Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

A comprehensive study of the phylodynamics of SARS-CoV-2 in Europe

Master Thesis

Cecilia Valenzuela Agüí

ceciliav@student.ethz.ch

Department of Biosystems Science and Engineering
Computational Evolution
ETH Zürich

Supervisors:
Prof. Dr. Tanja Stadler, Dr. Timothy Vaughan, Sarah Nadeau

March 24, 2021

Acknowledgements

I thank Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Abstract

The abstract should be short, stating what you did and what the most important result is.

TODO abstract

This project focus on the spatial dynamics of the early spread of SARS-CoV-2 in Europe. We apply a novel approach based on the Multi-type Birth Death phylodynamic model to infer structured population dynamics jointly with between-subpopulation transmission rates from viral genome sequences. The inferred epidemic trajectories for the combined outbreak responsible for the observed sequence data will allow us to better understand the entry into and early spread of SARS-CoV-2 in Europe.

Contents

Chapter 1

Introduction

We have many sources of information, but all of them are imperfect in some aspect. Bayesian phylodynamics incorporates different beliefs and data to get a detailed picture of the epidemic. Genetic sequences as an objective source of information about virus evolution and transmision together with viral phylogenetics, transpotation data and geographic data to understand the dynamics and case counts to guide the magnitutde of the numbers.

We lack absolute numbers of the pandemic time resoluted without reporting issues dependent. We lack understading of the dynamics, from where to where, how often, how long.

We can use travel data as a proxy, or we can incorporate it in a model. Lemey et al. Difference we get absolute numbers, not only the rates. And we use BD model instead of coalescent, better to describe beginning of a epidemic. While due to the complexite and time consuming model we dow scaled to europe and only thhe initial phase with the first countries to experiment relevant sars cov 2 epidemics.

the first reported cases were in China in early December [],

In a phylodynamic analysis, we aim to integrate genetic and epidemiological information to understand pathogen evolution and transmission dynamics. We can learn about the epidemic spread and the interactions between hosts from the imprint that these events leave in virus phylogenies. This is possible for RNA viruses, and in specific for SARS-CoV-2, because virus genetic evolution and epidemic processes are in the same time scale [?] [?]. More precisely, the field of phylodynamics studies how phylogenetic trees are being generated and infers the population parameters behind that process.

TODO introduction

- Importance of understanding the spread of the virus to prevent future outbreaks
- About phylodynamics/phylogeographics? Use of genetic sequences as a

source of information combined with other sources of information as travel data

- About what we know of the introduction of sars-cov-2 in Europe and early dynamics till 8 March
- About what we know of the case counts in Europe? Can sequences help?
- Introduce/formulate the questions: case counts, first introductions, ?migration vs within region transmission?, migration patterns, ?border closures.

Chapter 2

Material and Methods

SARS-CoV 2 genome data set. We assemble a dataset of 360 genetic sequences from December 2019 to March 8 2020, obtained from publicly available data on GISAID [Shu and McCauley, 2017] (accessed on November 2020). We follow the Nextstrain workflow for the curation of the dataset [Nextstrain-ncov, 2020]. Sequences with incomplete collection date, less than 27.000 bases in length or with more than 3.000 unknown bases are omitted. Also, sequences from known clusters of transmission or from the same patient are excluded. The resulting worldwide sequence dataset is aligned with MAFFT. The begining and the end of the alignment are masked respectively by 100 and 50 sequences as well as sites 13402, 24389 and 24390, identified by Nextstrain as prone to sequencing errors.

To focus on the early dynamics in Europe we select sequences from China, the origin of the epidemic; France and Germany, the European countries with the earliest cases; and Italy and Spain, the European countries with the biggest outbreaks in March. To take into account the dynamics in other regions of Europe we include a group of 50 sequences from other European countries. We limit our sample of Chinese genomes to sequences collected before January 23, the starting date of the lockdown in Hubei.

Due to the large (and unprecendented) number of available genetic sequences for SARS-CoV-2, we need to subsample the alignment. Each sequence is subsampled with a probability equal to the probability of having a reported case in that country the day of sample collection, inversely weighted by the probability of having a sequence in GISAID that day in the country. With this subsampling protocol, we aim to get a constant sampling proportion across the full period for each country.

This dataset of genetic sequences is the main source of information in our phylodynamic analysis. The goal is to infer the phylogenetic tree, i.e. the evolutionary tree-shaped relationship among the sequences, together with the epidemiological transmission parameters that gave rise to it. These parameters are defined within a population dynamic model and will inform us about the epidemic that the viral genetic sequences come from.

The multitype birth death model. To study the early dynamics of SARS-CoV-2 in the European countries, we use a simplified version of the multitype birth-death model described in [Kuhnert et al., 2016], following the analysis in [Nadeau et al.,]. Birth death models are compartmental population models with high flexibility that describe the process of epidemiological transmission. The stochastic formulation of these models are used in phylodynamic analyses [Stadler et al., 2012]. In the multitype version, we consider a structured population in types or subpopulations with characteristic within-subpopulation dynamics and migrations between them. In our case, the subpopulations are the different locations of the samples.

The process starts with one infected host in one of the subpopulations, e.g in subpopulation i, who can infect another individual at rate λ_i, become uninfectious at rate μ_i by death or recovery, migrate to another deme j at rate m_{ij} or be sequenced at rate ψ_i to become part of the phylogenectic tree. This process depicts the full transmission dynamics and specifically, the generation of the phylogenetic tree that we observed from our sequence data.

Under this model, we are able to compute the likelihood of the multitype birth-death parameters for a given tree. This likelihood is derived in [Kuhnert et al., 2016] by considering the probability of an individual evolved as observed in the tree. This derivation is analogous to the work in [Stadler et al., 2013], which is based on ideas from [Maddison et al., 2007].

We parameterize our model in terms of the effective reproductive number $R_i = \frac{\lambda_i}{\mu_i + \psi_i}$, a key value in epidemic control and understanding, the rate of becoming uninfectious $\delta_i = \mu_i + \psi_i$, the probability of and individual to be sequenced $s_i = \frac{\psi_i}{\mu_i + \psi_i}$ and the migration rates between locations $m_i j$.

Epidemic trajectories and structured trees. We will refer to the full sequence of transmissions, recoveries/deaths, migrations and sampling events as an epidemic trajectory. One set of sequences, and therefore one phylogenetic tree, is the product of one epidemic trajectory. Moreover, the tree represents a fraction of the events in the epidemic trajectory, those that involve the sampled individuals. In epidemiology, we aim to learn about the true epidemic trajectory since we usually have incomplete information caused by unreported cases or unknown transmission chains.

From the sequences metadata, we know the location of the tree tips. However, if we know the epidemic trajectory we also know the location of the lineage at any point in time in the tree. We will refer to the phylogenies with ancentral locations mapped onto the tree as structured trees [Vaughan et al., 2014]. In Fig 2.2 we show the epidemic trajectories corresponding to two different subpopulations and the structured tree of a set of samples. The change of ancestral location, represented by a change in color from blue to red in the tree, is caused by a migration event from one subpopulation to the other, depicted in the epidemic trajectories.

