Tutorial for the R package `seraphim` 1.0

Estimating lineage dispersal statistics

Simon Dellicour

December 26, 2024

The present tutorial describes how to use the R package "`seraphim`" [1, 2] to exploit phylogenetically informed movement data obtained through continuous phylogeographic reconstruction [3] in order to characterise the dispersal dynamics of the West Nile virus (WNV) lineages in North America [4]. In particular, we here use functions of the package to estimate several lineage dispersal statistics. See also the package manual for further detail on its different functions.

The R package "`seraphim`" is hosted on GitHub (`https://github.com/sdellicour/seraphim`) and the first step is to install it using the "install_github()" function of the "'`devtools`" package:

```
> install.packages("devtools"); library(devtools)
> install_github("sdellicour/seraphim/unix_OS") # for Unix systems
> install_github("sdellicour/seraphim/windows") # for Windows systems
```

Note that the installation of "`seraphim`" requires the preliminary installation of the following R packages: "ape", "doMC" (only available for Unix systems), "fields", "gdistance", "HDInterval", "ks", "phytools", "raster", "RColorBrewer", "rgeos", and "vegan". Once installed, the package has to be loaded as follows:

```
> library(seraphim)
```

This tutorial requires the following WNV example files also available on the GitHub repository of the package (`https://github.com/sdellicour/seraphim/tree/master/tutorials`): "WNV_gamma.trees", a file containing 100 annotated phylogenetic trees sampled from the post-burn-in posterior distribution of trees inferred for the WNV dataset of Pybus and colleagues [4] using the continuous phylogeographic approach developed by Lemey and colleagues [3].

This is out of the scope of the present tutorial but a detailed procedure on how to prepare and conduct a continuous phylogeographic reconstruction using the relaxed random walk (RRW) diffusion model [3] implemented in the software package BEAST 1.10 [5] is available on the BEAST community website (`https://beast.community/workshop_continuous_diffusion_yfv`).

Step 1: extracting spatio-temporal information in trees

The first step is to extract the spatio-temporal information embedded in annotated phylogenetic trees sampled from the posterior distribution of a continuous phylogeographic analysis. The tree file "WNV_gamma.trees" contains 100 trees sampled in the post-burn-in posterior distribution of trees. We will here use the "treeExtractions()" function to extract the information embedded in these 100 posterior trees. The "treeExtractions()" function first requires the definition of the following parameters: "localTreesDirectory" (name of the directory to create and where spatio-temporal information contained in each tree will be saved), "allTrees" (name of the tree file), "burnIn" (number of trees to discard as burn-in, i.e. a number defining a series of first trees in which no tree will be sampled – has to be set to "0" as burn-in trees are, in the present case, already discarded), "randomSampling" (boolean variable specifying if the trees have to be randomly sampled or sampled at the largest possible regular interval – not relevant in the present case as the trees have already been sampled), "nberOfTreesToSample" (number of trees to sample), "mostRecentSamplingDatum" (most recent sampling date in a decimal format) and "coordinateAttributeName" (attribute name used to indicate the geographic coordinates within the tree file).

```
> localTreesDirectory = "Extracted_trees"
> allTrees = scan(file="WNV_gamma.trees", what="", sep="\n", quiet=T)
> burnIn = 0
> randomSampling = FALSE
> nberOfTreesToSample = 100
> mostRecentSamplingDatum = 2007.63
> coordinateAttributeNam = "location"
```

Once all these parameters have been specifying, the "treeExtractions()" function can be launched as follows:

```
> treeExtractions(localTreesDirectory, allTrees, burnIn, randomSampling,
nberOfTreesToSample, mostRecentSamplingDatum, coordinateAttributeName)
```

Note that in the R script of this tutorial, we also describe how to use the "postTreeExtractions()" function of the package to extract the spatio-temporal information embedded in posterior trees inferred by such a Bayesian continuous phylogeographic analysis (and the use of this new function is actually recommended because, depending on the continuous phylogeographic inference settings, the "treeExtractions()" function can sometimes fail to perform the extractions, returning an error message).

