Tutorial for the R package `seraphim` 1.0

Transforming sampling coordinates to study the association between an environmental variable and the dispersal velocity of viral lineages

Simon Dellicour, Philippe Lemey

December 26, 2024

This tutorial describes how to use the R package "`seraphim`" [1] to transform sampling coordinates in order to study the association between a particular environmental variable (an "elevation" raster) and the dispersal/diffusion velocity of a rabies virus (RABV) epidemic in the North American raccoon population [2, 3]. See also the package manual for further detail on its different functions.

While the method presented in Dellicour *et al.* [3] and applied in the tutorial "Impact on diffusion velocity" is based on estimates obtained by phylogeographic inference, the method applied here aims at directly integrating the environment analysis prior to the Bayesian phylogeographic inference. Instead of assessing the association of environmental factors with phylogeographic estimates *a posteriori*, we will first measure environmental distances between our samples using the least-cost [4, 5] or Circuitscape path model [6]. The rational of this approach is that diffusion over distances that accommodate these factors should be more regular (characterised by less diffusion rate heterogeneity), or more Brownian, than over geographic distances that ignore these factors. Therefore, we will perform the phylogeographic diffusion analysis in a "new space" defined by these distances. One first possibility is to perform a multidimensional scaling (MDS) analysis based on pairwise environmental distances among sampling coordinates and consider the two main MDS dimensions as locations for each sequence in a bivariate diffusion model. A second alternative possibility is to perform a so-called "cartogram" transformation [7]. As illustrated in Figure 1, the principle of cartogram transformations is to transform the size of a series joint polygons according to values assigned to these polygons. Analogous to the MDS transformation, a cartogram transformation will increase the new Euclidean distance between localities associated with relatively high environmental distances, and decrease the new Euclidean distance between localities associated with relatively low environmental distances. We can then assess whether the diffusion process in the transformed space, i.e. defined by environmental distances (MDS transformation) or environmental values (cartogram transformation), indeed requires less heterogeneity as compared to a diffusion

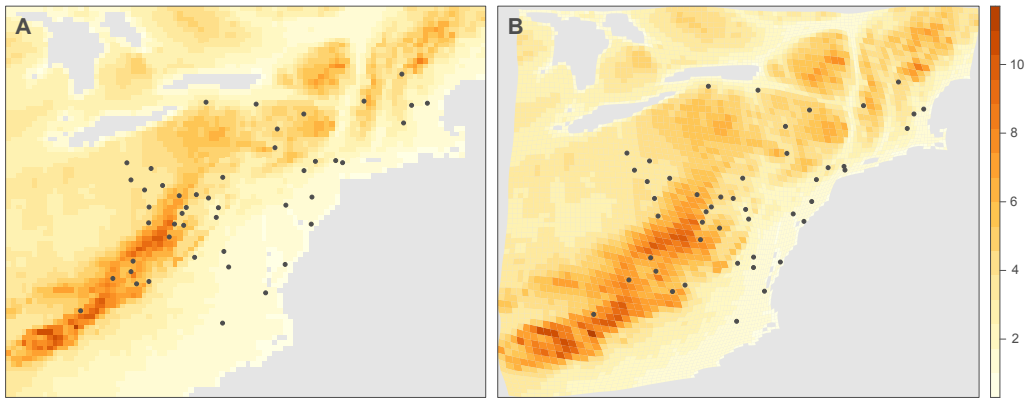

Figure 1: cartogram transformation (**B**) of an original elevation raster (**A**) along with the sampling coordinates of the raccoon rabies dataset (dark grey dots).

process in regular geographic space. For that purpose, we will adopt a relaxed random walk (RRW) process and quantify the heterogeneity of the lineage dispersal or diffusion velocity, i.e. the association between geographic distance and time duration associated with phylogeny branches. In this way, we can compare analyses based on distances defined by different environmental factors and compare the corresponding level of diffusion heterogeneity, the differences in which we will express as a Bayes factor support.

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) "Raccoon_rabies.fasta", a fasta file containing RABV sequences published by Biek and colleagues [2], and (ii) "Elevation_raster.asc", the environmental raster "elevation" for the study area.

