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

December 27, 2024

The present tutorial describes how to use the R package "`seraphim`" [1, 2] to exploit phylogenetically informed movement data obtained from a continuous phylogeographic reconstruction [3] in order to investigate the association between a particular environmental factor (an "elevation" raster) and the diffusion velocity of a rabies virus (RABV) spread in the North American raccoon population [1, 4]. At the end of the present tutorial, we also introduce the alternative isolation-by-resistance (IBR) analyses than be conducted with a very similar setting.

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 example files also available on the GitHub repository of the package (`https://github.com/sdellicour/seraphim/tree/master/tutorials`): (i) "RABV_gamma.trees", a file containing annotated phylogenetic trees sampled from a posterior distribution of trees inferred for the RABV dataset using the continuous phylogeographic approach developed by Lemey and colleagues [3]; and (ii) "Elevation_raster.asc", the environmental raster "elevation" encompassing the study area.

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 "RABV_gamma.trees" contains 5,001 trees sampled by the MCMC chain. We will here use the "treeExtractions()" function to extract the spatio-temporal information embedded 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 posterior 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="RABV_gamma.trees", what="", sep="\n", quiet=T)
> burnIn = 1001
> randomSampling = FALSE
> nberOfTreesToSample = 100
> mostRecentSamplingDatum = 2004.7
> coordinateAttributeName = "location"
```

Once all these parameters have been specifyied, 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). 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 [6].

Step 2: preliminary analysis of the environmental raster

When we have several different environmental rasters to test, this is useful to preliminary investigate which ones are potentially acting as 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 step. When the number of randomisation steps is set to zero, the function simply estimates the association between dispersal durations and environmental distances associated with each phylogenetic branch. In the context of this tutorial, we will investigate the association between the dispersal durations and the environmental distances computed for each branch using the least-cost path algorithm [7, 8] and while treating the "elevation" raster as a potential resistance factor. However, when we do not have any prior information about the impact of the environmental variables, it might make sense to test each environmental variable (raster) 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, negative values being 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 "NA"):

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

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

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

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 [7, 8] or Circuitscape [9] 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.

Note that users might want or need to test various transformations of the original raster file (e.g., rescaling or log-transformation of raster cell values). For instance, one can generate and test several distinct rasters by transforming the original raster cell values with the following formula: $v_t = 1 + k(v_o/v_{max})$, where v_t and v_o are the transformed and original cell values, and v_{max} the maximum cell value recorded in the raster [10]. In that case, the rescaling parameter k allows the definition and testing of different strengths of raster cell conductance or resistance, relative to the conductance/resistance of a cell with a minimum value set to "1" (e.g., $k = 10$, 100, and 1000).

Once the raster(s) to test is/are ready, 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 = "RABV_elevation_least-cost"
```