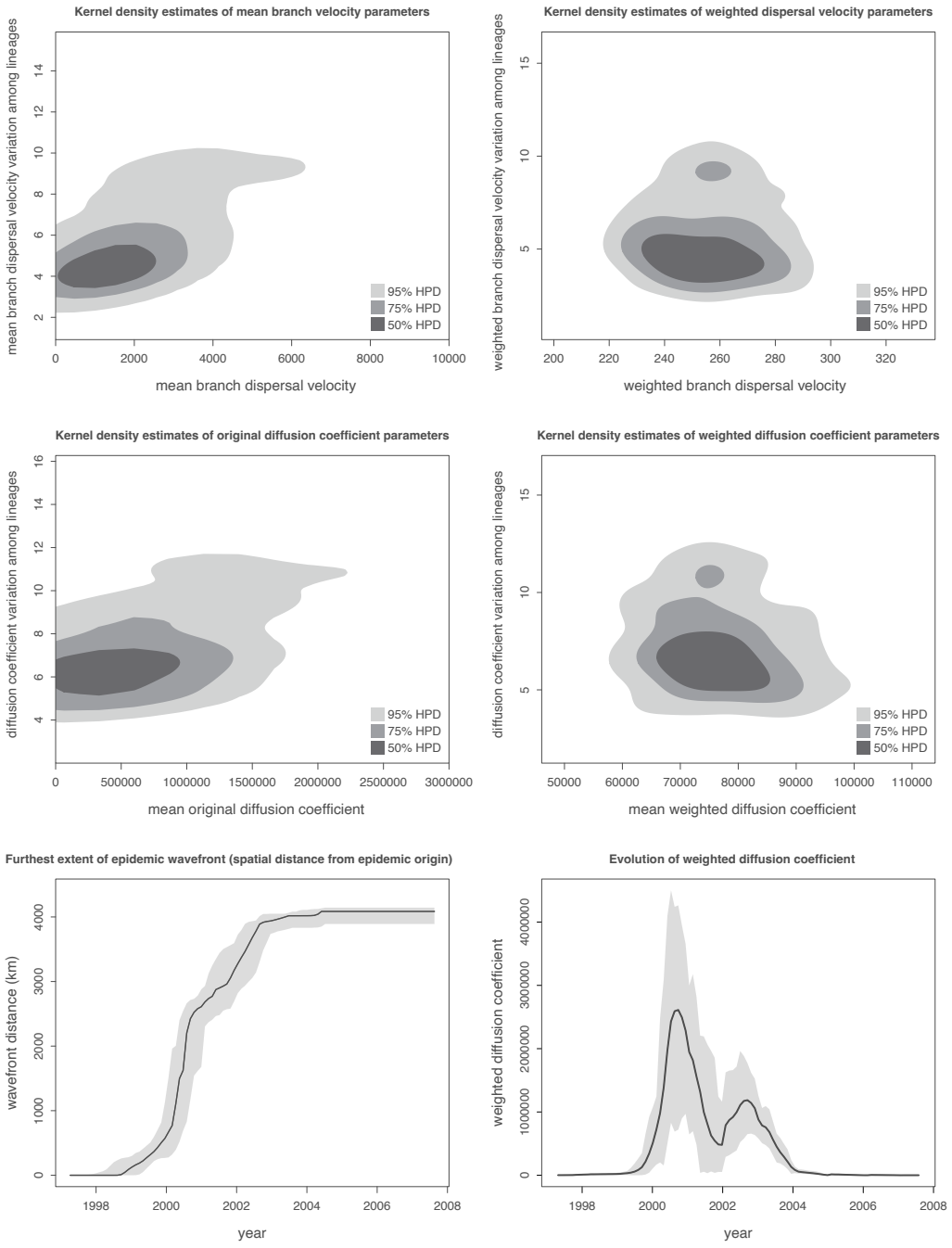

Figure 1: estimated dispersal statistics. For the four first graphs, the three contours show, in shades of decreasing darkness, the 50%, 75%, and 95% HPD regions via kernel density estimation. For the two last graphs, grey area corresponds to the 95% HPD interval of the estimated wavefront position and weighted diffusion coefficient estimates, respectively.