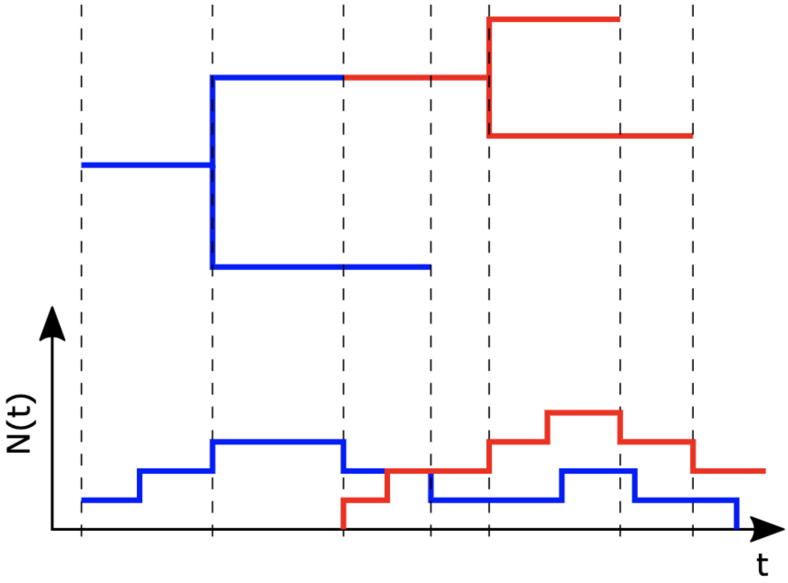

Figure 2.1: Example of structured tree (up) and epidemic trajectories (bottom) for two types. The horizontal axis represents time, and for the epidemic trajectories the vertical axis is the population size N(t). Figure from Tim Vaughan.

Use figure from EpinInf paper?

Bayesian Inference with BDMM-Prime. We want to infer the epidemic trajectories under the multitype birth-death model fit to a set of samples collected throughout SARS-CoV-2 early epidemics in Europe. In order to do this, we have to estimate the joint posterior distribution of the structured tree $\mathcal{T}_c$, the epidemic trajectory $\mathcal{E}$, the substitution model parameters θ and the multi-type birth death parameters η. This posterior probability can be expressed with Bayes' formula as:

$$P(\mathcal{T}_c, \mathcal{E}, \theta, \eta | A) = \frac{P(A|\mathcal{T}_c, \mathcal{E}, \theta, \eta) P(\mathcal{T}_c, \mathcal{E}, \theta, \eta)}{P(A)} \tag{2.1}$$

where A is the sequence alignment.

We make the following independence assumptions:

$$P(A|\mathcal{T}_c, \mathcal{E}, \theta, \eta) = P(A|\mathcal{T}, \theta) \tag{2.2}$$

$$P(\mathcal{T}_c, \mathcal{E}, \theta, \eta) = P(\mathcal{T}_c|\mathcal{T}, \eta) P(\mathcal{E}|\mathcal{T}_c, \eta) P(\mathcal{T}|\eta) P(\theta) P(\eta) \tag{2.3}$$

where $\mathcal{T}$ is the rooted time tree without ancestral locations. Thus, we can express Equation 2.1 in terms of the conditional distributions:

$$P(\mathcal{T}_c, \mathcal{E}, \theta, \eta | A) = P(\mathcal{T}_c | \mathcal{T}, \eta) P(\mathcal{E} | \mathcal{T}_c, \eta) \frac{P(A|\mathcal{T}, \theta) P(\mathcal{T}|\eta) P(\theta) P(\eta)}{P(A)} \tag{2.4}$$

$$= P(\mathcal{T}_c | \mathcal{T}, \eta) P(\mathcal{E} | \mathcal{T}_c, \eta) P(\mathcal{T}, \theta, \eta | A) \tag{2.5}$$

$P(\mathcal{T}_c|\mathcal{T}, \eta)$, $P(\mathcal{E}|\mathcal{T}_c, \eta)$ and $P(\mathcal{T}, \theta, \eta|A)$ are the posterior probabilities of the structured tree, the epidemic trajectory and the joint posterior of the phylogenetic tree (without ancestral locations) and the model parameters. The tree likelihood $P(A|\mathcal{T}, \theta)$ is the probability of the sequence alignment and can be efficiently evaluated using Felsenstein's pruning algorithm [Felsenstein, 1981]. The tree prior, $P(\mathcal{T}|\eta)$ also called phylodynamic likelihood is derived from the multitype birth-death model. $P(\theta)$ and $P(\eta)$ represent our prior belief in the distribution of the population and substitution model parameters.

We use a Markov chain Monte Carlo (MCMC) Metropolis-Hastings algorihtm to approximate this posterior. Since the MCMC only uses ratios of posterior probabilities, we avoid calculating the marginal likelihood $P(A)$. This algorithm is implemented in BDMM-Prime package [BDMM-Prime, 2020] for BEAST 2.6.3 [Bouckaert et al., 2019]. BDMM-Prime first samples from $P(\mathcal{T}, \theta, \eta|A)$, then in a second step these samples are augmented by sampling from $P(\mathcal{T}_c|\mathcal{T}, \eta)$ and $P(\mathcal{E}|\mathcal{T}_c, \eta)$ and adding these variables to obtain the overall posterior $P(\mathcal{T}_c, \mathcal{E}, \theta, \eta|A)$. In this way, $P(\mathcal{T}_c|\mathcal{T}, \eta)$ and $P(\mathcal{E}|\mathcal{T}_c, \eta)$ are only calculated for a subset of samples instead of every MCMC step.

While it is also possible to include $\mathcal{T}_c$ and $\mathcal{E}$ directly into the Bayes' rule as implemented in EpiInf package [Vaughan et al., 2019], the factorization of the posterior in these three terms allow us to use the standard birth-death-sampling tree prior implementations in BDMM to compute $P(\mathcal{T}, \theta, \eta|A)$ [Kuhnert et al., 2016] [Scire et al., 2020]. This will speed up the analysis with almost no overhead compared to the standard BDMM inference and can be used in pre-existing multitype analyses withoud additional MCMC.

Stochastic mapping of ancestral locations and epidemic trajectories To sample from $P(\mathcal{T}_c|\mathcal{T}, \eta)$, BDMM-Prime implements a stochastic mapping algorithm based on the work by [**?**]. A set of differential equations is numerically integrated over $\mathcal{T}$ to obtain the marginal probability of a location value at any point along each branch. From this probabilities, we can derive time-dependent rates that define the changes of the location along the tree according to a continuous-time Markov process. Then, we can simulate forward in time from the root of the tree the location trajectories down the tree edges. For a detailed derivation of the stochastic mapping of the algorithm refer to [Vaughan, 2021].

In the case of $P(\mathcal{E}|\mathcal{T}_c, \eta)$, following Bayes'rule:

$$P(\mathcal{E}|\mathcal{T}_c, \eta) = \frac{P(\mathcal{T}_c, \eta|\mathcal{E})P(\mathcal{E})}{P(\mathcal{T}_c, \eta)} = \frac{P(\mathcal{T}_c|\mathcal{E})P(\mathcal{E}|\eta)}{P(\mathcal{T}_c|\eta)} \tag{2.6}$$

We can easily compute the probability of a structured tree for a given epidemic trajectory $P(\mathcal{T}_c|\mathcal{E})$ as described in [Vaughan et al., 2019]. Each node in the structured tree must correspond to a compatible event in the epidemic trajectory for this probability to be nonzero. If we simulate an epidemic trajectory directly from $P(\mathcal{E}|\eta)$, it is very likely that the simulated events will not match with the events in the tree and $P(\mathcal{T}_c|\mathcal{E})$ will be 0. To avoid this problem, we simulate from $P^*(\mathcal{E}|\eta)$, which guarantees trajectories with non-zero probabilities by enforcing the tree events. Provided we weight these trajectories accordingly we can efficiently sample from $P(\mathcal{E}|\mathcal{T}_c, \eta)$:

$$P(\mathcal{E}^{(a)}|\mathcal{T}_c, \eta) = w_a^* P^*(\mathcal{E}^{(a)}|\eta) \propto P(\mathcal{T}_c|\mathcal{E}^{(a)}) \frac{P(\mathcal{E}^{(a)}|\eta)}{P^*(\mathcal{E}^{(a)}|\eta)} P^*(\mathcal{E}^{(a)}|\eta) \tag{2.7}$$

The epidemic trajectories are simulated with an adaptive tau-leaping algorithm [Gillespie, 2000] in BDMM-Prime. This method is based on the Gillespie algorithm [Gillespie, 1977] but the simulation time is divided in small intervals and the number of events is drawn from a Poisson distribution for each interval. This algorithm is more efficient than Gillespie when we have a high number of individuals since we have to update the rates less often.