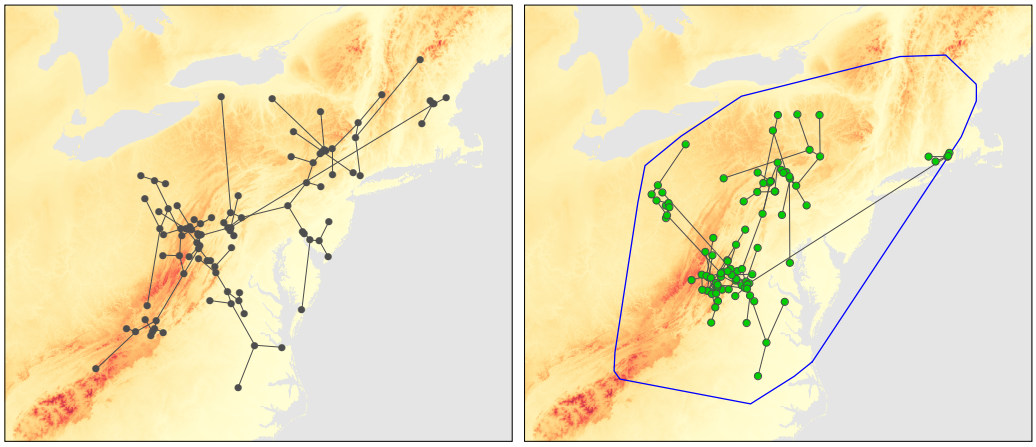

Figure 1: sampled and randomised trees mapped on the elevation raster. On the left: the original environmental raster (representing, in this case, elevation) upon which is superimposed the lineage dispersal events extracted from one annotated tree sampled from the posterior distribution of trees obtained through continuous phylogeographic inference. On the right: the result of one randomisation of branch 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 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 "fourCells" boolean parameter is used to specify if a given raster cell should be connected to either its four first-order ("TRUE") or eight first- and second-order ("FALSE"). The "avgResistance" parameter is not important at this stage and is 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)
```

Because at this stage the number of randomisations to perform is set to zero, the function will only generate a text file listing the results of the two linear regression (LR) analyses conducted for each phylogenetic tree: (i) the univariate linear regression between the dispersal durations (t) and the environmental distances (d) computed for a given environmental factor (for which the output files are labelled "LR1"), and (ii) the univariate linear regression between four times the dispersal durations ($4t$) and the squared environmental distances (d^2) computed for a given environmental factor (for which the output files are labelled "LR2"). While the first linear regression (LR1) aims to assess the impact of a given environmental factor on the heterogeneity of the lineage dispersal velocity ($t \sim d$), the second linear regression (LR2) aims to assess the impact of a given environmental factor on the heterogeneity of the diffusion velocity ($4t \sim d^2$), which is related to the diffusion coefficient identified as a metric more robust to the sampling intensity (i.e. the sampling size) than the lineage dispersal velocity ([13]). For both linear regressions, the function will thus generate a first text file gathering the following elements: the β regression coefficients and coefficient of determination R^2_{env} obtained from each linear regression, as well as the difference Q between R^2_{env} and R^2_{null} (the coefficient of determination obtained from the linear regression based on environmental distances computed on a "null" raster with uniform cell values equal to "1", the later distances corresponding to a proxy of the geographic distances; n.b.: Q was previously referred as D in [1]). Note that in the context of this tutorial and given the higher robustness of the diffusion coefficient metric ([13]), we will solely focus on the results related to the linear regression "LR2" between four times the dispersal durations ($4t$) and the squared environmental distances (d^2). For LR2 and as detailed below, the distribution of determination coefficients differences Q's clearly tends to be different from zero. This result indicates that the "elevation" raster treated as a resistance factor is an environmental variable that could have had an impact on the diffusion velocity of RABV lineages.

A variable can only be considered as potentially explanatory if both its distribution of regression coefficients and associated Q's distribution are positive [11]. As we can see below (and which can also, e.g., be visualised with histogram plots), the distributions of regression coefficients and Q values are here both clearly higher than zero. This can be formally assessed by analysing the generated text file to report the percentages of positive regression coefficients and Q values:

```
> tab = read.table("RABV_elevation_least-cost_linear_regression_results.txt",
header=T)
> LR_coefficients = tab[,"LR2_coefficients_elevation_R"]
> print(sum(LR_coefficients > 0))

100

> Qs = tab[,"LR2_Q_elevation_R"]
> print(sum(Qs > 0))

99
```

```
> Qs = tab[,"LR2_Q_elevation_R"]
> print(mean(Qs))

0.063
```

With a positive posterior distribution of Q values and an average difference (between R^2_{env} and R^2_{null}) $> 6\%$, these results thus indicate that the "elevation" raster treated as a potential resistance factor correspond to an environmental variable that could have had an impact on the diffusion velocity of RABV lineages inferred here. We can now go to the next step (step 3) to assess the statistical support associated with the distribution of Q values. As outlined above, it is important to note that assessing the statistical support 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, 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 statistical support 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, e.g., higher than 90 or 95%.

Step 3: assessing statistical support with a randomisation procedure

The final step is to assess the statistical support associated with the statistic Q, i.e. the statistic estimating an association between dispersal durations and environmental distances computed for each branch and based on the "elevation" raster teated as a potential resistance factor. 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 [12]. 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.* [14]). Bayes factor are automatically estimated by the "spreadFactors()" function when the "nberOfRandomisations" is at least set to "1". 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)
```

In the case of the elevation raster tested as a potential resistance factor, the BF is >20. It is then considered as a "strong" statistical support for $Q_{observed}$ (see Table 1 for the scales of interpretation of Bayes factor values).