Step 1: installing additional packages

In the context of this tutorial, two additional R packages have to be installed: the packages "`cartogram`" and "`sf`", both required to perform cartogram transformations in R. These two packages can be installed and loaded using the following command lines:

```
> install.packages("cartogram")
> install.packages("sf")
> library(cartogram)
> library(sf)
```

Step 2: rescaling the environmental raster

In this tutorial, we will test the impact of the "elevation" variable on the dispersal velocity of a RABV spread in the North American raccoon population. Before using the "elevation" raster to perform the MDS and cartogram transformations of the sampling coordinates, we will here (i) firstly modify its cell values so that they are rescaled between 0 and 10 and (ii) secondly add "1" to all the resulting cell values so that minimum raster cell values equal to one instead of zero:

```
> rast = raster("Elevation_raster.asc")
> k = 10
> M = max(rast[], na.rm=T)
> rast[!is.na(rast[])] = (rast[!is.na(rast[])]*(k/M))
> rast[!is.na(rast[])] = rast[!is.na(rast[])] + 1
```

The first operation is performed by defining the rescaling parameter k (here equal to 10) and avoids basing the MDS/cartogram transformation on high raster cell values. Of course, as the value assigned to this k parameter will have an impact on the outcome, several values should to be tested for this parameter (e.g., also 100 and 1000) in order to investigate its impact. Indeed, the 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. The second operation, i.e. adding "1" to all the cell values, aims to allow a comparison with an artificial raster with all the cell values equal to one. This "null" raster will be used to automatically compute environmental distances with the selected path model (least-cost or Circuitscape path model). The environmental distances computed on the "null" raster, which will be a proxy of the geographical distance associated with each branch, will then be used to automatically perform MDS/cartogram transformation. It is exactly this comparison of the outputs obtained from the transformations based on environmental distances computed on the "null" and environmental rasters that will allow us to analyse the impact of the environmental factor with the present approach (see below).

Step 3: first cartogram transformation test

Before using it to transform the sampling coordinates, a cartogram transformation of the environmental raster of interest can be conducted and visualised using the following lines of code:

```
> envVariable = rasterToPolygons(rast)
> names(envVariable) = "envVariable"
> envVariable = st_transform(st_as_sf(envVariable), 3857)
> cartogram = cartogram_cont(envVariable, "envVariable", itermax=5)
> plot(cartogram)
```

Note that to conduct a cartogram transformation based on an environmental raster, this raster has first to be transformed as a series of rectangular polygons (with one polygon per raster cell) using the function "rasterToPolygons()" and then to be projected using the function "st_transform()".

Step 4: performing MDS & cartogram transformations

MDS and cartogram transformation of sequence sampling coordinates can be performed with the"mdsTransformation()" and "cartogramTransformation()" functions, respectively. The "mdsTransformation()" function transforms sampling coordinates according to multi-dimensional scaling (MDS) analyses based on pairwise environmental distances computed among sampling points. A MDS analysis is performed for each specified environmental raster and original sampling coordinates are replaced by new coordinates found on two first MDS dimensions. Original sampling coordinates have to be contained in sequence names available in a fasta file or in a tab-delimited text file only containing these coordinates (in the later case, the first and second columns of the tab-delimited text file have to contain the genomic sequence IDs and collection dates, respectively, then followed by a third and fourth columns respectively listing the sampling latitude and longitude values). The "cartogramTransformation()" function first performs cartogram transformation of environmental rasters based on their cell values. As detailed above, cartogram transformations are performed using the "cartogram_cont()" function of the package "`cartogram`" and while considering each raster cell as a distinct polygon for which an environmental value is assigned (i.e. the original raster cell value). In a second step and for each cartogram transformation that has been performed, this function then modifies sampling geographic coordinates accordingly. As for the "mdsTransformation()" function, these sampling coordinates have to be contained in sequence names available in a fasta file or in a tab-delimited text file.