We simulate a fixed number of trajectories, called particles, for each set of population parameters η and structured tree $\mathcal{T}_c$. The trajectories are simulated until some point and are weigthed as described in Equation 2.7. We resample using importance sampling to get rid of low-weighted trajectories and then resume the simulation of all the particles from the sampled trajectory ending point. This process is repeated until the end of the simulation. Then, we sample one single trajectory from the set of final trajectories with important sampling and record each event and its timing to future analysis [Vaughan, 2021]. All these steps are implemented by BDMM-Prime trajectory logger and we use it in our analysis.

We conducted two different analyses. One with constant migration rates as in the standard multi-type birth death model and the other one with time-changing migration rates informed by travel data. We use the BDMM-Prime package [BDMM-Prime, 2020] implemented in BEAST 2.6.3 [?]. We run 10 parallel chains for analysis, of 10^7 iterations each with different initial values. These chains are assesed for convergence using Tracer v.1.7.1 and then combined after removing a 10% burnin. We check that the effective sample size (ESS) is greater than 200 for all parameters. The stochastic mapping of structured trees and epidemic trajectories is performed in a second step using the trace and tree logs from the MCMC. We use 300, 1000, 3000 and 10000 particles in the tau-leaping simulation of epidemic trajectories, and a tolerance of 0.03 for selecting tau leap length.

Incorporating air travel data and geographic distances. In order to inform our migration rates with external information, and to avoid having to estimate a large number of migration rates we implement a generalized linear model (GLM) for the migration matrix based on [Lemey et al., 2014] and [**?**]. This GLM model parametrizes the migration rates as a log linear function of a number of potential predictors, Equation 2.8. For each predictor x_i, the GLM parameterization includes a coefficient β_i, which quantifies the effect size of the predictor, and a binary indicator variable δ_i, that allows the predictor to be included or excluded from the model. The parameter c represents the overal magnitude of migration. This means that if every indicator is 0, every migration rate will be equal to this parameter. In the MCMC, we estimate the effect size of each of the predictors, as well as their inclusion probability $\mathbb{E}[\delta_i]$.

$$m_{ij} = c\exp(\sum_{i=1}^{p} \delta_i \beta_i xij) \tag{2.8}$$

In a similar way to the GLM formulation in [Lemey et al., 2020] we consider as predictors air travel data and geographical distance between locations. We define three time periods in our analysis: from the origin of the epidemic to January 23 (Hubei lockdown), from January 23 to end of February, from March 1 to March 8. The international travel during these months decreased drastically due to the increasing awareness about the SARS-CoV-2 pandemic as is shown in Figure **??**. Thus, we would expect the migration rates to change during the time of the analysis.

The air travel data is obtained from EUROSTATS transport datasets *avia_paexcc* and *avia_paincc*, in particular the passengers carried departures values for December 2019, January, February and March 2020. We compute the average daily flux of passengers for each of the time periods. The geographical distance is defined as the great circle pairwise distance between the centroids of the countries, computed based on the Natural Earth project (the 1:50m resolution version) world map 2013. The predictors are log transformed, a pseudocount is added to make all values positive, and then they are standardized, following the description in [Lemey et al., 2014].

Model specifications and priors. Since we focus on the early epidemic outbreak, we expect unimpeded spread of the virus that can be described by the exponential growth of the infected population and no significant decrease in the number of susceptibles over time [Boskova et al., 2014]. Thus, we assume a constant basic reproductive number R_0 for each of the European countries before the lockdown in the Lombardy region in March 8. However, in China the first wave

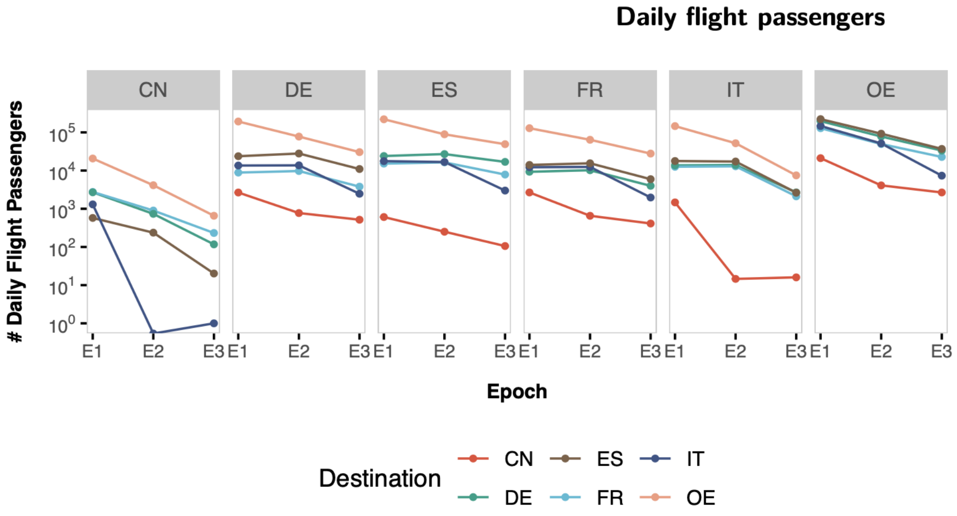

Figure 2.2:

improve figure

of the epidedmic had almost concluded in March so we can not make the same assumption. We fixed the reproductive number in China and include three different periods, changing on January 23 and February 10 based on [Pan et al., 2020] [Wu et al., 2020] [Xiao et al., 2020]. Before January 23, we use the estimation in [Park et al., 2020] [Billah et al., 2020] to fix the basic reproductive number R_0 in China to 2.9 . From January 23 to February 10 and from February 11 to March 8, we calculate the average effective reproductive number based on the estimations in [COVID-19-Re, 2020] [Huisman et al., 2020] and obtain the R_e in China to be 1.1 and 0.43 respectively.

In Table 2.1 we show the values and prior distributions used in the Bayesian inference of the multitype birth death model and substitution parameters. Several of these model specifications are identical to those from [Nadeau et al.,].

Data availability. The SARS-CoV-2 genetic data can be downloaded from www.gisaid.org [Shu and McCauley, 2017]. Supplementary Table **??** lists accession numbers for the genetic sequences used in this study. Reported SARS-CoV-2 case counts are obtained from the country official agencies indicated in Table **??** . Flight data can be obtained from EUROSTATS https://ec.europa.eu/eurostat/.

todo

todo

Code availability and analyses reproducibility. All the code is available at ... To ensure the reproducibility of the analysis, we have implemented the whole workflow with Snakemake [Köster and Rahmann, 2012]. The workflow is a modified version of Nextrain workflow [Nextstrain-ncov, 2020] with the additional rules for BEAST 2 analysis and results processing.

add snakemake graph

Parameter	**Value or Prior**	**Rationale**
Nucleotide substitution model	HKY + Gamma	Unequal transition/transversion rates, unequal base frequencies, rate heterogeneity among sites with 4 categories
Clock rate	0.0008	Approximatelly 24 mutations per year \cite{10}
Location of origin	Hubei	Putative pandemic origin
Time of origin	Lognormal (-1, 0.2)	Median 26 October, 95% IQR 22 August to 8 December 2019
European countries reproductive number	Lognormal (0.8, 0.5)	Median 2.2, 95% IQR 0.8 to 5.9
China reproductive number	Origin - Jan 23: 2.9 Jan 23 - Feb 10: 1.1 Feb 11 - Mar 8: 0.43	Fixed based on \cite{}.
Becoming uninfectious rate	36.5 y-1	Period between infections and becoming uninfectious assumed exponentially distributed with a mean of 10 days
Sampling proportion		Upper bounds based on reported cases:
China	Uniform (0, 0.12)	
France	Uniform (0, 0.07)	
Germany	Uniform (0, 0.07)	
Italy	Uniform (0, 0.01)	
Spain	Uniform (0, 0.08)	
Other European	Uniform (0, 0.03)	
Sampling start date	December 23, 2019	Just before first sample from China
Sampling end date (only China)	January 23, 2020	Only included Chinese sequences before Lockdown in Hubei
Migration rates (no GLM analysis)	Lognormal(0,1)	Median, IQR
Migration rates (GLM analysis)	Uniform(0,50)	7.3 days as mean of exponential time to travel for every individual
Migration rates change times (GLM analysis)	December 23, 2019 and March 1, 2020	
Global scaler GLM parameter	December 23, 2019	Median, IQR
Coefficients GLM predictors	Normal (0, 2)	Median 0, IQR
Binary indicators GLM predictors	0 - 1 equiprobable	BitFlipOperator.

