DOI: 10.1111/1755-0998.12697

Article Type: Resource Article

EcoGenetics: an R package for the management and exploratory analysis of spatial data in landscape genetics

Leandro G. Roser[1], Laura I. Ferreyra[2,4], Beatriz O. Saidman[3,4] and Juan C. Vilardi[2,4]

[1] Universidad Nacional de San Martín. Instituto de Investigaciones Biotecnológicas (IIB-INTECH). Buenos Aires, Argentina.

[2] Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Departamento Ecología, Genética y Evolución. Genética de Poblaciones Aplicada (GPA). Buenos Aires, Argentina.

[3] Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Departamento Ecología, Genética y Evolución. Genética de Especies Leñosas (GEEL). Buenos Aires, Argentina.

[4] CONICET - Universidad de Buenos Aires. Instituto de Ecología, Genética y Evolución (IEGEBA). Buenos Aires, Argentina.

Correspondence: Leandro Roser, Fax: (5411)4006-1559; E-mail: learoser@gmail.com

EcoGenetics: R package for landscape genetics

Abstract

The integration of ecology and genetics has become established in recent decades, in hand with the development of new technologies, whose implementation is allowing an improvement of the tools used for data analysis. In a landscape genetics context, integrative management of population information from different sources can make spatial studies involving phenotypic, genotypic and environmental data simpler, more accessible and faster. Tools for exploratory analysis of autocorrelation can help to uncover the spatial genetic structure of populations and generate appropriate hypotheses in searching for possible causes and consequences of their spatial processes. This paper presents *EcoGenetics*, an R package with tools for multi-source management and exploratory analysis in landscape genetics.

Keywords: ecology, genetics, landscape genetics, package, R, spatial autocorrelation

Introduction

From the second half of the 20th century, knowledge about populations has been growing due to important technological and theoretical advances. The information age, associated with the massive use of computers, has brought significant changes in science, with generation of large amounts of data, an increased capability for data visualization, exploration and processing, and the development of computational statistics, which have a direct impact on research activities. These changes have resulted in the emergence of landscape genetics (Manel *et al*. 2003), a discipline that amalgamates population genetics, landscape ecology and geostatistics (Storfer *et al*. 2007).

A typical workflow in landscape genetics (Fig. 1) utilizes a series of steps where data are 1) imported into a suitable software, 2) stored, organized and formatted, 3) explored, transformed and visualized, 4) modeled, and finally 5) communicated. The organization of the data is an activity that demands an important portion of the researcher's time as often 80% of data analysis is spent on the process of cleaning and preparing the information (Dasu & Johnson 2003; Wickham 2014). The generation of clean and organized datasets is a fundamentally important task during the analysis process. The R Project for Statistical Computing (R Core Team 2016) provides a powerful environment to accomplish this purpose.

Exploration of spatial information of populations often relies on assumptions that do not fit the raw data well. Spatial and temporal structuring, two fundamental components in the functioning of ecosystems (Levin 2000), are quantitatively manifested in correlations between observations of a variable, violating the independence required by standard statistical tests. This phenomenon of spatial dependence is caused by a combination of endogenous and exogenous population processes. Spatial patterns in observations generated by endogenous processes, such as ecological drift and random dispersal (Legendre & Legendre 2012), appear as "Spatial Autocorrelation" (SA) in the data. "Autocorrelation" refers to the fact that a variable is correlated with itself (auto), while "spatial" indicates that this correlation depends on the location of observations. Spatially autocorrelated observations are characterized by being more or less similar (SA positive and negative, respectively) than would expected by chance (Fortin & Dale 2005; Legendre 1993). Spatial dependence promoted by exogenous processes occurs when a particular variable reflects the structure of one or more variables that are themselves autocorrelated. For example a spatial pattern in plants might be related to a moisture gradient (Fortin & Dale 2005; Legendre & Legendre 2012). Population geneticists work on the basis that spatial distribution of genetic variation often differ strongly from randomness or uniformity (Epperson 1993). This condition is captured by the term "Spatial Genetic Structure"

(SGS), defined as the non-random distribution of genetic variability in space. SA is an underlying foundation in population ecology and genetics, therefore, development of theory, methods and tools for assessing SA is a common objective to both disciplines.

Global SA statistics and correlograms have become popular tools among population geneticists and ecologists after the publications of Sokal & Oden (1978a, b). Many research used correlograms to evaluate the relation between genetic similarity and geographical distance (Arnaud 2003; Loiselle 1995; Smouse & Peakall 1999). SA methods were particularly applied to the analysis of fine-scale SGS (Hardy & Vekemans 1999; Vekemans & Hardy 2004). Local SA statistics as the LISA (Anselin 1995) and Getis-Ord's G_i and G_i* (Getis and Ord 1992; Ord and Getis 1998) are relatively recent tools that allow to discover local clusters of autocorrelated observations (Anselin 1995) using alleles or other variables.

