Tutorial for the R package seraphim 1.0

Using phylogenetically-informed movement data to study the dispersal tendency across and towards specific environmental conditions

Simon Dellicour, Philippe Lemey

July 3, 2022

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 study the association between a particular environmental variable (a human population density raster) and the dispersal events of the rabies virus (RABV) in Iran [4]. In particular, the objective is to test if viral lineages tended to remain in and/or to disperse to areas associated with higher human population density. 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", "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) "Extracted_trees", a folder containing "csv" extraction files, and (ii) the human population density raster named "Pop_density.asc". Extraction files were obtained by extracting the spatio-temporal information contained in trees sampled from the posterior distribution of continuous phylogeographic inferences [3] (see for instance the tutorial "impact on dispersal velocity" for an example of such an extraction step). In these extraction files, each line corresponds to a specific phylogeny branch that 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 phylogenetic branch therefore represents a conditionally independent viral lineage dispersal event [5].

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 [6] is available on the BEAST community website (`https://beast.community/workshop_continuous_diffusion_yfv`).

Here, we will use these extraction files to investigate if RABV lineages tended to remain in and/or to disperse towards areas of higher human population density. For that purpose, we will use the "spreadFactors()" function that is also used in the tutorial "impact on dispersal velocity" but this time, we will specify that we do not want to use a path model ("pathModel = 0"). By doing so, we specify that we do not want to use a path model to compute environmental distances and analyse their correlation with branch dispersal durations. In this example, we will also specify that we want to test the environmental raster as a potential "conductance" factor ("resistances = list(FALSE)"). Indeed, as stated above, we here want to investigate if areas associated with higher human population density tended to "attract" dispersal events of RABV lineages. It is important to note that the "resistance/conductance" terminology is based on the path model specification. In the present context where we do not use/set a path model, "resistance factor" or "conductance factor" mean that we test the corresponding environmental variable as a factor repulsing or attracting tree nodes, respectively.

The other parameters of the "spreadFactors()" function have to be specified as follows:

```
> localTreesDirectory = "Extracted_trees"
> nberOfExtractionFiles = 900
> envVariables = list(raster("Pop_density.asc"))
> pathModel = 0
> resistances = list(FALSE)
> avgResistances = list(FALSE)
> fourCells = FALSE
> nberOfRandomisations = 1
> randomProcedure = 3
> outputName = "Pop_density"
> showingPlots = FALSE
```

Although we focus on only one raster file in this case, 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 "avgResistance" and "fourCells" parameters are not at all used for this analysis but cannot be left unspecified. As for the "outputName" string, it 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 (but in that case, the function will run much slower).

The function can then be launched using the following command:

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

With "pathModel" set to "0", the "spreadFactors()" function will compute (i) the value E, which is the mean of the environmental values extracted at the nodes' position, and (ii) the ratio R defined as the proportion of branches for which the environmental value recorded at the oldest node position is higher than the environmental value recorded at the youngest node position. While E measures the tendency of tree nodes to remain located in lower/higher environmental values, R rather measures the tendency of lineages to disperse towards lower/higher environmental values. These two metrics are computed for each tree sampled from a posterior distribution, and we therefore obtain a posterior distribution for E and R. Finally, each of these two posterior distributions is compared to a null distribution of the same metric computed after having randomised phylogenetic node positions within the study area, under the constraint that branch lengths, tree topology and root position are unchanged ("randomProcedure = 3"). This approach only requires one randomisation per sampled tree and leads to the approximation of a Bayes factor (BF) support for each statistic. For a particular environmental factor e tested as a factor attracting lineages, the Bayes factor BF_e associated with the statistic E is approximated by the posterior odds that $E_{estimated} > E_{randomised}$ divided by the equivalent prior odds (the prior probability for $E_{estimated} > E_{randomised}$ 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 $E_{estimated} > E_{randomised}$, i.e. the frequency at which $E_{estimated} > E_{randomised}$ in the samples from the posterior distribution. The prior odds is 1 because we have an equal prior expectation for $E_{estimated}$ and $E_{randomised}$. The formal estimate of posterior predictive odds is analogous to approximating BFs in case two alternative hypotheses exist, such for the inclusion of rate parameters or predictors in BSSVS procedures (Bayesian stochastic search variable selection [7, 8]). Bayes factor are automatically approximated by the "spreadFactors()" function when the "nberOfRandomisations" is at least set to "1". As output, the function will generate a text file reporting the different BF values. Alternatively, if the environmental factor was tested as a factor repulsing lineages, BF_e would be approximated by the posterior odds that $E_{estimated} < E_{randomised}$ divided by the equivalent prior odds.

The same approach is used to approximate Bayes factor supports for the statistic R. Whether the environmental factor is tested as a factor attracting or repulsing lineages, the posterior BF_e for R is approximated by the posterior odds that $R_{estimated} < R_{randomised}$ (attracting lineages) or that $R_{estimated} > R_{randomised}$ (repulsing lineages) divided by the equivalent prior odds.

The function will create two distinct text files reporting the Bayes factor supports associated with each tested environmental factors; one for the statistic E ("Pop_density_direction1_Bayes_factors.txt") and one for the statistic R ("Pop_density_direction2_Bayes_factors.txt"). In the case of the human population density raster tested as a factor attracting RABV lineage dispersal events in Iran, the BF support is >20 for both statistic, which can be considered as "strong" evidences of the statistical significance of $E_{estimated}$ and $R_{estimated}$ [9].

As these tests are directly based on the environmental values extracted at internal and tip node positions, their outcome can be particularly impacted by the nature of sampling [10]. Indeed, half of the node positions, i.e., the tip node positions, are directly determined by the sampling. To assess the sensitivity of the tests to heterogeneous sampling, one could e.g. repeat these tests while only considering internal tree nodes. Since internal nodes are phylogeographically linked to tip nodes, discarding tip branches will, however, only mitigate the direct impact of the sampling pattern on the outcome of the analysis. Overall, those tests remain influenced by sampling effort and pattern, and should then be considered more as a description of the environmental context of inferred virus lineage dispersal rather than a robust test of the impact of those conditions on dispersal [4].

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] Dellicour S, Troupin C, Jahanbakhsh F, Salama A, Massoudi S, Moghaddam MK, Baele G, Lemey P, Gholami A, Bourhy H (2019). Using phylogeographic approaches to analyse the dispersal history, velocity, and direction of viral lineages application to rabies virus spread in Iran. *Molecular Ecology* 28: 4335-4350

[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] 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.

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

[8] 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.

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

[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.