Table 2.1: Caption

finish table and caption

Chapter 3

Results

The origin of SARS-CoV-2 in Europe While the first case in Europe was reported to ECDC by France authorities on January 25, the first European wave is suspected to have started earlier but been undetected, mistaken for other common repiratory diseases in winter. In our model, we assume the origin of SARS-CoV-2 in Europe to be an imported case from China to any of the European countries. We can select a single event of interest from the inferred epidemic trajectories, and summarise its time and location. To study the origin of SARS-CoV-2 in Europe, we analyze the time and destination of the first imported case and estimate the origin of the European epidemic on January 2020 (95% CCI), Figure 3.1 . The actual reported date to ECDC is in the tail of the 95% credible interval. As shown in Figure 3.2 A, we place the destination country of this first imported case in Italy or France, with x of probability each.

with final results

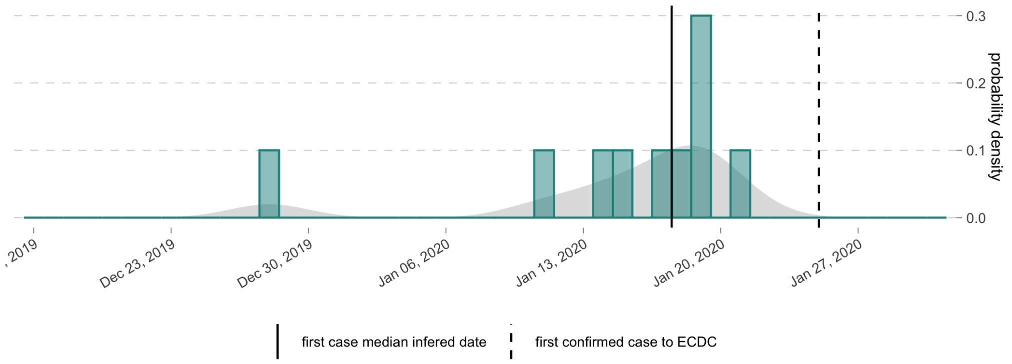

Figure 3.1: First case of SARS-CoV-2 in Europe. In the histogram, we count how many of the inferred epidemic trajectories started the European epidemic each day. From these counts, we estimate the probability density of the origin in Europe (in grey). The median date is drawn with the solid vertical line, and the date of the first reported European case to ECDC with a dashed vertical line.

Likewise, we analyze the first ten imported cases to Europe to estimate the main source of exported SARS-CoV-2 cases early on and their destination. China is the main source of these first introductions, supporting the hypothesis of the

European epidemic seeded by several and not just one imported cases from China, Figure 3.2 B. The role of Italy as the receptor of these first cases is highlighted even more looking at several cases. Early introductions in Spain have a very low probabilty, looking like Spain was not a key player in the origin of SARS-CoV-2 in Europe. Figure 3.2 C.

with final results

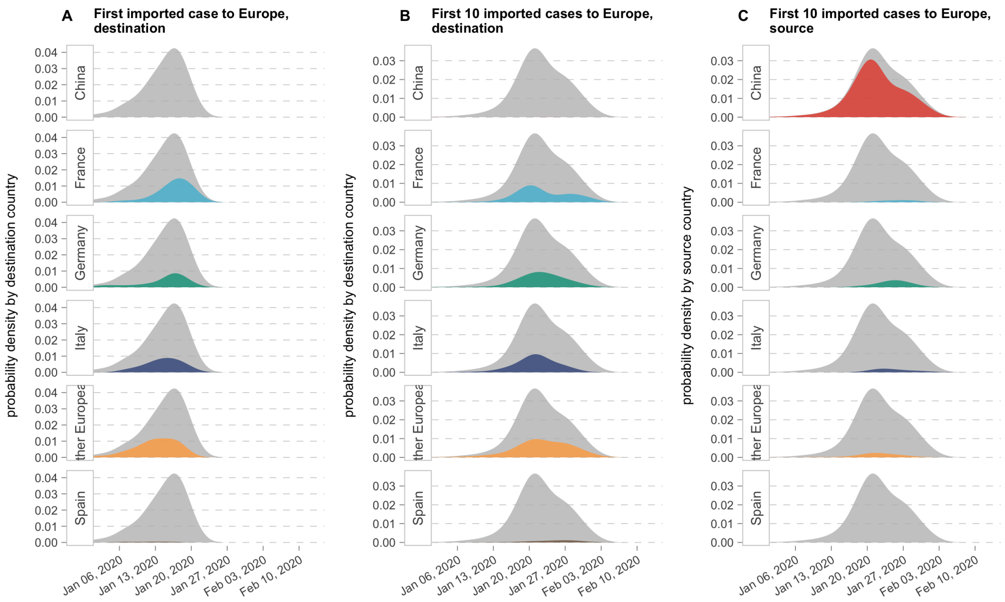

Figure 3.2: First imported cases of SARS-CoV-2 to Europe. **A** We estimate the probability distribution of the time for the first imported case to Europe (in grey). Each row highlights the probability of this first case going to the indicated country (in color). **B** In a similar way, we considered the time of the ten first imported cases into any of the European country, and plot the time probability distribution (in grey). In the rows, we plot the probability of each country being the destination country of any of the first 10 imported cases to Europe (in color). **C** Same probability distribution of the time for the 10 first imported cases to any European country (in grey) than in B, but in each row we highlight in color the probability of these cases coming from a specific country.

The early spread of SARS-CoV-2 among the European countries

Since in our model we consider the different European countries, we can also estimate from the infered epidemic trajectories the origin of the epidedmic in each particular country. In Figure we have the origin time for each country compared to the ddate of the first reporrted case to ECDC by that country. In average, we predict a delay in detection of x days for the Eurropean countries, out of the 95%CCI for France, Germany and Italy. Origin in Spain close to the official detection day. We estimate that the epidemic in any of the European countries was seeded by an imported case from China as seen in Figurem with some small prrobability for Germany and Italy of having orriginatetdd from a case from each other, and mostly half of the rpobability in the case of Spain from the epidemdic started with a case imported from one of the others Euroopean countries. Same than before, we can look to the first 10 imported cases to each country and now we see that the transmission between european countries adquire much moer importance, so once SARS-CoV-2 entered Europe, the transmission between european countrries were soon more important than the cases coming from China. This could relate also to the decrease in tthe trravel to and from China seen in figure.

Among the European countries, France and Germany have the earlier time of introduction, followed by Italy, Other European deme and Spain. We can compare this inferred dates of introduction with the day each country reported its first case to ECDC. In all cases, the reported day was later than the median of the inferred distribution, but it is inside the 95% interval.

The first case in each European country could have been imported from China or from other European country with an ongoing epidemic that started earlier. However, in our analysis we obtain a much higher support for China being the most probable source of the first case for all European demes, Figure **??** C.

Even if the first case in each European country came from China, the timing of the introductions in the European countries relative to each other probablu expanded across several days, defining and order of countries with the time of their first case.

We can analyze if this order of countries, defined by the time in which the first case occured, is shared among the majority of the inferred population trajectories, Figure **??**. We obtain that the first European country with a SARS-CoV-2 case was more likely France or Germany, followed by Italy and Other European deme. While Spain was more likely the last European country with a case among the ones included in the analysis.