The need of integrating geographic, ecological and genetic data requires a platform for the joint management of information from different sources. This paper presents a package for landscape genetics called *EcoGenetics*, designed under the R language and statistical environment. The aim of the package is to provide flexible tools using a SA analytic framework to integrate, manage and explore spatially explicit population data, tailored to ecological and genetics research. With the definition of a new class of object for multi-source data storage and manipulation (`ecogen`), and a set of exploratory functions, the package facilitates the analysis of single and multiple variables from a broad range of data sources, in an attractive data visualization environment.

Storing and organizing multi-source population data: ecogen objects

EcoGenetics was constructed under the object-oriented S4 system of R. The "ecogen" class, a central element of the package, was designed for efficient and straightforward handling of multi-source information in the different stages of analysis. An object of this class is a data structure that behaves like an ordered "stack" of information. Each layer of the stack represents data from a different source stored in a slot, as described in Fig. 2A. Table 1 summarizes the main functions available for `ecogen` objects, which are overviewed in the following sections. A tutorial based on examples is available online in GitHub at *https://leandroroser.github.io/EcoGenetics-Tutorial.*

Construction of ecogen objects

First, with the command `data(eco.test)`, five data frames and one `ecogen` object (`eco`) are added into the workspace:

```
> library("EcoGenetics")
> data(eco.test)
> ls()
[1] "coordinates" "eco" "environment" "genotype" "phenotype" "structure"
```

Each data frame in the workspace represents information from a different source for 225 individuals: geographic coordinates (`coordinates`), codominant genetic markers (`genotype`), phenotypic variables (`phenotype`), environmental variables (`environment`), and population structure information (`structure`). The object `eco` was constructed with this set of data frames, using the `ecogen` constructor as follows:

```
> eco <- ecogen(XY = coordinates, P = phenotype, G = genotype, E = environment, S = structure, type = "codominant", order.G = TRUE)
> eco # The object is shown as a panel in the console (Fig. 2B)
```

The construction of a valid `ecogen` object requires that all data frames with information from different sources have the same structure: columns represent different variables (according to data source), rows represent individuals, which must be in the same order, and row names identify the corresponding individuals. The constructor is flexible and offer options for data ordering and row names assignment, which are described in the online tutorial. Both codominant/dominant markers are accepted as genetic data. The package uses an internal *genind* object (Jombart, 2008), modified to work as a transitional data structure between data frames and the content of the `G` (genotypes) and `A` (alleles) data slots. Arguments available for the importation of genetic data are detailed in the *EcoGenetics* documentation. Specifically a description of the structure of an `ecogen` object can be obtained with the `help("ecogen")` command.

Use of accessor functions with ecogen objects

All the S4 objects of *EcoGenetics* have a set of accessor functions assigned, whose role is to get and set the content of the slots. Accessors have the following notation: a prefix (`ecoslot.`) followed by the name of the corresponding slot plus the name of the ecogen object in parentheses. For the object `eco` of the example, the corresponding accessors are: `ecoslot.XY(eco)`, `ecoslot.P(eco)`, and so on (Table 1). The correct assignment of content to the slot of an existent object is made with accessors; these special functions ensure a basic pre-processing and checking of the data when used in assignment operations. The object of the previous section can also be obtained with an approach based in accessors:

```
> eco.temp <- ecogen()
> ecoslot.XY(eco.temp) <- coordinates; ecoslot.P(eco.temp) <- phenotype
> ecoslot.E(eco.temp) <- environment; ecoslot.S(eco.temp) <- structure
> ecoslot.G(eco.temp, order.G = TRUE) <- genotype # ecoslot.G "set" mode
```

`Note: ordered genotypes in slot G`

Accessors use is detailed in the online tutorial.

Algebra of ecogen objects

A set of operations are defined for the class "ecogen" with the purpose of multi-source data manipulation. These operations can be classified into "subset", "split" and "combine" methods (Fig. 3 and Table 1, Manipulation functions). Other useful standard functions are defined in Table 1.

Conversion of ecogen objects from/to other formats

EcoGenetics is able to import and export `ecogen` objects from/to other data formats, as listed in Table 1. The functions `eco.convert` and `eco.format` help to perform several operations with genetic data for easy conversion into other formats. Conversion operations are detailed in the online tutorial.