In the context of this tutorial, environmental distances will be computed for each branch using the least-cost path model [4, 5] and the "elevation" raster will be tested as a potential resistance factor. The different parameters of the "mdsTransformation()" and "cartogramTransformation()" functions have to be specified as follows:

```
> input = read.dna("Raccoon_rabies.fasta", format="fasta")
> envVariables = list(rast)
> resistances = c(TRUE)
> avgResistances = c(TRUE)
> fourCells = FALSE
> pathModel = 2
> outputName = "RABV"
> OS = "Unix"
```

The "input" object has to be either a list of sequences in a fasta format or a tab-delimited text file containing the geographic coordinates to transform. In the first case, the fasta object has to be obtained by reading a fasta file with the "read.dna()" function of the "ape" package: read.dna("file_name.fasta", format="fasta"). In the second case, the text file has to contain only four columns: (i) the genomic sequence IDs, (ii) their collection dates, and their sampling (iii) latitude and (iv) longitude coordinates to transform. 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 "fourCells" parameter specifies if a given raster cell should be connected to either its four first-order or eight first- and second-order neighbouring cells (see the "`gdistance`" package implementing the least-cost path algorithm or Circuitscape program manuals for further details). The "avgResistance" 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 "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 [4, 5]) or "3" (Circuitscape path model [6]). The "outputName" string will be used as a prefix to name the different outputs of the function. The information about the nature of the operating system is only useful when the function has to call the "Circuitscape" Python package [6] (see the `seraphim` manual for further details).

```
> mdsTransformation(input, envVariables, pathModel, resistances,
avgResistances, fourCells, outputName, OS)
> cartogramTransformation(input, envVariables, resistances, outputName)
```

For each environmental raster that is provided, the "mdsTransformation()" function first computes pairwise environmental distances among all sampling coordinates. As stated above, these pairwise environmental distances are then used to perform MDS analyses and transform the original geographic coordinates by taking new values on the two main MDS dimensions. Depending on the type of input file, new fasta or text files are generated, one per MDS analysis that is performed. If new fasta files are generated, only the sequence names are modified: longitude and latitude coordinates are updated according to the MDS transformation and thus indicate the new position of each sampled sequence in the transformed space. If text files are generated, these files will then contain the transformed coordinates. Similarly, the "cartogramTransformation()" function also performs one cartogram transformation per provided environmental raster and creates a new fasta/text file per cartogram transformation. In these files, longitude and latitude coordinates are thus updated according to the cartogram transformation and thus indicate the new position of each sampled sequence in the transformed space. Note that in this tutorial, the input file is in a fasta format.

Step 5: running a continuous phylogeographic analysis using BEAST and based on each generated fasta file

The fourth step is to generate a BEAST ".xml" file for each distinct fasta file generated by the "mdsTransformation()" and "cartogramTransformation()" functions. Initial versions

of these input files can be generated in BEAUti as for a continuous phylogeographic analysis [8] by specifying a RRW diffusion model (Cauchy, gamma or lognormal) for the partition of latitude and longitude values (see the related BEAST tutorial for further details on the BEAUti settings for this analysis). 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 [8] implemented in the software package BEAST 1.10 [**?**] is available on the BEAST community website (`https://beast.community/workshop_continuous_diffusion_yfv`). After having generated a first version of these different ".xml" files with BEAUti, the estimation of the "distTime.correlation" statistic has to be manually added in each ".xml" file:

```
<continuousDiffusionStatistic id="distTime.correlation"
statistic="distanceTimeCorrelation" mode="correlationCoefficient"
greatCircleDistance="false">
<multivariateTraitLikelihood idref="location.traitLikelihood"/>
</continuousDiffusionStatistic>
```

In addition, the ".xml" files also have to be modified to ask for logging the correlation values:

```
<continuousDiffusionStatistic idref="distTime.correlation"/>
```

See also the ".xml" examples files provided along this tutorial.