Along the same lines, we could ask if this order of countries is mantained when instead of looking at the first case in the country we look into at first case exported (migration) from that country to other European country. In Figure **??** we observe a similar pattern to the first cases order, with Germany, France and Italy being the countries in the first positions in more than half of the inferred trajectories and never in the last position, and Spain as the country in last position in almost every epidemic trajectory.

Detection timing and response to the European epidemics

The time between the first case in the country, i.e. first incoming migration event and the first case from within-region transmission is of x (y-z) with similar values for all demes?. The time from the first incoming migration to the first outgoing migration is longer with a median of x (y-z). (This could be interesting

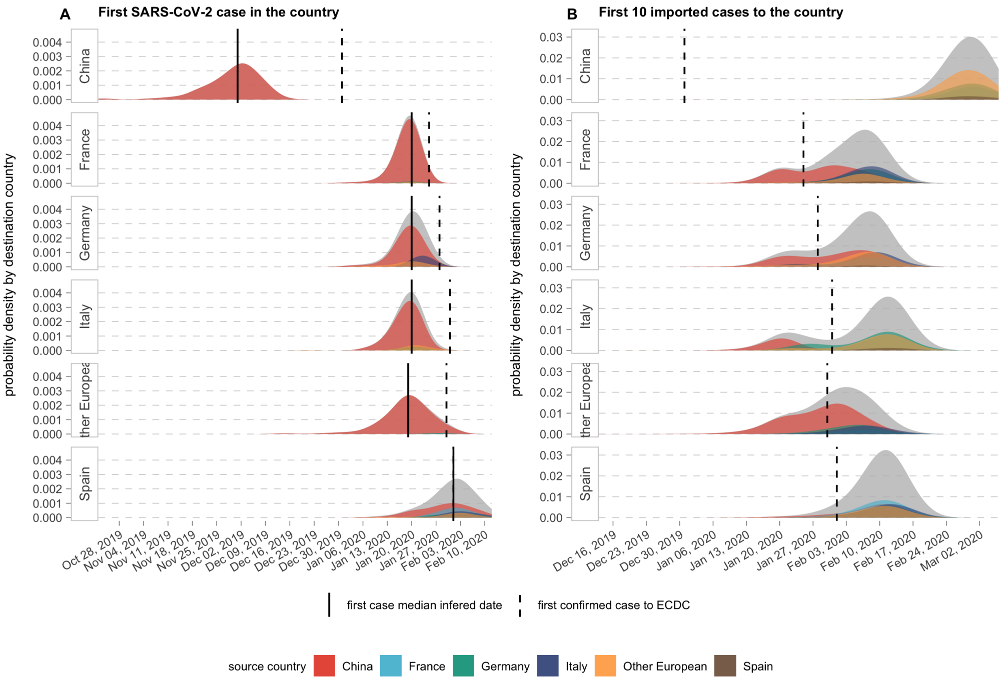

Figure 3.3: Country-level origin of SARS-CoV-2 epidemic. For China, we consider the origin of the epidemic as the first local case while for the European countries we select the first imported case into the country. **A** We estimate the probability distribution of the date for the first case in each of the locations included in the analysis (in grey). We highlight the conditional probability for the source of the first case in each country (in color). **B** The ten first imported cases into the country are selected, and we estimate the probability distribution of their occurrence. As in A, we show the conditional probability of the source country.

to say if we should focus or not the screening and testing capacities to detect incoming migrations or if when we have evidence of cases in the population we shoud follow a more general strategy to find cases in the population according to the model. Is it different for each country?)

(How well did the countries detecting the first cases, there were already within region transmision?) We compare the date of the first reported rate of each country with the date in which within-region transmission for that country started according to the model. We see some differences among countries, France and Germany had in all inferred epidemic trajectories ongoing within-region transmission when first cases werre reported, while x% of epidemic trajectories did not have had a within region transmission case when the first case in Spain was reported.

We can also compare the date of the first reported case with the date of the first outgoing migration from the country. (This could be interesting to say if a

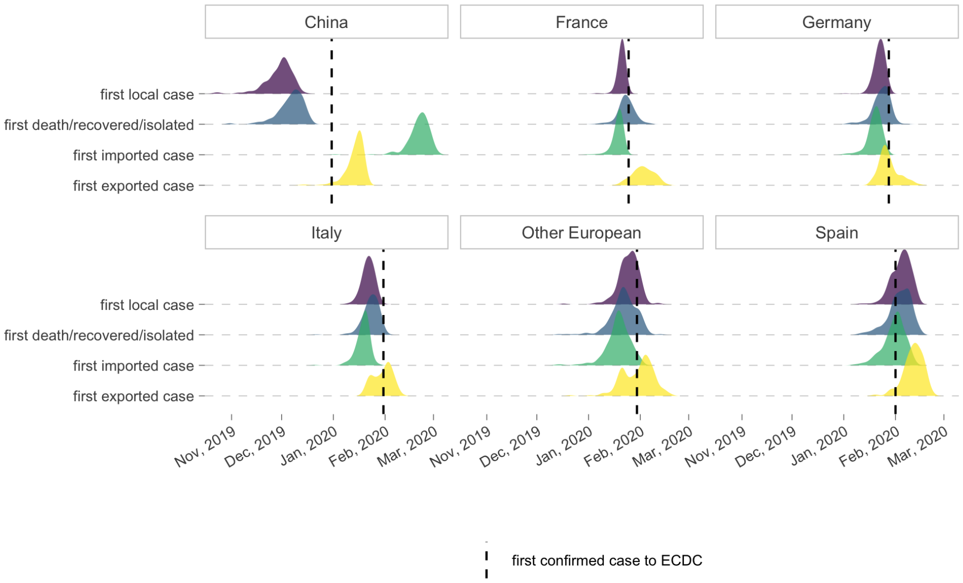

Figure 3.4: Timing of key first epidemic events. For each country, we select the key epidemic events: first local transmission case, first death, recovery or isolated case (same event category for the birth-death model), first imported and first exported case. We summarise the timing of these events from the inferred trajectories and obtain a probability distribution for each of them. The first detection of SARS-CoV-2 in each country reported to ECDC is shown with a dashed line.

extreme measure closing borders with the first case could be effective to impede transmission to other countries: percentage of trajectories whre transmission to Europe would have been avoided. For other countries we can look at how many migrations events could have been avoided (and how many not) if the country closed borders after first reported case according to the model. Not realistic measure, extreme case.)

Burden of SARS-CoV-2 infections in Europe The inferred epidemic trajectories contain the information about the total number of cases until 8th of March. For the European countries, we obtain an inferred number of total cases above the number of confirmed cases to ECDC, consistent with known limited test availability of the first wave and previous studies results [?] [Wu et al., 2020]. These cases counts correspond to x-y times higher than the number of reported cases. Italy is the country with the highest infered number of cases x, followed by Spain, France and Germany. The infered number of cases for China is below the reported number of cases.

Include values

limitations of the model, sequence information, partial outbreak dynamics

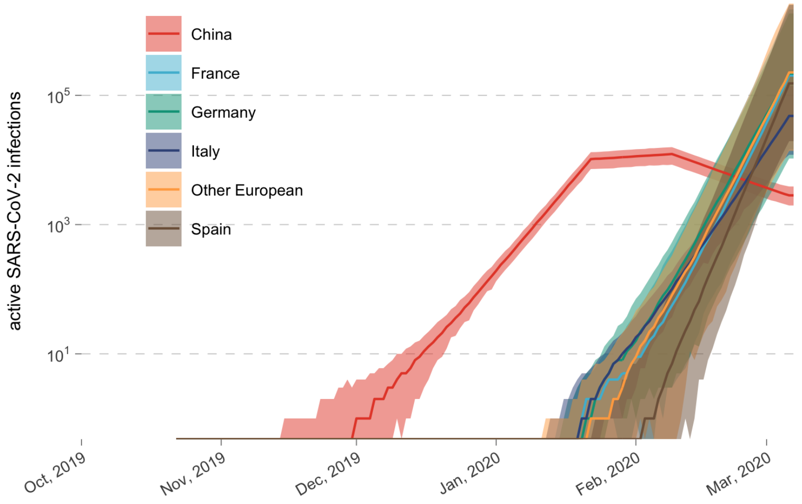

Figure 3.5: Active SARS-CoV-2 infections by country from the origin of the pandemic to March 8, 2020. We compute the median number and 95% credible interval of active SARS-CoV-2 infections for each analysis day by country. We plot the median value for the country with a coloured line and the 95% credible interval of active infections with a shaded area.