Interactive data exploration

EcoGenetics has main and auxiliary functions for data exploration. The core for exploratory analysis consists of a family of six main functions, as described in Table 2. Some functions compute several related statistics of the listed analyses for multiple variables. The set of characteristics provided by the package avoids the need to use different programs and loops for a different statistic or with multiple variables, thus reducing the number of programs required for a similar task. The several functions of the package have original methods for the presentation and extraction of the results. Plot methods make extensive use of the *ggplot2* package (Wickham 2009) and JavaScript-based packages for interactive data visualization, as *plotly* (Sievert *et al*. 2016) for an interactive extension of *ggplot2* graphs. The following sections describe the different tools of the package.

Spatial weights

Spatial weights matrices are central elements in SA analysis. A spatial weights matrix W is a square positive matrix that defines the strength of the spatial relations between observations, assigning the value w_{ij} to the connection between individuals (i, j). In a binary weighting scheme, W corresponds to an adjacency matrix ("connection network") indicating if the individual pairs (i, j) are connected (w_{ij} = 1) or not (w_{ij} = 0). In other situations, W is a matrix assigning a value to the connection (i, j) using a model for the spatial relations (*e.g.*, following an exponential decay with distance, up to a threshold distance d, where the weights are set to 0). Spatial weights matrices are obtained in *EcoGenetics* using the function `eco.weight`. Several weights construction methods are available in the function. Different plotting methods, interactive and non-interactive, are also available for weight objects. These aspects are detailed in the online tutorial.

Correlogram analysis

A correlogram is a plot for a correlation coefficient as a function of the inter-individual distance. The definition of SA implies that "nearby observations are more or less similar than expectations by chance". In a situation where a spatial pattern shows positive SA (as happens with many biological processes) the values of autocorrelation will tend to decrease with distance. The correlogram can then be used to characterize the spatial pattern. A trend from positive to negative values indicates a gradient in the data; fluctuation around the expected value of the statistic indicates patchiness. A description of other patterns can be found in Fortin & Dale (2005). Omnidirectional correlograms are constructed without taking into account a particular spatial direction. This standard method assumes that the autocorrelation patterns vary similarly with distance in all directions (isotropy). *EcoGenetics* also includes an approach to explore patterns that vary with direction (anisotropy) by the "bearing correlogram" method (Rosenberg 2000). The method can be used to construct directional correlograms and explore whether the data is likely to hold the isotropic assumption or not. For the creation of bearing correlograms, the weights matrices used in the analysis are rescaled by a factor that varies between 0 and 1, related with the direction pointed by the vector *v* connecting each pair of individuals (*i*, *j*). Each w_{ij} of the weights matrix *W* is recomputed as $w'_{ij} = w_{ij} \cos^2(\alpha_{ij} - \Theta)$, where α_{ij} is the angle that *v* forms with the positive *x* axis (due East) in counterclockwise direction, and Θ the angle of the reference vector pointing in the direction of analysis, also with respect to due East. When $\alpha_{ij} = \Theta$, w_{ij} is weighted by 1 and $w'_{ij} = w_{ij}$, On the contrary, when $\alpha_{ij} = \Theta \pm \pi/2$, w_{ij} is weighted by 0 and $w'_{ij} = 0$.

For single variables, the package supports the construction of correlograms based in Moran's *I* (Moran 1950), Geary's *C* (Geary 1954) and Bivariate Moran's *Ixy* (Reich *et al.* 1994) statistics (Table 2) by means of the function `eco.correlog`. Multivariate approaches for phenotypic traits can be obtained with Mantel and partial Mantel correlograms (Oden & Sokal 1986; Sokal 1986) using the function `eco.cormantel`. A multivariate method for genetic data is available with the function `eco.malecot`. Default options are set for codominant markers, using a kinship matrix based on Nason's F_{ij} (Loiselle *et al*. 1995). A custom kinship matrix for codominant/dominant markers can also be imported. A plot can be obtained for all the standard correlograms with the function `eco.plotCorrelog` (Fig. 4A). Two types of output format can be selected for `eco.correlog` and `eco.cormantel`, when used to construct bearing correlograms for several successive angles. Angles can be fixed, and for each one a table is constructed containing distances and values of the statistic in columns (the independent and dependent variables of standard correlograms, respectively). In the second format, distance classes are fixed and for each one a table is constructed with angles and values of the statistic in columns. In this latter case, an angular correlogram or Bearing Plot (Falsetti

& Sokal, 1993; Rosenberg 2000) of the statistic in function of the successive angles can be constructed for each distance class with the function `eco.plotCorrelogB (`Fig. 4B).

