Tutorial for the R package `seraphim` 1.0

Estimating dispersal/epidemiological statistics

Simon Dellicour

July 7, 2019

The present tutorial describes how to use the package "`seraphim`" (for "studying environmental rasters and phylogenetic informed movements" [1, 2]) to characterise the dispersal dynamics of the West-Nile virus (WNV) lineages in North America [3]. In particular, we here use functions of the package in order to estimate several dispersal/epidemiological statistics. The first step is to install "`seraphim`" 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
```

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

```
> library(seraphim)
```

This tutorial requires the WNV example file also available on the GitHub repository (`https://github.com/sdellicour/seraphim/tree/master/tutorials`): "WNV_gamma.trees", a file containing 100 phylogenetic trees sampled from the post-burn-in posterior distribution of trees inferred for the WNV dataset using the method of Lemey *et al.* [4] in Pybus *et al.* [3].

Step 1: extracting spatio-temporal information in trees

The first step is to extract the spatio-temporal information contained in phylogenetic trees. This kind of trees have to be in a Newick format and can, for instance, be inferred by the continuous phylogeographic method implemented in BEAST [4] with a "gamma"

relaxed random walk model. 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 contained 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 datum 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=TRUE)
> 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)
```

Step 2: estimation of several dispersal statistics

The second step of this tutorial consists in using the extracted information to estimate a series of epidemiological statistics using the "spreadStatistics" function. So far, estimations of four 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.* [3], and the weighted coefficient $D_{weighted}$ as defined Trovão *et al.* [5]. If we consider n phylogeny branches, these four 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 in km) and the time elapsed (usually in years) on each phylogeny branch.

In addition to these statistics, the function also estimates the evolution of two maximal wavefront distances, as well as the evolution of the dispersal velocity 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 dispersal velocity evolution plot (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 = [728.7, 6324.7]
Median value of weighted dispersal velocity = 255.7
95% HPD = [227.3, 286.7]
Median value of original diffusion coefficient (Pybus et al. 2012) = 413018
95% HPD = [209403, 2409666]
Median value of weighted diffusion coefficient (Trovao et al. 2015) = 75977.1
95% HPD = [63542.1, 92286.7]
```

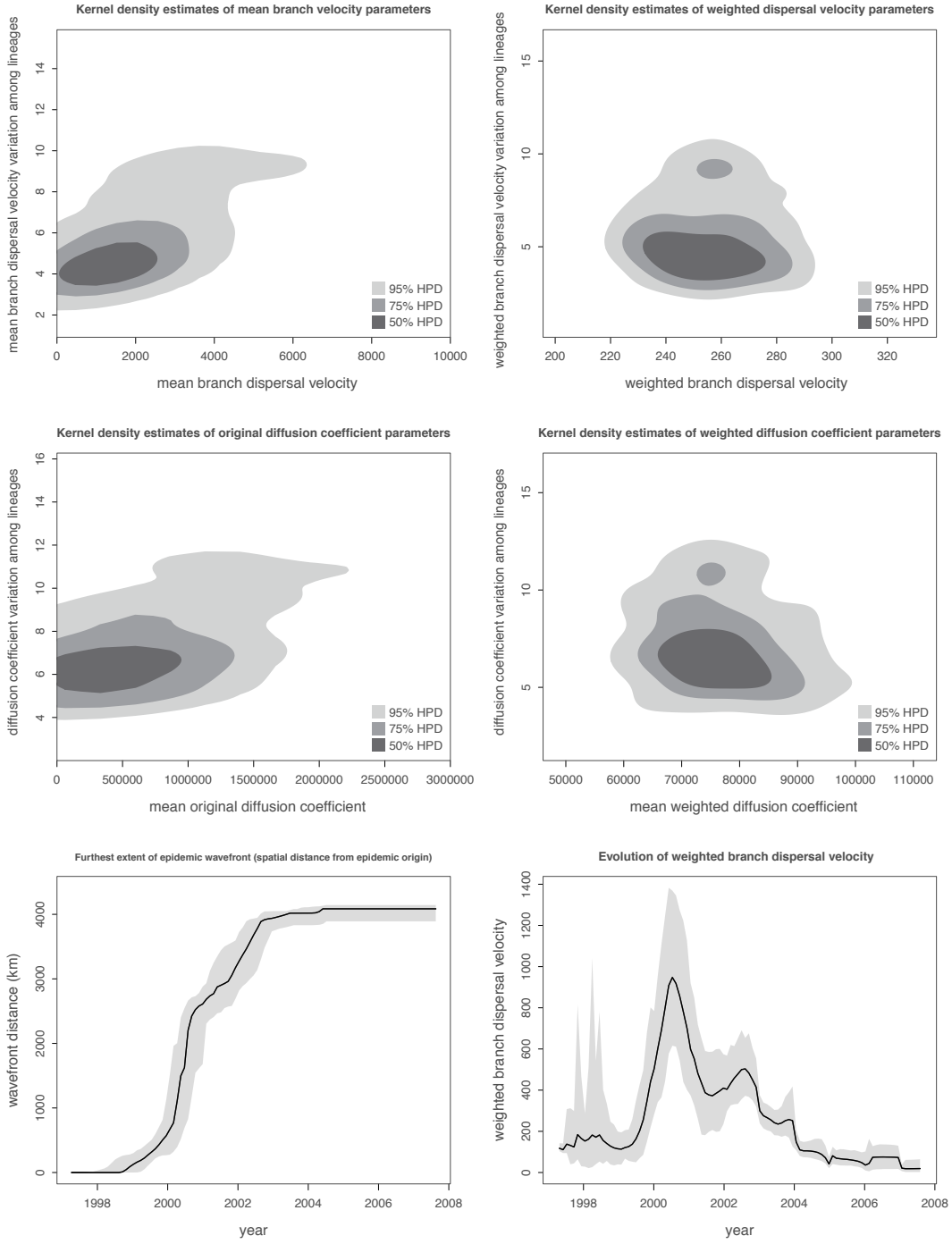

Figure 1: estimated dispersal/epidemiological statistics. For the four first graphs, the three contours show, in shades of decreasing darkness, the 50%, 75%, and 95% HPD regions via kernel density estimation. For the two last graphs, grey area corresponds to 95% credible regions of the estimated wavefront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

As displayed in Figure 1, the function will also generate and save six different graphs: the kernel density estimates of the mean branch dispersal velocity parameters (branch dispersal velocity variation among branches vs. mean branch dispersal velocity), the kernel density estimates of the weighted diffusion 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] and [2]. 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] 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.

[4] 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.

[5] 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.