In Figure 3.5 and/or 3.6 we compare the total number of inferred cases by day to the total cumulative number of cases that have been reported to ECDC that same day. Our inferred case counts follow a exponential growth earlier in time than the reported curve and with higher number of cases, being the difference bigger for later times in the epidemic.

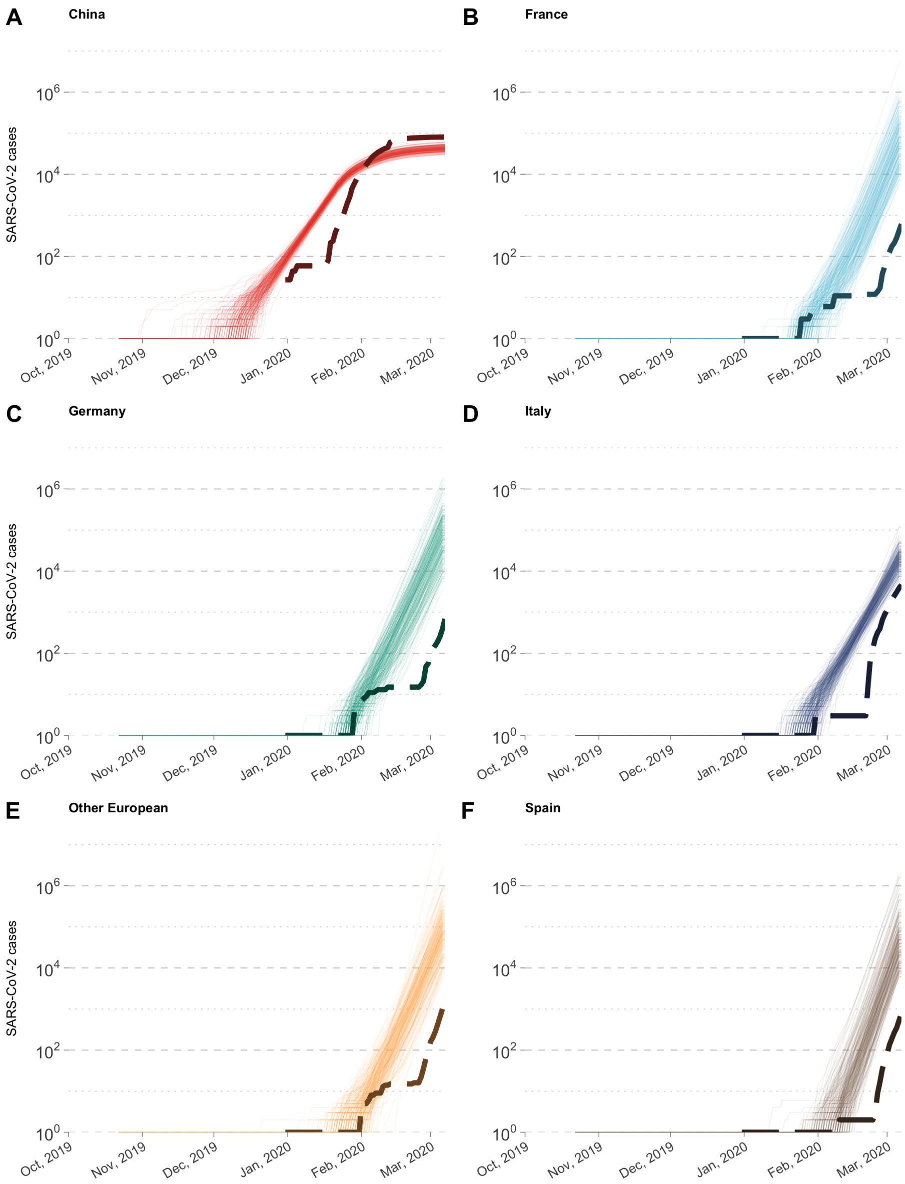

Figure 3.6: SARS-CoV-2 inferred cases and ECDC confirmed cases by country over time. To compare with the official ECDC confirmed cases in a analogous time frame, we consider the no longer infectious cases inferred by the model that appear after a mean period of ten days post-infection. Each line represent the cumulative number of inferred cases in log scale for each of the 500 analysed epidemic trajectories with time in the x-axis. The ECDC confirmed case counts are shown with dashed lines.

Why is till March 20?

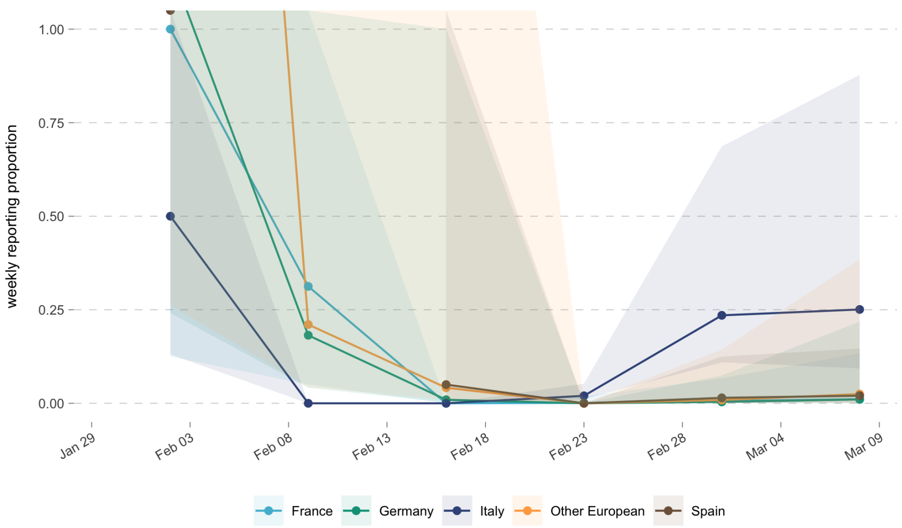

Figure 3.7: Weekly reporting proportion of new SARS-CoV-2 cases. We compute the number of new confirmed cases by week reported to ECDC and divide it by the median number of inferred cases (no longer infectious cases) in the same week. For each week, the point represents the median reporting proportion, and the shaded area the 95% credible interval.

We can think of a reporting rate as the number of reported cases relative to the total number of cases inferred by the model. This reporting rate decreases with time for all European countries, except for Italy that increases again around March, Figure 3.7.

Imported cases vs local transmission

TODO

From the epidemic trajectories, we can extract the information about how many cases are within-region transmission and how many are migrations from other countries. The cumulative number of transmission events and migrations, represented in Figure **??** increases exponentially over time. An incomming migration into every European deme happens always before within-region transmission, seeding the epidemic. Within-region transmision accounts for most of the cases in the countries from late january onwards (when first cases were being reported in Europe).

get the proportion of within-region transmission and incoming migrations?

International transmission patterns Figures **??** and 3.9.