Global SA analysis

Global statistics allow a global survey of the presence of SA in a data set. For uni- and bivariate approaches, *EcoGenetics* is able to compute and test the Moran's *I*, Geary's *C*, Join-Count and Bivariate Moran's *Ixy* statistics with the function `eco.gsa` (Table 2 and Fig. 4C). Multivariate methods are based on Mantel (Mantel 1967) and partial Mantel (Smouse *et al*. 1986) statistics (function `eco.mantel`). The use of Mantel test, widely adopted among population geneticists, is currently under active debate (Legendre & Fortin 2010; Guillot & Rousset 2013; Legendre *et al*. 2015). Mantel test assess the hypothesis of absence of relationship between values in two dissimilarity matrices. Alternatives to this method are discussed in Legendre *et al*. (2015), who showed with simulated data that the power of Mantel test for detecting SA is low. This paper also indicated that regression using dbMEM (distance based Moran Eigenvectors Maps) should be more powerful than a Mantel test conducted with dissimilarity matrices for modeling the relationship between a response dataset (as a genetic matrix) and the geographical distance. *EcoGenetics* provides the possibility of performing a Mantel test with truncated distance matrices, an alternative with higher power than the classical Mantel test when there is a specific ecological or genetic dispersal model in mind (Legendre *et al*. 2015). A model can be proposed for example in a situation where the effect of distance among sites can only be perceived up to a certain distance where contagion, dispersal of propagules in plants, or migration in animals, no longer creates spatial correlation (Legendre *et al*. 2015). The function `eco.mantel` also accepts a weights object obtained with the function `eco.bearing`, which generates a directional weights object to compute a bearing Mantel test as performed by Falsetti & Sokal (1993). See the online tutorial for examples.

Local SA analysis

Local SA analysis is based on the computation of local SA statistics to study the similarity of each individual with its neighbors. This methodology allows to discover local clusters of autocorrelated observations ("hot" and "cold" spots, Anselin 1995) and can provide maps showing how SA varies geographically (Sokal & Thomson 2006). In addition, for Moran's *I* and Geary's *C* the local values represent a decomposition of the corresponding global statistic (Anselin 1995). This decomposition enables the identification of those groups of individuals that contribute most to the global analysis (Sokal & Thomson 2006). For each individual, a SA statistic is obtained using a weights object specifying its spatial relationship with others. Local SA statistics (Table 2) are computed and tested with the function `eco.lsa`. Plots for single or multiple variables can be obtained with the function `eco.plotLocal (`Figs 4D,E).

Integration of *EcoGenetics* in the R ecosystem

Using ecogen objects to analyze the relations among data from different sources

Different methods have been developed in several R packages for modeling multivariate multi-source data, including trend-surface analysis (Legendre 1990), distance-based Moran's eigenvector maps (dbMEM, Borcard & Legendre 2002; Borcard *et al*. 2004; Dray *et al*. 2006) or asymmetric eigenvector maps (AEM, Blanchet *et al*. 2008). The analyses are carried out by using geographic functions derived from points coordinates as explanatory variables in multiple regression, multi-scale ordination, canonical analysis or variation partitioning among environmental and spatial components (Borcard *et al*. 1992; Borcard & Legendre 1994; Wagner 2004). The packages *adespatial* (Dray *et al.* 2016) and *vegan* (Oksanen *et al.* 2016) offer a set of tools in R to work around these tasks. The use of `ecogen` objects combined with the function `eco.formula`, included in the package, can ease significantly the work with these tools. This function can create complex expressions with `ecogen` objects, acting as a proxy between the variables stored in `ecogen` objects and any other function able to use a formula as argument. The next examples illustrate the use of `eco.formula` with the function `rda` of the package vegan. The arguments that `rda` can take are `X` (a matrix of response variables, *e.g.*, phenotypic traits), `Y` (predictor variables, *e.g.*, environmental variables or alleles), and `Z` (conditioning variables, *e.g.*, geographic coordinates or dbMEM's). In its simplest version, when `Y` and `Z` are missing, the function performs a principal components analysis. If `Y` and `Z` are provided, the function performs a redundancy analysis.

```
> require(adespatial); require(vegan)

> # PCA analysis

> pc <-rda(ecoslot.P(eco), scale=TRUE)

> # RDA analysis, using dbMEMs as conditioning variables. First compute

> # dbMEMs for eco in adespatial and store the result in slot C

> eco_distance <- dist(ecoslot.XY(eco)) # create distance matrix

> ecoslot.C(eco, use.object.names = TRUE) <- dbmem(eco_distance, thresh
= 1, MEM.autocor = "positive") # compute dbMEMs and store in slot C

> # Perform a RDA with vegan, using the first 20 dbMEMs, and the function
eco.formula. U() is an auxiliary function interpreted by eco.formula; it
includes in the formula all the variables of the slot within parentheses
(see the help file of eco.formula for details). The function "Condition" is
not interpreted by eco.formula: it is used to pass as conditional variables
the dbMEMs between parentheses to "rda"

> my_formula <- eco.formula(eco, P1 + P2 + P3 ~ E1 + E2 + U(A) +
Condition(U(C[, 1:20])))

> my_formula # watch formula content

> rda(my_formula)
```