Step 6: estimating Bayes factor values by comparing the log files

The last step is to investigate the association between the lineage dispersal durations and distances. This can for instance be done by comparing the "distTime.correlation" values, i.e. the correlations between the geographic distance and duration (in time units) associated with each phylogeny branch (one value per logged tree), among the different BEAST log files. Two distinct comparisons have to be performed, one related to the MDS transformation and one related to the cartogram transformation. The first comparison will be between the log file obtained from the BEAST analysis based on sampling coordinates modified according to the MDS transformation based on the environmental (here "elevation") raster and the log file obtained from the BEAST analysis based on sampling coordinates modified according to the MDS transformation based on the "null" raster. The second comparison will be between the log file obtained from the BEAST analysis based on sampling coordinates modified according to the cartogram transformation based on the environmental (here "elevation") raster and the log file obtained from the BEAST analysis based on original sampling coordinates. In practice, these two comparisons are performed by estimating Bayes factors (BFs), with one BF value returned per comparison. For the MDS transformation case, the BF_e for a particular environmental factor e is defined by the posterior odds or that $distTime.correlation_{mdsNullRaster} < distTime.correlation_{mdsEnvRaster}$ divided by the equivalent prior odds (the prior probability for $distTime.correlation_{mdsNullRaster} < distTime.correlation_{mdsEnvRaster}$ 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 $distTime.correlation_{mdsNullRaster} < distTime.correlation_{mdsEnvRaster}$, i.e. the frequency at which $distTime.correlation_{mdsNullRaster} < distTime.correlation_{mdsEnvRaster}$ in the sampled posterior distribution. See equation (6) in Lemey *et al.* [9] for a similar approach. Similarly, for the cartogram transformation case, the BF_e for a particular environmental factor e is defined by the posterior odds or that $distTime.correlation_{originalCoordinates} < distTime.correlation_{cartogram}$ divided by the equivalent prior odds (the prior probability for $distTime.correlation_{originalCoordinates} < distTime.correlation_{cartogram}$ is considered to be 0.5). BFs can be estimated in R with the following commands:

```
> cor1 = read.table("Raccoon_rabies_res/RABV_MDS_nullRaster_LC.log",
header=T)[201:2200,"distTime.correlation"]
> cor2 = read.table("Raccoon_rabies_res/RABV_MDS_elevation_LC.log",
header=T)
> d = cor2-cor1
> p = (sum(d>0))/length(d)
> BF_MDS = p/(1-p); BF_MDS

28.41

> cor1 = read.table("Raccoon_rabies_res/RABV_original_coordinates.log",
header=T)[201:2200,"distTime.correlation"]
> cor2 = read.table("Raccoon_rabies_res/RABV_cartogram_elevation.log",
header=T)[201:2200,"distTime.correlation"]
> d = cor2-cor1
> p = (sum(d>0))/length(d)
> BF_cartogram = p/(1-p); BF_cartogram

5.12
```

As detailed above, we here discarded the burn-in from the BFs' estimations. In the case of the elevation raster tested as a resistance factor in the context of this tutorial, the BF is >20 and then considered as a "strong" evidence of the impact of that particular environmental factor on the dispersal velocity (see Table 1 for the scale of interpretation of Bayes factor values).

Note that in the present example, the association between the lineage dispersal durations and distances is investigated through the computation (directly by the software package BEAST) of a correlation between the branch durations and geographic distances, which thus assesses the homogeneity of the lineage dispersal velocity. Given that such a lineage dispersal metric has been identified as less robust than than a diffusion coefficient metric to the sampling intensity (i.e. the sampling site) [9], it is actually better to instead favor an investigation of the homogeneity of the branch diffusion coefficient. This can for instance be done through the comparison of post hoc linear regressions performed between four times the dispersal durations ($4t$) and the squared Euclidean distances (d^2) computed in

the different transformed spaces – see the analyses detailed in the following GitHub repo for an example: `https://github.com/sdellicour/landscapephylogeography`.

Table 1: scale of interpretation of Bayes factors (BF) according to Jeffreys [10] and Kass & Raftery [11].

Scale of interpretation defined by Jeffreys [10]			Scale of Kass & Raftery [11]	
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, Faria N, Lemey P, Pybus OG (2016b). SERAPHIM: studying environmental rasters and phylogenetically-informed movements. *Bioinformatics* 32 (20): 3204-3206.

[2] 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.

[3] 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.

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

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

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

[7] Dougenik JA, Chrisman NR, Niemeyer DR (1985). An algorithm to construct continuous area cartograms. *The Professional Geographer* 37: 75-81.

[8] 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.

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

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

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

[12] 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.