VEME workshop 2019, Hong Kong
Tutorial for the R package `seraphim` 1.0

Using continuous phylogeographic analysis to investigate the impact of environmental factors on the dispersal velocity of viral lineages

Simon Dellicour

July 7, 2019

The present tutorial describes how to use the package "`seraphim`" [1, 2] to exploit phylogenetically informed movement data in order to investigate the association between a particular environmental factor (an "elevation" raster) and the dispersal velocity of rabies virus (RABV) spread in North American raccoon populations [1, 3]. See also the package manual for further details. 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 RABV example files: "RABV_gamma.trees", a file containing the phylogenetic trees inferred for the RABV dataset using the method of Lemey *et al.* [4], and "Elevation_raster.asc", the environmental raster "elevation".

Step 1: extracting spatio-temporal information from 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]. The tree file "RABV_gamma.trees" contains 5001 trees sampled by the MCMC chain. We will here use the "treeExtractions" function to extract the spatio-temporal information contained in 100 post-burin-in trees randomly sampled in this posterior distribution. 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" (all the posterior trees), "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), "randomSampling" (boolean variable specifying if the trees have to be randomly sampled or sampled at the largest possible regular interval), "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="RABV_gamma.trees", what="", sep="\n", quiet=TRUE)
> burnIn = 1001
> randomSampling = FALSE
> nberOfTreesToSample = 100
> mostRecentSamplingDatum = 2004.7
> coordinateAttributeName = "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: estimating dispersal/epidemiological statistics

The second optional 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 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.* [5], and the weighted coefficient $D_{weighted}$ as defined by Trovão *et al.* [6]. 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 measured in kilometers) 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. The function will both estimate values and generate/save graphs. It simply 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, the "showingPlots" boolean variable specifying if the different graphs should be displayed or not (in addition to be saved by the function), and (vi) the "outputName" string (prefix) to give to the different output files.

```
> nberOfExtractionFiles = 100
> timeSlices = 100
> onlyTipBranches = FALSE
> showingPlots = FALSE
> outputName = "RABV_raccoon"
```

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

```
> spreadStatistics(localTreesDirectory, nberOfExtractionFiles,
timeSlices, onlyTipBranches, showingPlots, outputName)
```

In addition to text files summarising the different results, the function will generate and save six different graphs: the kernel density estimates of the mean branch velocity parameters (branch velocity variation among branches *vs* mean branch velocity), the kernel density estimates of the weighted diffusion velocity parameters (branch velocity variation among branches *vs* weighted diffusion 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.

Step 3: preliminary analysis of the environmental raster

When we have several different environmental rasters to test, this is useful to preliminary investigate which ones are potential resistance or conductance factors and then focus on a restricted set of selected raster files. This first analysis is directly based on the environmental raster cell values and without performing any randomisation steps. When the number of randomisation steps is set to zero, the function simply estimates the correlation between dispersal durations and environmental distances associated with each phylogenetic branch. In the context of this tutorial, we will estimate the correlation between the dispersal durations and the environmental distances computed for each branch using the least-cost method [7, 8] and while treating the "elevation" raster as a potential resistance factor. Note that when we do not have any prior information about the impact of the environmental variables, it might make sense to test each factor once as a resistance and once as a conductance factor.

Before specifying the different parameters for this analysis, we will first modify the cell values of the "elevation" raster so that there isn't any negative value on the grid, which is not allowed when using the different path models (least-cost or Circuitscape path model) proposed in "`seraphim`". In addition, we will increase all the cell values by "1" so that minimum cell values equal to "1" instead of "0" (note that this operation will not affect cells with a "no data" value):

```
> rast = raster("Elevation_raster.asc")
> rast[rast[]<0] = 0
> rast[] = rast[] + 1
```

We can then plot the resulting raster using the "plot" from the package "`raster`" or the customised "rasterPlot" function from the package "`seraphim`":

```
> plotRaster(rast, addAxes=T, addLegend=T)
```

The aim of the latter modification is to allow a comparison with an artificial raster where all the cell values equal to "1". This "null" raster will be used to compute environmental distances with the selected path model (least-cost or Circuitscape path model). In addition to environmental distance(s), each phylogenetic branch will then be also associated with an environmental distance computed on a "null" raster, which will be a proxy of the geographical distance associated with each branch.