Using ecogen objects and EcoGenetics methods in interaction with population genetics packages

Genetic population analyses can be performed by several R packages. The package *adegenet* (Jombart 2008) defines the "genind" class to manipulate genetic data and includes methods for population genetics and multivariate analysis. From a `genind` object, *adegenet* can construct connection networks (*e.g.* Delaunay triangulation) that can be imported into *EcoGenetics* as `eco.weight` objects. The packages *gstudio* and *popgraph* (Dyer 2014) offer tools to work around exploratory analysis of population genetic data, with a generalized *ggplot2* environment. The package *hierfstat* (Goudet & Jombart 2015) allows estimating hierarchical *F*-statistics and basic stats. The package *poppr* (Kamvar *et al.* 2014) has an interface to the `amova` functions of *ade4* (Dray & Dufour 2007) and *pegas* (Paradis 2010), and provides methods to obtain genetic distances between individuals and the construction of minimum spanning networks. Examples of combined operations using *EcoGenetics* and functions of these packages are described in the online tutorial.

Testing

In addition to testing throughout development, *EcoGenetics* exploratory functions were cross-compared with other available programs to ensure a correct performance, including the R packages *spdep* (Bivand & Piras 2015), *vegan* (Oksanen *et al*. 2016), *ecodist* (Goslee & Urban 2007), *ncf* (Bjornstad *et al*. 2016) and the Python spatial library *PySAL* (Rey & Anselin 2010). A series of benchmark tests (Supplemental Materials) were conducted in R using a laptop with Linux, a 2.20GHz Intel Core i7 CPU and 16GB of 1600MHz RAM. The tests of core functions of the package showed a performance comparable to other R programs (Supplemental Materials). For most population datasets, processing should be fast on a PC or laptop.

Conclusion

EcoGenetics contains tools for the integration of multi-source data and a generalized framework for exploratory SA analysis, with a cutting-edge data visualization environment. The package provides capabilities for analysis of several variables in many of its routines and a high flexibility during the analysis to support different configurations. Given the need of data integration, fast and easy manipulation of multi-source datasets, and adequate exploratory methods with the increasing rate in information volume, we expect our package to be widely useful in population genetics and ecology research.

Data accessibility

The stable version of the package and the reference manual are available in the CRAN repository (*https://cran.r-project.org/web/packages/EcoGenetics*). The sample datasets used in this work are included in the package. An introductory tutorial is available in GitHub at *https://leandroroser.github.io/EcoGenetics-Tutorial/*.

Supporting Information

File S1 R script to run the benchmark tests performed in this paper. The program generates an HTML report with the running time of the tested functions.

File S2 Non-parallel version of the vegan function "mantel_correlog".

File S3 Elapsed time (seconds) for the benchmark tests.

File S4 PDF version of the online tutorial.

Table 1 A summary of the functions available for `ecogen` objects. The first column ("Global objective") indicates the main purpose of the group of corresponding functions, with the specific objectives in the second column.