Step 2: estimation of several dispersal statistics

The second step of this tutorial consists in using the spatio-temporal information extracted from posterior trees to estimate a series of dispersal statistics using the "spreadStatistics()" function. So far, estimations of five statistics are implemented: the mean branch dispersal velocity v_{branch}, the weighted branch dispersal velocity $v_{weighted}$, the original diffusion coefficient $D_{original}$ defined by Pybus *et al.* [4], the weighted coefficient $D_{weighted}$ defined by Trovão *et al.* [6], as well as the isolation-by-distance (IBD) signal measured either as (i) the Spearman correlation (r_S) between the patristic and great-circle geographic distances computed for each pair of tip nodes, (ii) the Pearson correlation (r_P1) between the patristic and great-circle geographic distances computed for each pair of tip nodes, or (iii) the Pearson correlation (r_P2) between the patristic and the log-transformed great-circle geographic distances computed for each pair of tip nodes. Of note, the latter IBD signal statistics do not rely on a continuous phylogeographic reconstruction but are thus directly estimated from the great-circle distance between tip nodes and the patristic distances computed from the tree topology.

If we consider n phylogeny branches, the four first statistics are defined as follows:

$$v_{branch} = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i}{t_i} \qquad v_{weighted} = \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} t_i}$$

$$D_{original} = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i^4}{t_i^2} \qquad D_{weighted} = \frac{\sum_{i=1}^{n} d_i^4}{\sum_{i=1}^{n} t_i^2}$$

where d_i and t_i are, respectively, the geographic distance travelled (great-circle distance measured in kilometers) and the time elapsed (usually in years) on each phylogeny branch. For a given tree, branches with short duration will have respectively less of an impact on $v_{weighted}$ and $D_{weighted}$ than on v_{branch} and $D_{original}$, and therefore also on the resulting variance among $v_{weighted}$ and $D_{weighted}$ values across all trees. Compared to v_{branch} and $D_{original}$, $v_{weighted}$ and $D_{weighted}$ can respectively allow a better discrimination among epidemics with different diffusivity because it is associated with a smaller variance [7].

It is worth noting that, in a recent study [8], we used simulations to evaluate the robustness of these different dispersal metrics to the sampling effort. Our results reveal that the diffusion coefficient and isolation-by-distance signal metrics appear to be the most robust to the number of samples considered for the phylogeographic reconstruction. We thus recommend using the diffusion coefficient metrics rather than the branch dispersal velocity ones to evaluate and compare the dispersal capacity of lineages whose dispersal history has been inferred through continuous phylogeographic inference.

In addition to these lineage dispersal statistics, the "spreadStatistics()" function also estimates the evolution of two maximal wavefront distances, as well as the evolution of the diffusion coefficient estimates through time. The function will both estimate values and generate/save graphs. It requires the user to specify (i) the directory in which extracted spatio-temporal information has been saved (see above), (ii) the number of extraction of files to use (this number cannot be higher that the number of extractions performed in the previous step), (iii) the number of distinct time slices ("timeSlices") that will be used