After these preliminary raster manipulations, the different parameters of the "spreadFactors" function have to be specified as follows:

```
> envVariables = list(rast)
> pathModel = 2
> resistances = list(TRUE)
> avgResistances = list(TRUE)
> fourCells = FALSE
> nberOfRandomisations = 0
> randomProcedure = 3
> outputName = "Elevation_least-cost"
```

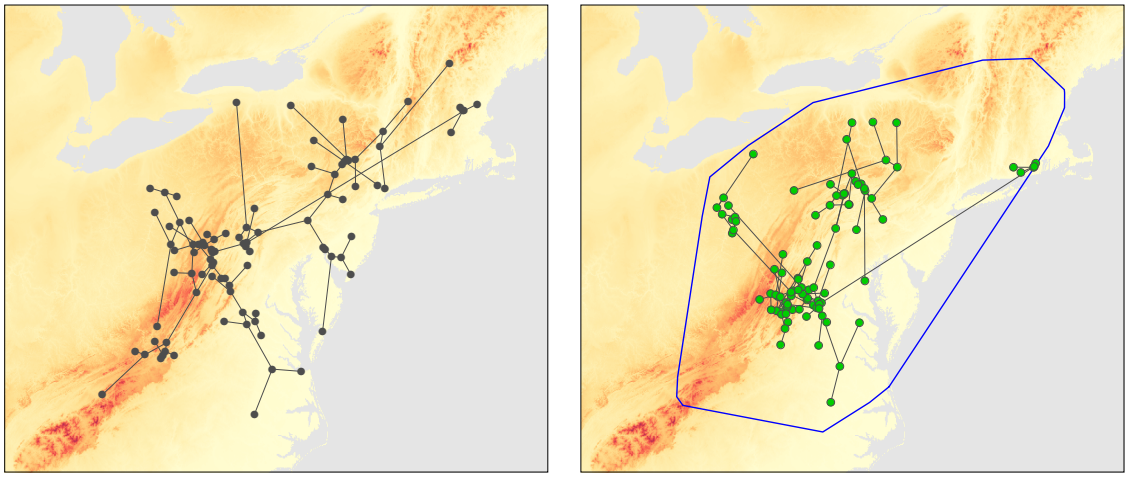

Figure 1: sampled and randomised trees mapped on the elevation raster. On the left: the original environmental raster (representing, in this case, elevation) upon which is superimposed the movement events extracted from one spatiotemporally-referenced phylogeny. On the right: the result of one randomisation of node positions. This randomisation procedure is performed within a minimum convex hull (shown in blue), which is defined by the node locations of all selected phylogenies.

Even if in this particular case we focus on only one raster file, the "envVariables" object has to be a list of raster files and the "resistances" object has to be a vector of boolean variables specifying if each raster has to be treated as a resistance ("TRUE") or a conductance ("FALSE") variable. The "pathModel" variable specifies which path taken model has to be used to compute the environmental distances associated with each branch: "1" (straight-line path model), "2" (least-cost path model [7, 8]) or "3" (Circuitscape path model [9]). The "avgResistance" and "fourCells" parameters are not important at this stage and are only used with the Circuitscape

path model (see the package manual as well as the manual of Circuitscape for further details). The "randomProcedure" is not important at the moment but has to be created; simply set it equal to "3" (default, see below). Like for the "spreadStatistics" function, the "outputName" string will be used as a prefix to name the different outputs of the function. Once all these parameters have been specifying, the function can be launched using the following command:

```
> spreadFactors(localTreesDirectory, nberOfExtractionFiles, envVariables,
pathModel, resistances, avgResistances, fourCells, nberOfRandomisations,
randomProcedure, outputName)
```

The function will generate a text file containing some statistic values (one per statistic and per sampled/extracted tree) measuring the correlation between dispersal durations and environmental distances computed for each branch. These statistics are the "environmental" determination coefficient (estimated from the univariate linear regression between the dispersal durations and the environmental distances associated with each branch) and the statistic Q (difference between the environmental determination coefficient and the "spatial" determination coefficient estimated from the univariate linear regressions between the dispersal durations and the geographical distances associated with each branch; n.b.: Q was previously referred as D in [1]). Note that, as mentioned above, the geographical distance is computed using the selected path taken model on a "null" raster with uniform cell values equal to one. For instance, as we can see in Figure 2, the distribution of determination coefficients differences Q's clearly tends to be different from zero. This result indicates that the "elevation" raster treated as resistance is a potential factor that could have had an impact on the dispersal rate of the RABV epidemic.