Global objective	Specific objective		Function name
Configuration	Constructor		`ecogen`
	Set `ecogen` slots		Data store: `ecoslot.P<-`, `ecoslot.G<-`, `ecoslot.E<-`, `ecoslot.S<-` Results store: `ecoslot.OUT<-`
	Set `ecogen` names		`names<-`
Access	Data slots: using accessors		`Data store: ecoslot.P, ecoslot.G, ecoslot.A, ecoslot.E, ecoslot.S, ecoslot.C`
	Data slots: using brackets		"`[[`", using "P", "G", "E", "A", "S" (*e.g.*, `eco[["P"]]`)
	Result slot: using accessors		Results store: `ecoslot.OUT`
	Result slot: using brackets		"`[[`", using "OUT"
	Creation of formula with elements in object		`eco.formula`
Manipulation	**Subset**	Integer subset	"`[`"
		Logical subset	
		Subset by group	`eco.subset`
	Split	Split into list of `ecogen` objects	`eco.split`
		Split into `ecogen` objects in workspace	`eco.split`
	Combine	Bind by row	`eco.rbind`
		Bind by column	`eco.cbind`
		Merge objects	`eco.merge`
Visualization	Show object		`show`
Description	Object names		`names`
	Number of rows in slots		`nrow`
	Number of columns in slots		`ncol`
	Dimension of object		`dim`
	Check if object is of class "ecogen"		`is.ecogen`
	Coercion to list		`as.list`
Conversion	**Import**	**Export**	
	genind	genind (Jombart 2008)	`ecogen2genind` / `genind2ecogen`
	gstudio	Gstudio (Dyer 2016)	`gstudio2ecogen` / `ecogen2gstudio`
	Genepop	Genepop (Rousset 2008)	`genepop2ecogen` / `ecogen2genepop`
	SPAGeDi	SPAGeDi (Hardy & Vekemans 2002)	`spagedi2ecogen` / `ecogen2spagedi`
	---	Hierfstat (Goudet & Jombart 2015)	`ecogen2hierfstat`
	---	Geneland (Guillot *et al.* 2005)	`ecogen2geneland`

Table 2 Family of SA analysis functions included in *EcoGenetics*. The functions are based in one of two approaches: (1) a single statistic or (2) a correlogram. In the first case, the statistic can be computed globally across all samples, or locally for each sample and a set of relations with others, as determined by a connection network. In the second case, a statistic is recursively computed over a set of intervals. Supported data for each analysis are indicated in reference to the slots of ecogen objects: `P` (phenotypic variables), `G` (genotypes, binary data frame for dominant data, complete genotypes for codominant markers), `A` (counts per allele for codominant markers [0, 1, 2]), and `E` (environmental variables). Join-count analysis supports categorical data.

Function	Statistic name	Reference	# Var[1]	Sup. data[2]	Approach[3]	Method	Plot function
`eco.gsa`	Join-count	Moran 1948; Cliff & Ord 1981	U	*G*(dom/codom), *A* (codom), categorical data	SS	Global	Incorporated (univariate) / `eco.plotGlobal` (multivariate)
	Moran's *I*	Moran 1950		*P*, *G* (dom), *A* (codom), *E*			
	Geary's *C*	Geary 1954					
	Bivariate Moran's I_{xy}	Reich *et al*. 1994	B				
`eco.lsa`	local Moran's *I*	Anselin 1995	U	*P*, *G* (dom), *A* (codom), *E*	SS	Local	`eco.plotLocal`
	local Geary's *C*	Anselin 1995					
	local Getis-Ord's G_i and G_i*	Getis & Ord 1992; Ord & Getis 1995					
`eco.mantel`	ordinary Mantel test	Mantel 1967	M	*P*, *G* (dom), *A* (codom), *E*	SS	Global	Incorporated
	partial Mantel test	Smouse *et al*. 1986					
`eco.correlog`	Moran's *I*	Moran 1950; Sokal & Oden 1978	U	*P*, *G* (dom), *A* (codom), *E*	ODC / DC	Global	`eco.plotCorrelog` / `eco.plotCorrelogB`
	Geary's *C*	Geary 1954; Sokal & Oden 1978					
	Bivariate Moran's I_{xy}	Reich *et al*. 1994	B				
`eco.cormantel`	ordinary Mantel test	Oden & Sokal 1986	M	*P*, *G* (dom), *A* (codom), *E*	ODC / DC	Global	`eco.plotCorrelog` / `eco.plotCorrelogB`
	partial Mantel test	Oden & Sokal 1986					
`eco.malecot`	Loiselle's kinship coefficent (F_{ij}) – custom statistic	Loiselle *et al*. 1995; Kalisz *et al*. 2001; Born *et al*. 2012	M	*G* (dom), *A* (codom)	ODC / DC	Global	`eco.plotCorrelog`

1. Number of variables. *U* univariate, B bivariate, *M* multivariate
2. Supported data. *codom* codominant markers, *dom* dominant markers
3. *SS* single statistic, *ODC* omnidirectional correlogram, *DC* directional correlogram (Rosenberg 2000)

Fig. 1 A typical population data analysis routine. The representation is inspired in the Chapter 1.2 of Wickham and Grolemund (2016). Solid lines indicate the main direction of the pipeline, and striped lines other usual directions (as exportation and reorganization of the data after exploration). The routine usually requires several cycles of data exploration and modeling (gray box) before the communication of results.

