VEME workshop 2022, Panama City
Tutorial for the R package `seraphim` 1.0

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

Simon Dellicour

July 9, 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 investigate the association between a particular environmental factor (an “elevation” raster) and the dispersal velocity of a rabies virus (RABV) spread in the North American raccoon population [1, 4]. 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: (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 from trees

The first step is to extract the spatio-temporal information embedded in annotated phylogenetic trees sampled from the posterior distribution of the 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 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=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 [6].

Step 2: estimating dispersal statistics (optional)

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.* [6], and the weighted coefficient $D_{weighted}$ defined by Trovão *et al.* [7]. 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. For a given tree, branches with short duration will have respectively less of an impact on $v_{weighted}$ and $D_{weighted}$ than on v_{branch} and $D_{original}$, and therefore also on the resulting variance among $v_{weighted}$ and $D_{weighted}$ values across all trees. Compared to v_{branch} and $D_{original}$, $v_{weighted}$ and $D_{weighted}$ can respectively allow a better discrimination among epidemics with different diffusivity because it is associated with a smaller variance [8].

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 here generate and save six different graphs: the kernel density estimates of the mean branch velocity parameters (branch dispersal velocity variation among branches *vs* mean branch dispersal velocity), the kernel density estimates of the weighted branch dispersal 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 (inferred root location). 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 of the tree.

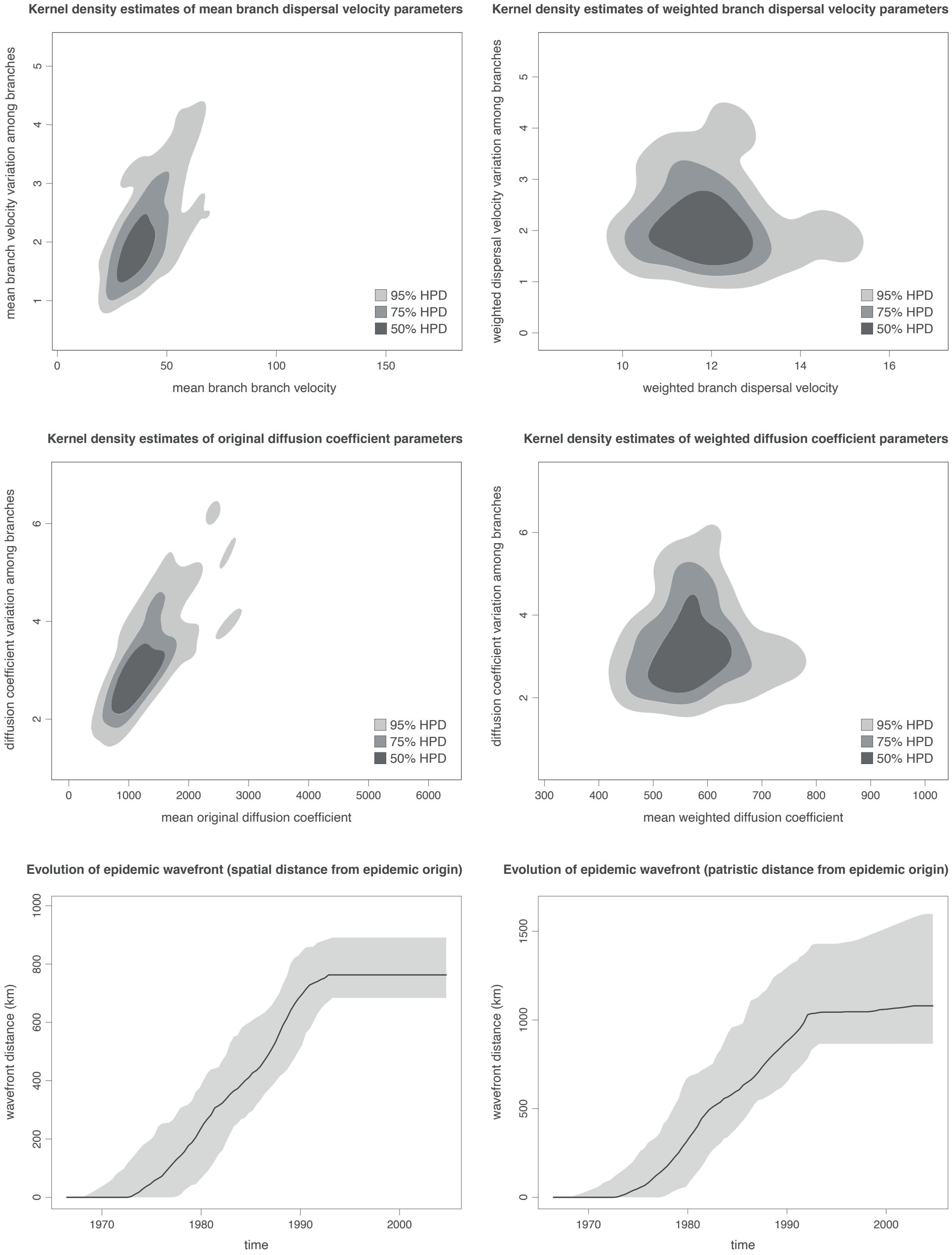

Figure 1: estimated dispersal statistics. For the four first graphs, the three contours show, in shades of decreasing darkness, the 50%, 75%, and 95% HPD regions via kernel density estimation. For the two last graphs, grey areas corresponds to 95% credible regions of the estimated wavefront position.