TODO

Similar information in both plots, but in the chord plots instead of a daily evolution the time period is split on three (as in the GLM analysis). Chord plots

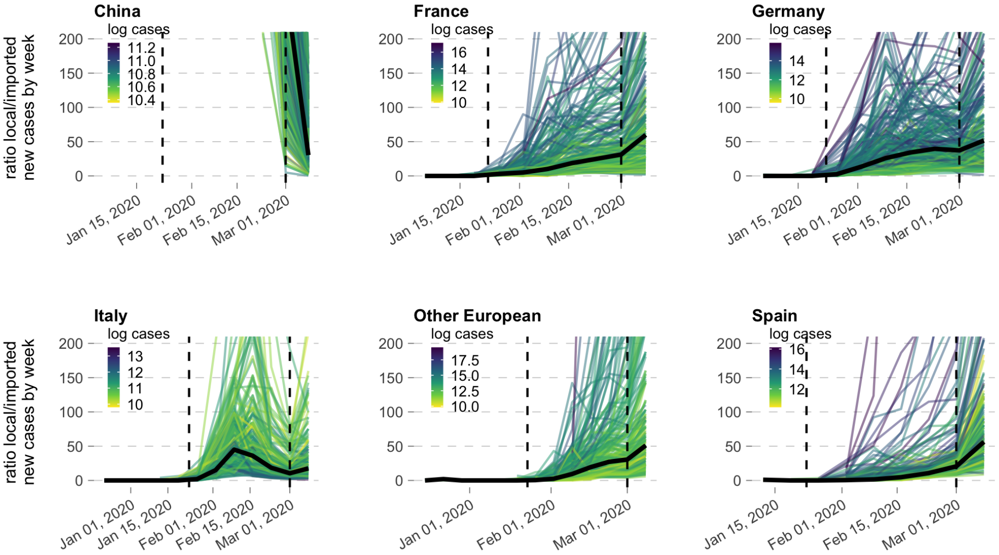

Figure 3.8: Evolution of the ratio of local transmissions by imported cases for the SARS-CoV-2 early epidemic in China, France, Germany, Italy, Spain and other European countries. We compute for each day the ratio of new local infections by new imported cases for each of the analysed epidemic trajectories. Each line depicts the evolution of this ratio in one epidemic trajectory, coloured by the total number of estimated infections by the model. The black line is the median value for the ratio local/imported cases.

are nicer and easier to understand I think, but barplots shows the great detail of the results of the model. Another advantage of the chrod plot is that the it shows mean absolute values and not only relative values.

Hubei-China is the majority source of migrations for all countries till February and then some patterns emerge (we expect more interesting results with the added info in GLM analysis).

Epidemiological parameters

TODO

Not sure if it is necessary. But maybe include, briefly, the values of estimate R0, migration and sampling rates? Maybe a table in the appendix? Or the plots with the priors.

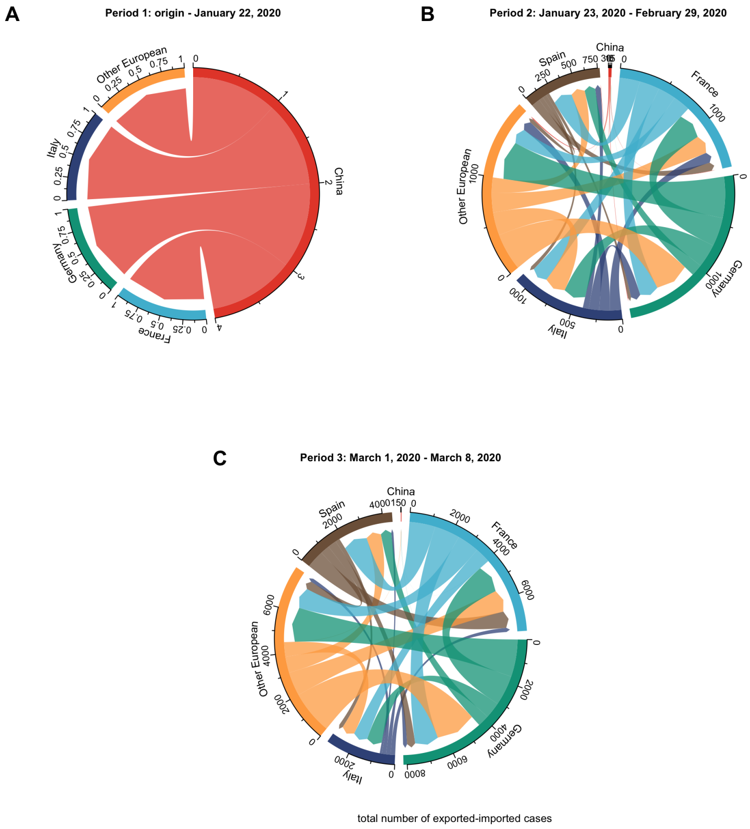

Figure 3.9: International imported case dynamics by analysis constant-migration rate period. One chord diagram is drawn for each of the main periods of the analysis defined by the change times of the migration rates in the model. Each flux represents the infections exported from one country to another during the full period. The median value of infections for each flux is indicated in the plot.

Number of particles

TODO, or into the discussion??

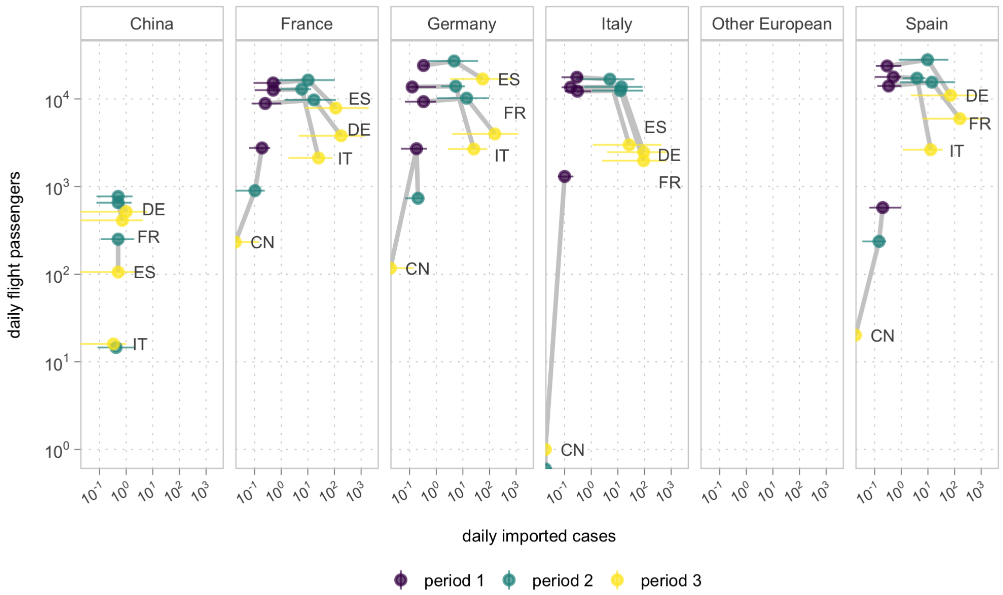

Figure 3.10: Evolution of daily imported cases and flight passengers by constant migration rate period of the analysis. In each subfigure, we take the median number of daily imported cases by period in the country (x-axis) and the average of daily flight passengers from Eurostat for each source country.

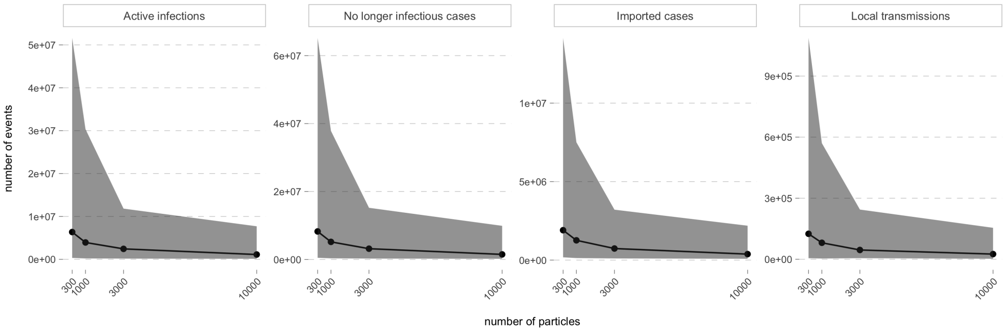

Figure 3.11: Impact of the number of particles on the stochastic simulation of epidemic trajectories. We simulate epidemic trajectories with a tau-leaping algorithm from our infered parameters of the birth-death model and the phylogenetic trees. We use 300, 1000, 3000 and 10000 particles, i.e. number of trajectories simulated for each set of estimated parameters. In the y-axis, We plot the median number of total active infections, no longer infectious cases, imported cases and local transmissions, and the 95% credible interval (shaded area) computed from the epidemic trajectories events and the number of particles in the x-axis.