Fig. 2 Structure of an `ecogen` object. (A) `ecogen` data panel, as displayed in the R console. The sample data shown ("eco", included in the package), is composed of a set of 255 individuals. The panel indicates the dimensions of each slot and additional data. The data frames constituting the content of data slots can be extracted with an accessor function, as indicated in the first lines of the panel. The following slots conform the object: slot *XY*, storing a data frame with geographic coordinates; slot *P*, storing a phenotypic data frame; slot *G*, storing a genotypic data frame; slot *A*, containing as matrix of counts per allele the information of *G* (only available for codominant markers); slot *E*, storing an environmental data frame; slot *S*, storing a data frame with structure information (hierarchies) assigned to the individuals; slot *C*, for a custom data frame; and slot *OUT*, containing a list for storage of results; (B) Abstract representation of an `ecogen` object with the four basic data frames (*XY*, *P*, *G*, *E* and *S*), where the complete set of layers behave as a "stack". A point with (*x*, *y*) coordinates crossing the stack is indicated with an arrow.

Fig. 3 Operations available for `ecogen` objects. (A) Object subsetting using single brackets; (B) Object subsetting by group (`eco.subset`); (C) Object splitting into groups (`eco.split`); (D) Union of objects by rows (`eco.rbind`); (E) Union of objects by columns (`eco.cbind`); (F) Intersection of two objects (`eco.merge`).

Fig. 4 Examples of graphical outputs for different functions of the package. (A) Correlogram for multiple variables using the Moran's *I* statistic, with confidence intervals obtained by jackknife; (B) Bearing Plot for Moran's *I*. Periodic function of the Moran's *I* and the distance classes against compass direction; (C) Plot for multiples variables with the global Moran's *I*; (D) Plot for a univariate local analysis with the G_i* statistic in a simulated grid. The X and Y coordinates of the points were ranked, using the ranks as the new X and Y axes. This normalization step allows to plot the points in a scale-independent fashion to show all of them in a single image without overlapping. Red and blue filled circles indicate significant positive and negative values of autocorrelation, respectively. The significance was estimated by a permutation test with correction for multiple comparisons; (E) Plot for a multiple-variable extension of the analysis in image C. A matrix of individuals (horizontal axis) x variables (vertical axis) x G_i* values is represented as a heatmap. Positive and negative values with significant autocorrelation are indicated with red and blue colors.

Acknowledgements

This research was supported by fundings from Universidad de Buenos Aires (UBACYT 20020130100043BA), Agencia Nacional de Promoción Científica y Tecnológica (PICTO OTNA 00081 and PICT 2013 0478), and Consejo Nacional de Investigaciones Científicas y Técnicas (PIP

CONICET Nº 11220130100191CO) granted to B.O.S. and J.C.V. The authors wish to express their gratitude to three anonymous reviewers for their valuable comments and suggestions which contributed significantly to the improvement of the paper. The authors are also indebted to Dr. Peter Felker who greatly contributed to improve the English style.

References

Anselin L (1995) Local indicators of spatial association-LISA. *Geographical Analysis*, **27**, 93–115.

Arnaud J (2003) Metapopulation genetic structure and migration pathways in the land snail *Helix aspersa*: influence of landscape heterogeneity. *Landscape Ecology*, **18**, 333–346.

Bivand R, Piras G (2015) Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software*, **63**, 1–36.

Bjornstad O (2016) *ncf: Spatial Nonparametric Covariance Functions. R package version 1.1-7*. https://CRAN.R-project.org/package=ncf.

Blanchet F, Legendre P, Borcard D (2008) Modelling directional spatial processes in ecological data. *Ecological modelling,* **215**, 325–336.

Borcard D, Legendre P, Drapeau P (1992) Partialling out the spatial component of ecological variation. *Ecology*, **73**, 1045–1055.

Borcard, D, Legendre P (1994) Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). *Environmental and Ecological statistics*, **1**, 37–61.

Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, **153**, 51–68.

Borcard D, Legendre P, Avois-Jacquet C, Tuomisto H (2004) Dissecting the spatial structure of ecological data at multiple scales. *Ecology*, **85**, 1826–1832.

Born C, Le Roux P, Spohr C, McGeoch M, Van Vuuren B (2012) Plant dispersal in the sub- Antarctic inferred from anisotropic genetic structure. *Molecular ecology*, **21**, 184–194.

Cliff A, Ord J (1981) Spatial Processes, Models and Applications. Pion, London.

Dasu T, Johnson T (2003) Exploratory Data Mining and Data Cleaning. John Wiley & Sons, Hoboken, New Jersey.

Dray S, Legendre P, Peres-Neto P (2006) Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling*, **196**, 483–493.

