Tutorial for the R package `seraphim` 1.0

Using phylogenetically-informed movement data to study the association between an environmental variable and the dispersal velocity of a virus spread

Simon Dellicour

June 18, 2019

The present tutorial describes how to use the package "`seraphim`" [1, 2] to use phylogenetically informed movement data to study the association between a particular environmental variable (an "elevation" raster) and the dispersal velocity of the rabies virus (RABV) spread in North American raccoon populations [1, 3]. See also the package manual for further details. The first step is to download the package (`http://evolve.zoo.ox.ac.uk/Evolve/Software.html`) and place the "seraphim_1.0_VERSION.beta.tar.gz" file in a R workspace directory. The package can then be installed from this archive file using the following R command:

```
> install.packages("seraphim_1.0_VERSION.tar.gz", repos=NULL, type="source")
```

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 example files also available at `http://evolve.zoo.ox.ac.uk/Evolve/Software.html`. These files are "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 in trees

We will first extract the spatio-temporal information contained in phylogenetic trees. This kind of trees have to be in an annotated Newick format and can, for instance, be inferred by the continuous phylogeographic method implemented in BEAST [4]. The tree file "RABV_gamma.trees" contains 5,001 trees sampled by the MCMC chain. We will here use the "treeExtractions" function to extract the 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" (name of the ".trees" 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), "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)
```

After this extraction step, each phylogenetic branch can be considered as a vector defined by its start and end location (latitude and longitude), and its start and end dates (in decimal units). Each phylogeny branch therefore represents a conditionally independent viral lineage dispersal event [5].

Step 2: estimating the correlation between branch durations and environmental distances

We will now compute and assign an "environmental distance" to each of the vectors obtained in step 1. An "environmental distance" is here defined as a distance metric weighted according to the values of an environmental variable at each location [1, 6]. Two distinct path models can a priori be used to compute the environmental distance allocated to each phylogeny branch for a given environmental raster: (i) the "least-cost" path model, which uses a least-cost algorithm to determine the route taken between the

start and end points [7, 8], and (ii) the "random walk" or "Circuitscape" path model, which uses circuit theory to accommodate uncertainty in the route taken [9]. Note that for these path models, the tested environmental factor has to be considered either as a conductance factor (i.e., as a variable that facilitates movement) or as a resistance factor (i.e., impedes movement). Once the environmental distances have been computed, we can then estimate, for each of the 100 posterior trees, the correlation between the phylogeny branch durations (branch lengths in units of time) and their associated environmental distance [1, 6]. Specifically, we estimated the statistic $Q = R^2_{env}$ - R^2_{null}, where R^2_{env} is the coefficient of determination obtained when branch durations are regressed against environmental distances computed on the environmental raster, and R^2_{null} is the coefficient of determination obtained when branch durations are regressed against environmental distances computed on a "null" raster, that is, the environmental raster with a value of "1" assigned to all the cells (except cells with no original data). The Q statistic (previously referred as "D" in [1]) therefore represents how much variation in lineage movement is explained when spatial heterogeneity in the environmental variable is taken into account, above and beyond that explained by distance alone [1]. Therefore, when $Q > 0$, distances weighted according to a heterogeneous environmental raster are correlated more strongly with branch duration than distances computed on a "null" raster (which represents geographical distance alone). Since one Q value was calculated for each sampled posterior tree, we then obtained a distribution of 100 Q values.

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 branch durations and environmental distances associated with each phylogenetic branch. In the context of this tutorial, we will estimate the correlation between the branch 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. The different parameters of the "spreadFactors" function have to be specified as follows:

```
> envVariables = list(raster("Elevation_raster.asc"))
> pathModel = 2
> resistances = list(TRUE)
> avgResistances = list(TRUE)
> fourCells = FALSE
> nberOfRandomisations = 0
> randomProcedure = 3
> outputName = "Elevation_least-cost"
> showingPlots = FALSE
> nberOfCores = 1
> OS = "Unix"
```

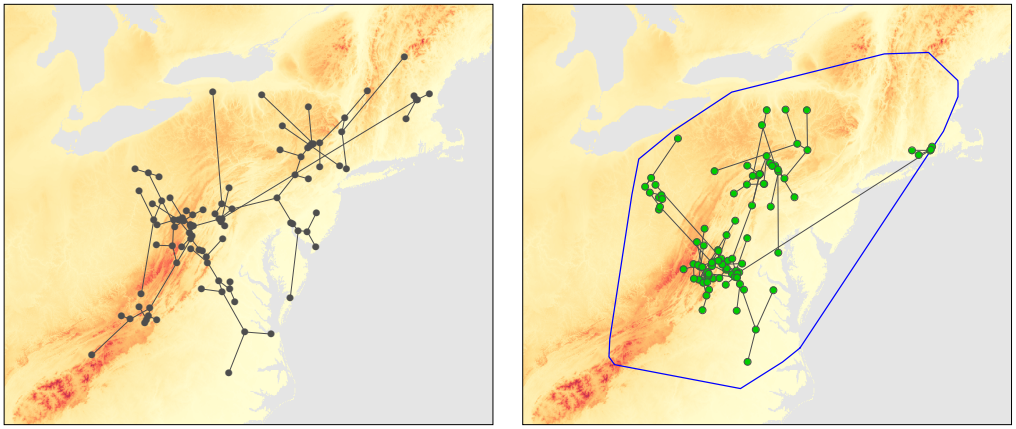

Figure 1: sampled and randomised trees mapped on the elevation raster. On the left: the original environmental raster (representing, in this case, elevation) upon which is superimposed the movement events extracted from one spatiotemporally-referenced phylogeny. On the right: the result of one randomisation 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" value is not important at the moment but has to be created (simply set it equal to "1"). Like for the "spreadStatistics" function, the "outputName" string will be used as a prefix to name the different outputs of the function. If the boolean parameter "showingPlots" equals "TRUE", the function will generate and save several graphs like the one displayed in Figure 1 (but in that case, the function will run much slower). Finally, we also have to specify the number of cores ("nberOfCores") to use and the operating system on which the function will run. At this stage, parallelisation of the code is not useful and then simply set the number of cores to one. The information about the nature of the operating system is only useful when the function has to call the "Circuitscape" Python package [9] (see the `seraphim` manual for further details). Before launching the function, we will modify the raster cell values so that minimum raster cell values equal to one instead of zero (note that this operation will not affect cells with a "no data" value). For that purpose, we will add a value of one to all the cells (except the ones with a "no data" value):

```
> envVariables[[1]][] = envVariables[[1]][] + 1
```

The aim of this modification is to allow a comparison with an artificial raster with all the cell values equal to one. This "null" raster will be used to compute environmental

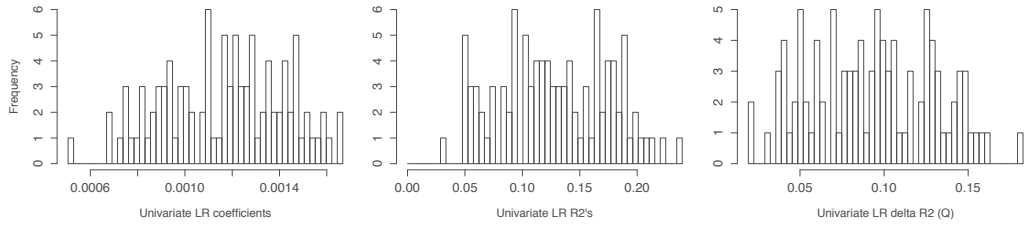

Figure 2: results of the linear regressions performed between branch durations and least-cost distances computed on the elevation raster. Each value on the histograms corresponds to one sampled tree.

distances with the selected path taken model (straight-line, 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 this step, the function can be launched using the following command:

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

The function will generate a text file containing some statistic values (one per statistic and per sampled/extracted tree) measuring the correlation between branch durations and environmental distances computed for each branch. These statistics are the linear regression coefficient and "environmental" determination coefficient R^2_{env} estimated from the univariate linear regression between the branch durations and the environmental distances associated with each branch, as well as the statistic Q, which is thus the differences between R^2_{env} and R^2_{null} (R^2_{null} being the "null" determination coefficient estimated from the univariate linear regressions between the branch 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 here computed using the selected path model on a "null" raster with uniform cell values equal to "1".

A variable can only be considered as potentially explanatory if both its distribution of regression coefficients and associated Q's distribution are positive [10]. As we can see in Figure 2, the distributions of regression coefficients and Q values are here both clearly higher than zero. Alternatively, we can also read and analyse the generated text file to report the percentages of positive regression coefficients and Q values:

```
> tab = read.table("Elevation_least-cost_LR_results.txt", header=T)
> LR_coefficients = tab[,"Univariate_LR_coefficients_Elevation_raster_R"]
> print(sum(LR_coefficients > 0))

100

> Qs = tab[,"Univariate_LR_delta_R2_Elevation_raster_R"]
> print(sum(Qs > 0))

