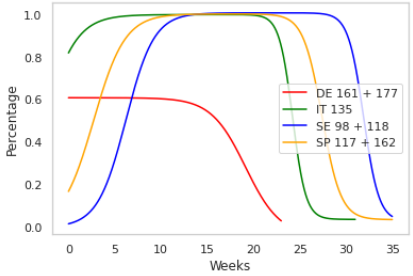
1.



B.1.1.7 (Alpha)

2020W51 to 2021W33

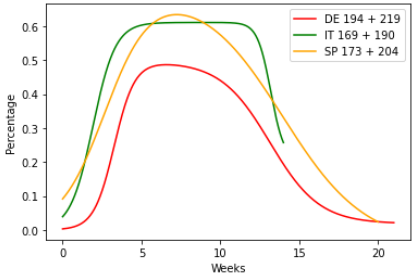
2.



B.1.617.2 (Delta)

2021W19 to 2022W06

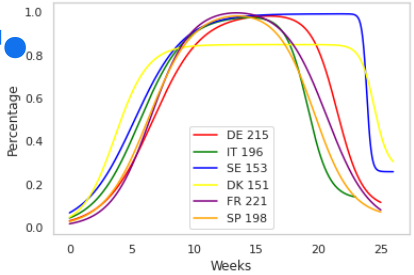
3.



Omicron (B.1.1.529)

2021W48 to 2022W17

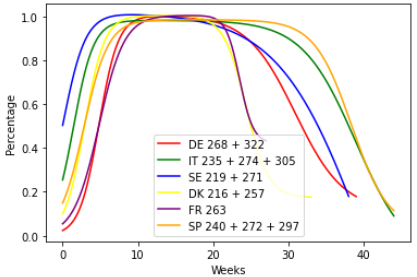
4.



BA.2 (Omicron)

2021W50 to 2022W28

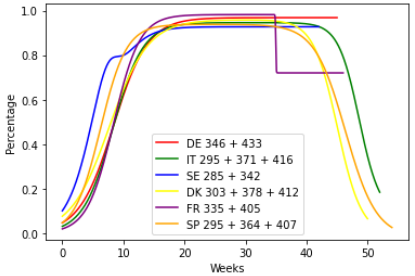
5.



BA.2.86 (Omicron)

2022W18 to 2023W14

6.

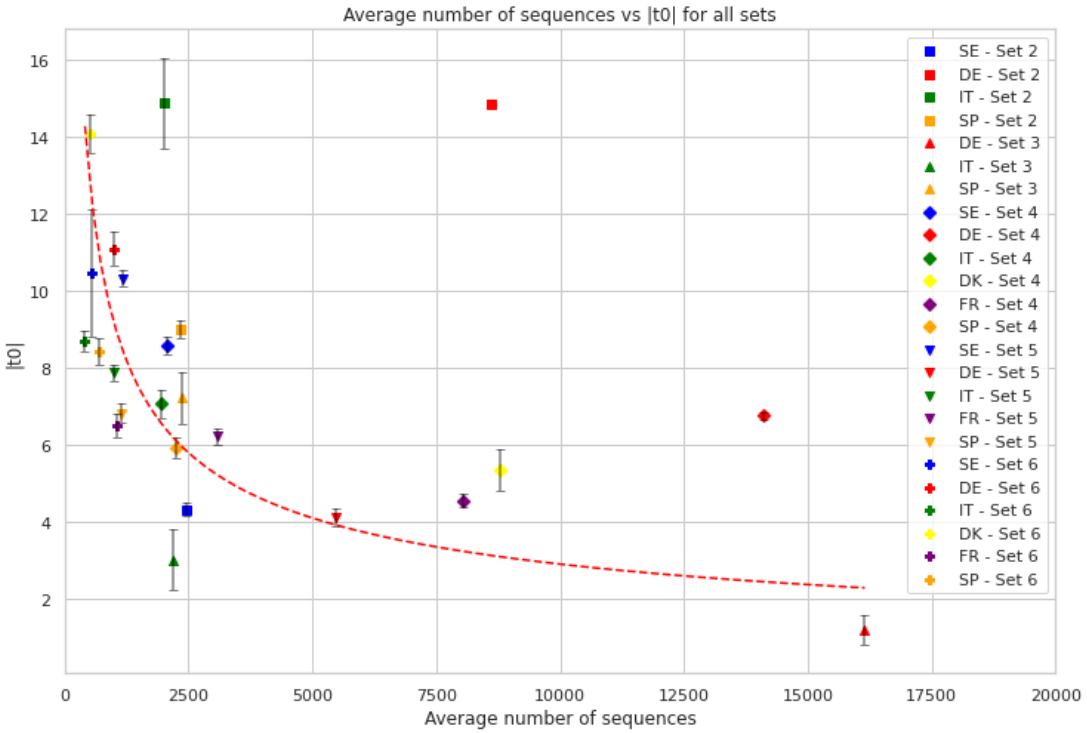


XBB.1.5 (Omicron)

2022W52 to 2024W00

t0 and Early Detection

By ensuring enough sequences per week, we can reliably detect new variants within a short window.



Predicting Dominance

We could predict which chains will become dominant within 2-3 months, using the parameters based on the first 4 weeks of data.

Practical Implications

Combining effective genomic sequencing with early prediction models can significantly improve variant monitoring systems, serving as a low-cost, efficient surveillance tool.

Future applications

Early warning system

The parameter $t0$ and quantitative indications on the sequencing campaign could be integrated into public health surveillance systems to provide early warnings of new variant emergence.
This would allow health authorities to respond faster, possibly before a variant becomes widespread.

AI integration

Leveraging machine learning algorithms on top of these predictive models could enhance their accuracy and speed, allowing for real-time variant tracking. Artificial Intelligence could refine the detection of stable variant chains using early-stage data.

Thanks!

DOES ANYONE HAVE ANY QUESTIONS?

marika.davanzo01@universitadipavia.it

marikadavanzo