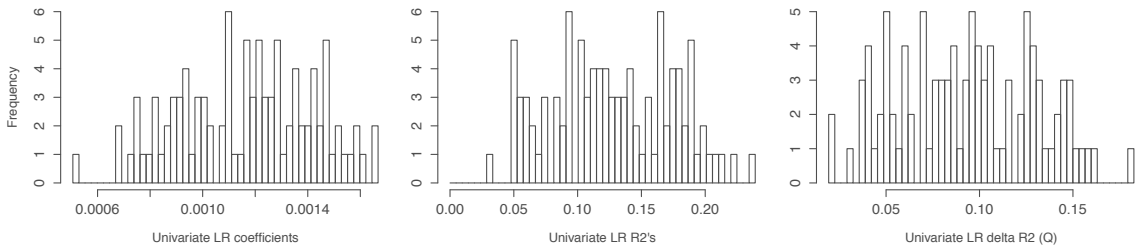

Figure 2: results of the linear regressions performed between dispersal durations and least-cost distances computed on the elevation raster. Each value on the histograms corresponds to one sampled tree.

Step 4: test based on a randomisation procedure

The final step is to test the level of significance of the statistic Q estimating the correlation between dispersal durations and environmental distances computed for each branch and based on the "elevation" resistance raster. Here, we will use the randomisation of phylogenetic node positions, which was already specified above ("randomProcedure = 3"). The approach described below is based on a single randomisation step performed for each sampled tree and returns a Bayes factor value per tested environmental factor [10]. The BF_e for a particular environmental factor e is approximated by the posterior odds that $Q_{observed} > Q_{randomised}$ divided by the equivalent prior odds (the prior probability for $Q_{observed} > Q_{randomised}$ is considered to be 0.5):

$$BF_e = \frac{p_e}{1-p_e} / \frac{0.5}{1-0.5} = \frac{p_e}{1-p_e}$$

where p_e is the posterior probability that $Q_{observed} > Q_{randomised}$, i.e. the frequency at which $Q_{observed} > Q_{randomised}$ in the samples from the posterior distribution. The prior odds is "1" because we have an equal prior expectation for $Q_{observed}$ and $Q_{randomised}$. The formal estimate of posterior predictive odds is analogous to computing BFs in case two alternative hypotheses exist, such for the inclusion of rate parameters or predictors in BSSVS procedures (Bayesian stochastic search variable selection; see equation (6) in Lemey *et al.* [11]). Bayes factor are automatically estimated by the "spreadFactors" function when the "nberOfRandomisations" is at least set to "1". In the case of the elevation raster tested as a resistance factor in the context of this tutorial, the BF is >20. It is then considered as a "strong" evidence of the statistical significance of $Q_{observed}$ (see Table 1 for the scales of interpretation of Bayes factor values).
In practice, we just have to set the number of randomisation steps per sampled tree to "1":

```
> nberOfRandomisations = 1
```

Once this new parameter is specified, the "spreadFactors" function can be re-launched with the same command:

```
> spreadFactors(localTreesDirectory, nberOfExtractionFiles, envVariables,
pathModel, resistances, avgResistances, fourCells, nberOfRandomisations,
randomProcedure, outputName)
```

Table 1: scale of interpretation of Bayes factors (BF) according to Jeffreys [12] and Kass & Raftery [13].

Scale of interpretation defined by Jeffreys [12]			Scale of Kass & Raftery [13]	
BF values	$log_{10}(BF)$	Strength of evidence	BF values	Strength of evidence
$3.16 - 10$	$0.5 - 1$	substantial	$3 - 20$	positive
$10 - 31.62$	$1 - 1.5$	strong	$20 - 150$	strong
$31.62 - 100$	$1.5 - 2$	very strong	>150	very strong
>100	>2	decisive		

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] Biek R, Henderson JC, Waller LA, Rupprecht CE, Real LA (2007). A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *PNAS* 104: 7993-7998.

[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] 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.

[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] Dijkstra EW (1959). A note on two problems in connexion with graphs. *Numerische Mathematik* 1: 269-271.

[8] Van Etten J (2012). R package gdistance: distances and routes on geographical grids. R package version 1.12.

[9] McRae BH (2006). Isolation by resistance. *Evolution* 60: 1551-1561.

[10] 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*, *in press*.

[11] Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology* 5.

[12] Jeffreys H (1961). Theory of Probability (3rd edition). Oxford University Press, Oxford.

[13] Kass RE, Raftery AE (1995). Bayes Factors. *Journal of the American Statistical Association* 90: 791.