Table 1: scale of interpretation of Bayes factors (BF) according to Jeffreys [15] and Kass & Raftery [16].

Scale of interpretation defined by Jeffreys [15]			Scale of Kass & Raftery [16]	
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		

Alternative analysis: investigation of the isolation-by-resistance patterns

Analytical steps 2 and 3 described above can also and alternatively be conducted with the "isolationByResistance()" function that is based on the exact same parameters specification. While the "spreadFactors()" function can be used to investigate the association between branch dispersal/diffusion velocity and environmental distances (i.e. the impact of environmental factors on the heterogeneity of the dispersal/diffusion heterogeneity of viral lineages), the "isolationByResistance()" function will instead investigate the association between the phylogenetic (i.e. patristic) and environmental distances between each pair of tip nodes, i.e. the impact of environmental factors on the isolation-by-distance (IBD) pattern, which can be characterised as an isolation-by-resistance (IBR) analysis. This function will estimate and assess the statistical support of an alternative correlation Q statistic this time defined as the difference between (i) the coefficient of determination R^2_{env} obtained from the univariate linear regression between the patristic distances and the log-transformed environmental distances computed on a given environmental factor and for each pair of tip nodes, and (ii) the coefficient of determination R^2_{null} obtained from the univariate linear regression between the patristic distances and the log-transformed environmental distances computed on the corresponding null raster (environmental distances computed on such a null raster being again considered as a proxy for the geographical distance, i.e. an IBD setting). In the present case, the Q statistic thus measures to what extent a given environmental factor could potentially explain a deviation from

the IBD pattern. As it is not based on the analysis of phylogenetic branch positions inferred through a phylogeographic reconstruction, the IBR analysis can therefore not be considered as a landscape phylogeographic approach emphper se. While a emphlandscape phylogeographic approach aims to analyse the impact of environmental factors on the dispersal dynamic of lineages, the IBR analysis is indeed conceptually closer to a emphlandscape genetic approach aiming to uncover the environmental factors impacting the inter-individual genetic differentiation. Because this analysis is however based on pairwise patristic distances computed on (time-scaled) phylogenetic trees, we propose to characterise it as a *landscape phylogenetic approach*, i.e. an analytical approach aiming to investigate the impact of environmental factors on the phylogenetic distance or divergence time (in the context of time-scaled phylogenetic inference) between invidividuals. Moreover, the "isolationByResistance()" function uses the same randomisation procedures as implemented in the "spreadFactors()" function to assess the statistical support of the Q statistic. Given that those procedures are based on the randomisation of branch positions inferred on the map by a continuous phylogeographic inference, the IBR analysis implemented here still require such a phylogeographic reconstruction, even if the outcome of the phylogeographic inference itself is thus not involved in the computation of the Q statistic defined here.

As mentioned above, the "isolationByResistance()" function can be launched using the exact same settings as the "spreadFactors()" function:

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

> tab = read.table("RABV_elevation_least-cost_linear_regression_results.txt",
header=T)
> LR_coefficients = tab[,"LR_coefficients_elevation_R"]
> print(sum(LR_coefficients > 0))

100

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

100

> Qs = tab[,"LR2_Q_elevation_R"]
> print(mean(Qs))

0.116
```

In the case of the elevation raster tested as a potential resistance factor, the posterior distribution estimated for the Q statistic is clearly positive and has a mean value of 12%. Furthermore, the BF obtained for this analysis is again >20, confirming that this environmental factor can indeed be associated with a deviation from the IBD pattern.

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] 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.

[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] 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.

[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, Lequime S, Vrancken B, Gill MS, Bastide P, Gangavarapu K, Matteson NL, Tan Y, du Plessis L, Fisher AA, Nelson MI, Gilbert M, Suchard MA, Andersen KG, Grubaugh ND, Pybus OG, Lemey P (2020). Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nature Communications* 11: 5620.

[11] 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.

[12] 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.

[13] 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.

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

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

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