100
```

These results thus indicate 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. We can now go to the next step (step 3) to assess the level of significance of the distribution of Q values. At this stage, it is important to note that testing the significance of the Q distribution does not really make sense if the regression coefficient and/or the Q distributions are not positive. Indeed, in the first case, this would mean that branch durations are negatively correlated with environmental distances and, in the second case, this would mean that considering environmental distances computed on the environmental raster rather than on the "null" raster does not improve the linear regression fit (and this even if the Q distribution is potentially significant). This second step of the workflow thus also aims at selecting environmental factors for which the significance of the Q distribution has to be tested by the randomisation procedure of step 3. In the case where Q distributions are not entirely positive, a solution can be to only select environmental factors for which the proportion of positive Q is higher than 90 or 95%.

Step 3: testing the significance with a randomisation procedure

The final step is to test the level of significance of the distribution of Q values against a null model of no impact of the environmental factor. To generate an appropriate null distribution for Q, we will use a randomisation procedure presented in [1, 2] and in which phylogenetic node positions were randomised within the study area, under the constraint that branch lengths, tree topology and root position are unchanged. This randomisation procedure was already specified above ("randomProcedure = 3"). Below, we describe two different approaches to interpret the randomisation results. The first approach is based on several (100) randomisation steps and has been used in [1] to study the impact of environmental factors on the RABV dataset considered in this tutorial. With this first approach, we report a proportion of posterior trees for which a significant correlation between dispersal duration and environmental distance is identified. In the second approach, each sampled posterior tree is randomised only once to generate the equivalent value under the null hypothesis; this results in a null distribution of Q values that can be compared directly to the posterior distributions of estimated Q values to report Bayes factor like support [6].

3.1. First approach: reporting a proportion of p-values < 0.05

First, we have to specify the number of randomisation steps we want to perform:

```
> nberOfRandomisations = 100
```

Once these new parameters are specified, the "spreadFactors" function can be re-launched:

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

The randomisations can be very time consuming (especially when analysing several rasters and/or when selecting the Circuitscape model). In that case, "the spreadFactors" function can run on several cores by specifying a number of cores higher than one with the "nberOfcores" parameter. The parallelisation is performed using the "doMC" package and only works with an R script launched from a terminal (on Unix terminal, use the command "R < myScript.r −−no−save"). Again, the function will generate and save a text file. To know the exact proportion of p-values smaller than 0.05 estimated from the comparison between randomised and estimated Q values:

```
> tab = read.table
("Elevation_least-cost_randomisation_results.txt", header=T)
> a = tab[,"Uni_LR_delta_R2_p.values_Elevation_R"]
> print(length(a[a[]<0.05])/nberOfExtractionFiles)

0.81
```

Based on the randomisation of phylogenetic node positions, 81 out of 100 sampled trees are associated with a significant p-value. This number of p-values smaller than 0.05 mean that for 81% of the trees sampled in the posterior distribution, elevation is significantly associated with a slower rabies spread in North American raccoon populations.

3.2. Second approach: reporting a Bayes factor

Reporting proportions of p-values <0.05 has several disadvantages: (i) it requires several randomisation steps, (ii) it is cut-off dependent and therefore only captures a part of the randomisation results (for example, an important proportion of p-values could be higher than but very close to 0.05), and (iii) the proportion metric may in itself be difficult to interpret. For all these reasons, we implement the approximation of Bayes factor (BF) supports, with one BF value returned per tested environmental factor [6]. This second approach only requires one randomisation per sampled tree. The BF_e for a particular environmental factor e is approximated by the posterior odds that $Q_{estimated} > Q_{randomised}$ divided by the equivalent prior odds (the prior probability for $Q_{estimated} > Q_{randomised}$ 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 $Q_{estimated} > Q_{randomised}$, i.e. the frequency at which $Q_{estimated} > Q_{randomised}$ in the samples from the posterior distribution. The prior odds is 1 because we have an equal prior expectation for $Q_{estimated}$ 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 approximated 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 $Q_{estimated}$ (see Table 1 for the scales of interpretation of Bayes factor values).

When one or several tested environmental factors is/are associated with a strong support (BF $>$ 20), it is then interesting to focus on their impact by analysing/comparing the estimated distributions of Q values. For instance, in the specific case of the "elevation" raster tested as a resistance factor for the raccoon RABV spread in North America, this distribution indicates that considering the environmental rather than the "null" raster increases the determination coefficient of around 10%.

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] 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.

[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] Jacquot M, Nomikou K, Palmarini M, Mertens P, Biek R (2017). Bluetongue virus spread in Europe is a consequence of climatic, landscape and vertebrate host factors as revealed by phylogeographic inference. *Proceedings of the Royal Society B: Biological Sciences* 284: 20170919.

[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.