Chapter 4

Discussion

TODO discussion

- Case counts discussion and reporting rates, second wave in Europe. Other estimations in other studies.
- First introduction other studies.

  Once the first european imported cases, everything happen quickly. If we look at the evolution of thte epidedmic in China, longer ttime between first case andd exportedd cases. Greater connectivity among Europe spread faster the epidemic.

  Origin of Spain very close to official reporting date, while now the officiall case counts in Spain have been updadted to included cases detected early on by the surveillance system. This coulld be highly dependant on our samples and the dates of the sequences.
- Migration patterns other studies.
- Discuss assumptions of the model, caveats and possible improvements.
- Number of particles

Chapter 5

Conclusion

TODO conclusion

Bibliography

[BDMM-Prime, 2020] BDMM-Prime (2020). Github - tgvaughan/bdmm-prime: Modified implementation of the bdmm multi-type birth-death model for beast 2.

[Billah et al., 2020] Billah, M. A., Miah, M. M., and Khan, M. N. (2020). Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence. *PLOS ONE*, 15:e0242128.

[Boskova et al., 2014] Boskova, V., Bonhoeffer, S., and Stadler, T. (2014). Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. *PLOS Computational Biology*, 10:e1003913.

[Bouckaert et al., 2019] Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N. D., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Plessis, L. D., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C. H., Xie, D., Zhang, C., Stadler, T., and Drummond, A. J. (2019). Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS Computational Biology*, 15.

[COVID-19-Re, 2020] COVID-19-Re (2020). Covid-19 re.

[Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376.

[Gillespie, 1977] Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions.

[Gillespie, 2000] Gillespie, D. T. (2000). Approximate accelerated stochastic simulation of chemically reacting systems.

[Huisman et al., 2020] Huisman, J. S., Scire, J., Angst, D. C., Neher, R. A., Bonhoeffer, S., and Stadler, T. (2020). Estimation and worldwide monitoring of the effective reproductive number of sars-cov-2. *medRxiv*, page 2020.11.26.20239368.

[Kuhnert et al., 2016] Kuhnert, D., Stadler, T., Vaughan, T. G., and Drummond, A. J. (2016). Phylodynamics with migration: A computational framework to quantify population structure from genomic data. *Mol Biol Evol*, 33:2102–2116. Main reference for BDMM model<br/><br/>.

[Köster and Rahmann, 2012] Köster, J. and Rahmann, S. (2012). Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, 28:2520–2522.

[Lemey et al., 2020] Lemey, P., Hong, S. L., Hill, V., Baele, G., Poletto, C., Colizza, V., Áine O'Toole, McCrone, J. T., Andersen, K. G., Worobey, M., Nelson, M. I., Rambaut, A., and Suchard, M. A. (2020). Accommodating individual travel history and unsampled diversity in bayesian phylogeographic inference of sars-cov-2. *Nature Communications*, 11:5110.

[Lemey et al., 2014] Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C. A., Smith, D. J., Pybus, O. G., Brockmann, D., and Suchard, M. A. (2014). Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza h3n2. *PLoS Pathog*, 10:e1003932. Lemey, Philippe Rambaut, Andrew Bedford, Trevor Faria, Nuno Bielejec, Filip Baele, Guy Russell, Colin A Smith, Derek J Pybus, Oliver G Brockmann, Dirk Suchard, Marc A eng R01 AI107034/AI/NIAID NIH HHS/ R01 HG006139/HG/NHGRI NIH HHS/ 095831/Wellcome Trust/United Kingdom 092807/Wellcome Trust/United Kingdom HHSN266200700010C/PHS HHS/ DP1 OD000490-01/OD/NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. PLoS Pathog. 2014 Feb 20;10(2):e1003932. doi: 10.1371/journal.ppat.1003932. eCollection 2014 Feb.

[Maddison et al., 2007] Maddison, W. P., Midford, P. E., and Otto, S. P. (2007). Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, 56:701–710.

[Nadeau et al.,] Nadeau, S. A., Vaughan, T. G., Scire, J., Huisman, J. S., and Stadler, T. The origin and early spread of sars-cov-2 in europe.

[Nextstrain-ncov, 2020] Nextstrain-ncov (2020). Github - nextstrain/ncov: Nextstrain build for novel coronavirus sars-cov-2.

[Pan et al., 2020] Pan, A., Liu, L., Wang, C., Guo, H., Hao, X., Wang, Q., Huang, J., He, N., Yu, H., Lin, X., Wei, S., and Wu, T. (2020). Association of public health interventions with the epidemiology of the covid-19 outbreak in wuhan, china. *JAMA*, 323:1915.

[Park et al., 2020] Park, S. W., Bolker, B. M., Champredon, D., Earn, D. J. D., Li, M., Weitz, J. S., Grenfell, B. T., and Dushoff, J. (2020). Reconciling early-outbreak estimates of the basic reproductive number and its uncertainty: framework and applications to the novel coronavirus (sars-cov-2) outbreak.

[Scire et al., 2020] Scire, J., Barido-Sottani, J., Kühnert, D., Vaughan, T. G., and Stadler, T. (2020). Improved multi-type birth-death phylodynamic inference in beast 2. *bioRxiv*.

[Shu and McCauley, 2017] Shu, Y. and McCauley, J. (2017). Gisaid: Global initiative on sharing all influenza data – from vision to reality.

[Stadler et al., 2012] Stadler, T., Kouyos, R., VonWy, V., Yerly, S., Böni, J., Bürgisser, P., Klimkait, T., Joos, B., Rieder, P., Xie, D., Günthard, H. F., Drummond, A. J., and Bonhoeffer, S. (2012). Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution*, 29:347–357.

[Stadler et al., 2013] Stadler, T., Kühnert, D., Bonhoeffer, S., and Drummond, A. J. (2013). Birth-death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). *Proceedings of the National Academy of Sciences of the United States of America*, 110:228–233.

[Vaughan, 2021] Vaughan, T. G. (2021). Bdmm-prime work in progress.

[Vaughan et al., 2014] Vaughan, T. G., Kühnert, D., Popinga, A., Welch, D., and Drummond, A. J. (2014). Efficient bayesian inference under the structured coalescent. *Bioinformatics*, 30:2272–2279.

[Vaughan et al., 2019] Vaughan, T. G., Leventhal, G. E., Rasmussen, D. A., Drummond, A. J., Welch, D., and Stadler, T. (2019). Estimating epidemic incidence and prevalence from genomic data. *Mol Biol Evol*, 36:1804–1816. Vaughan, Timothy G Leventhal, Gabriel E Rasmussen, David A Drummond, Alexei J Welch, David Stadler, Tanja eng Evaluation Study Research Support, Non-U.S. Gov't Mol Biol Evol. 2019 Aug 1;36(8):1804-1816. doi: 10.1093/molbev/msz106. Particle filtering paper. Nicely written. Compartmental models explained. Re-read to understad method.

[Wu et al., 2020] Wu, J. T., Leung, K., and Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. *The Lancet*, 395:689–697.

[Xiao et al., 2020] Xiao, Y., Tang, B., Wu, J., Cheke, R. A., and Tang, S. (2020). Linking key intervention timing to rapid decline of the covid-19 effective reproductive number to quantify lessons from mainland china. *International Journal of Infectious Diseases*, 97:296–298.

Appendix A
Data

Appendix B
Priors

Appendix C

Supplementary Figures