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 step. 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 path algorithm [9, 10] 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 [9, 10] or Circuitscape [11] 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, you 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 [12]. 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"
```

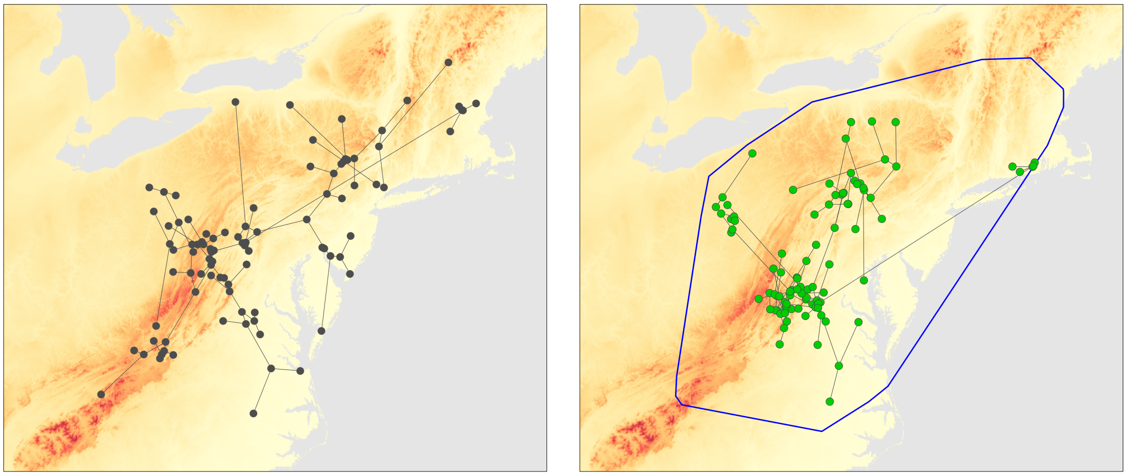

Figure 2: sampled and randomised trees mapped on the elevation raster. On the left: the original environmental raster (representing, in this case, elevation) upon which is superimposed the movement events extracted from one annotated tree sampled from the posterior distribution of trees obtained through 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 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 [9, 10]) or "3" (Circuitscape path model [11]). 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" coefficient of determination (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 coefficient of determination and the "spatial" coefficient of determination 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 "1". For instance, as we can see in Figure 3, 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 dispersal velocity of RABV lineages.

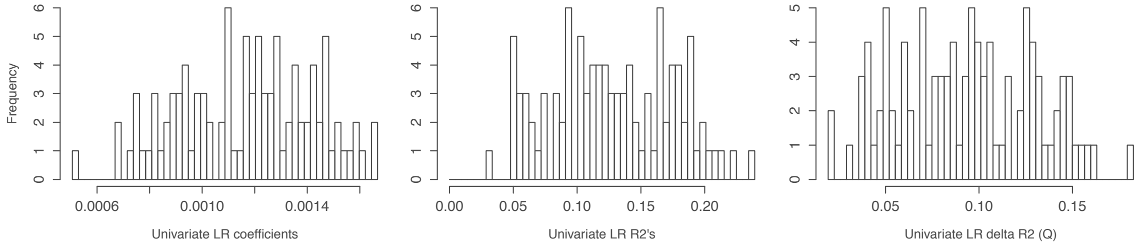

Figure 3: results of the linear regressions performed between dispersal durations and environmental distances computed with the least-cost path algorithm on the elevation raster. Each value on the histograms corresponds to one sampled tree.

A variable can only be considered as potentially explanatory if both its distribution of regression coefficients and associated Q's distribution are positive [13]. As we can see in Figure 3, 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[,"Univariate_LR_coefficients_elevation_R"]
> print(sum(LR_coefficients > 0))
```

```
100
```

```
> Qs = tab[,"Univariate_LR_delta_R2_elevation_R"]
> print(sum(Qs > 0))
```

```
100
```

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 dispersal rate of this RABV epidemic. We can now go to the next step (step 4) 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 third 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 4. 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 4: assessing the 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 the correlation 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 [8]. 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		

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] 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.

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

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

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

[12] 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.

[13] 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.

[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.