to generate the maximal wavefront distance evolution plots, (iv) the "onlyTipBranches" boolean variable indicating if statistics estimations have to be based on the tip branches only, (v) the "showingPlots" boolean variable specifying if the different plots have to be displayed or not, (vi) the "outputName" string (prefix) to give to the different output files, (vii) the number of cores ("nberOfCores") to use for the computations, and (viii) the sliding window, in units of time, that will be used to generate the diffusion coefficient evolution plots (optional).

```
> nberOfExtractionFiles = 100
> timeSlices = 100
> onlyTipBranches = FALSE
> showingPlots = FALSE
> outputName = "WNV"
> nberOfCores = 1
> slidingWindow = 1
> spreadStatistics(localTreesDirectory, nberOfExtractionFiles, timeSlices,
onlyTipBranches, showingPlots, outputName, nberOfCores, slidingWindow)

Median value of mean branch velocity = 1522.4
95% credible region = [696.8, 5846.0]
Median value of weighted dispersal velocity = 255.7
95% HPD = [226.5, 286.5]
Median value of original diffusion coefficient (Pybus et al. 2012) = 413018
95% HPD = [133517, 2102014]
Median value of weighted diffusion coefficient (Trovao et al. 2015) = 75977
95% HPD = [62776, 91542]
Median value of the isolation-by-distance (IBD) signal (rS) = 0.204
95% HPD = [0.148, 0.262]
Median value of the isolation-by-distance (IBD) signal (rP #1) = 0.221
95% HPD = [0.162, 0.275]
Median value of the isolation-by-distance (IBD) signal (rP #2) = 0.274
95% HPD = [0.242, 0.319]
```

As displayed in Figure 1, the function will also generate and save six different graphs: the kernel density estimates of the mean branch velocity parameters (branch dispersal velocity variation among branches *vs* mean branch dispersal velocity), the kernel density estimates of the weighted branch dispersal velocity parameters (branch dispersal velocity variation among branches *vs* weighted branch dispersal velocity), the kernel density estimates of original diffusion coefficient parameters (diffusion coefficient variation among branches *vs* original diffusion coefficient), the kernel density estimates of weighted diffusion coefficient parameters (diffusion coefficient variation among branches *vs* weighted diffusion coefficient), as well as the evolution of the maximal spatial and patristic wavefront distances from epidemic origin.

The maximal *spatial* wavefront distance corresponds to the straight-line distance (i.e. "as the crow flies") from to the estimated location of the root, and the maximal *patristic* wavefront distance corresponds to the distance computed as the sum of geographical distances associated with each branch connecting a given node to the root. Note that the latter one is now different from the maximal *patristic* wavefront distance as estimated

in [1]. Indeed, in the previous version of the package that was used for these studies, the maximal *patristic* wavefront distance was computed as the maximal *patristic* distance from any node location to the root at a given point in time. Now, the maximal *patristic* wavefront distance is defined as the *patristic* distance from the root to the node associated with the highest *spatial* distance from the root location at a given point in time. While the previous implementation still represents an interesting metric, it thus corresponds to another measure and we believe that our more recent implementation makes more sense for the study of actual wavefront evolution. In summary, in the current implementation, both maximal wavefront distances are related to the furthest extent of the wavefront but while the first one is computed as the *spatial* distance from the root location, the second one is computed as the *patristic* from the root location.

References

[1] Dellicour S, Rose R, Pybus OG (2016a). Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinformatics* 17: 82.

[2] Dellicour S, Rose R, Faria N, Lemey P, Pybus OG (2016b). SERAPHIM: studying environmental rasters and phylogenetically-informed movements. *Bioinformatics* 32 (20): 3204-3206.

[3] Lemey P, Rambaut A, Welch JJ, Suchard MA (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology & Evolution* 27: 1877-1885.

[4] Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, Crawford FW, Gray RR, Arinaminpathy N, Stramer SL, Busch MP, Delwart EL (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *PNAS* 109: 15066-15071.

[5] Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* 4: vey016.

[6] Trovão NS, Suchard MA, Baele G, Gilbert M, Lemey P (2015). Bayesian inference reveals host-specific contributions to the epidemic expansion of Influenza A H5N1. *Molecular Biology and Evolution* 32 (12): 3264-3275.

[7] Dellicour S, Rose R, Faria NR, Vieira LFP, Bourhy H, Gilbert M, Lemey P, Pybus OG (2017). Using viral gene sequences to compare and explain the heterogeneous spatial dynamics of virus epidemics. *Molecular Biology & Evolution* 34: 2563-2571.

[8] Dellicour S, Bastide P, Rocu P, Fargette D, Hardy OJ, Suchard MA, Guindon S, Lemey P (2024). How fast are viruses spreading in the wild? *PLoS Biology* 22: e3002914.