Dray S, Dufour, A-B (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, **22**, 1–20.

Dray S, Blanchet G, Borcard D, Guenard G, Jombart T, Larocque G, Legendre P, Wagner H (2016) *adespatial: Multivariate Multiscale Spatial Analysis. R package version 0.0-6.* https://CRAN.R-project.org/package=adespatial.

Dyer R (2014) popgraph: *This is an R package that constructs and manipulates population graphs. R package version 1.4.* https://CRAN.R-project.org/package=popgraph.

Dyer R (2016) gstudio: *Tools Related to the Spatial Analysis of Genetic Marker Data. R package version 1.5.0.*

Epperson B (1993) Recent advances in correlation studies of spatial patterns of genetic variation. *Evolutionary Biology*, **27**, 95–155.

Falsetti A, Sokal R (1993) Genetic structure of human populations in the British Isles. *Annals of Human Biology*, **20**, 215–229.

Fortin M, Dale M (2005) Spatial analysis: a guide for ecologists. Cambridge University Press, New York.

Geary R (1954) The contiguity ratio and statistical mapping. *The Incorporated Statistician*, **5**, 115–146.

Getis A, Ord J (1992) The analysis of spatial association by use of distance statistics. *Geographical Analysis*, **24**, 189–206.

Goslee C, Urban D (2007) The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, **22**, 1–19.

Goudet J, Jombart T (2015) *hierfstat: Estimation and Tests of Hierarchical F-Statistics. R package version 0.04-22.* https://CRAN.R-project.org/package=hierfstat.

Guillot G, Mortier F, Estoup A (2005) Geneland: A program for landscape genetics. *Molecular Ecology Notes*, **5**, 712–715.

Guillot G, Rousset F. (2013) Dismantling the Mantel tests. *Methods in Ecology and Evolution*, **4**, 336–344.

Hardy O, Vekemans X (1999) Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity*, **83**, 145–154.

Hardy O, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Kalisz S, Nason J, Hanzawa F, Tonsor S (2001) Spatial population genetic structure in *Trillium grandiflorum*: the roles of dispersal, mating, history, and selection. *Evolution*, **55**, 1560–1568.

Kamvar Z, Tabima J, Grünwald N. (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, **2**, e281.

Legendre P (1990) Quantitative methods and biogeographic analysis. In: *Evolutionary biogeography of the marine algae of the North Atlantic* (ed. Garbary D, South G), NATO ASI Series, Volume G22, pp. 9–34. Springer, Berlin.

Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673.

Legendre, P, Fortin M (2010) Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular ecology resources*, **10**, 831–844.

Legendre P, Legendre L (2012) Numerical ecology, 3rd English edition. Elsevier Science BV, Amsterdam.

Legendre P, Fortin M, Borcard D (2015) Should the Mantel test be used in spatial analysis?. *Methods in Ecology and Evolution*, **6**, 1239–1247.

Levin S (2000) Multiple scales and the maintenance of biodiversity. *Ecosystems*, **3**, 498–506.

Loiselle B, Sork V, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany*, **82**, 1420–1425.

Manel S, Schwartz M, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**, 189–197.

Moran P (1948) The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B*, **10**, 243–251.

Moran P (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Oden N, Sokal R (1986) Directional autocorrelation: an extension of spatial correlograms to two dimensions. *Systematic Biology*, **35**, 608–617.

Oksanen J, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Simpson G, Solymos P, Stevens M, Wagner H (2016) *vegan: Community Ecology Package. R package version 2.3-5.* Available at: https://CRAN.R-project.org/package=vegan.

Ord J, Getis A (1995) Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, **27**, 286–306.

Paradis E (2010) pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, **26**, 419–420

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

Reich R, Czaplewski R, Bechtold W (1994) Spatial cross-correlation of undisturbed, natural shortleaf pine stands in northern Georgia. *Environmental and Ecological Statistics*, **1**, 201– 217.

Rey S, Anselin L (2010) PySAL: A Python Library of Spatial Analytical Methods. In: *Handbook of Applied Spatial Analysis* (eds. Fischer M, Getis A), pp. 175–193. Springer-Verlag, Berlin.

Rosenberg M (2000) The bearing correlogram: a new method of analyzing directional spatial autocorrelation. *Geographical Analysis*, **32**, 267–278.

Rousset F (2008) Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Sievert C, Parmer C, Hocking T, Chamberlain S, Ram K, Corvellec M, Despouy P (2016) *plotly: Create Interactive Web Graphics via 'plotly.js'. R package version 4.5.6.* https://CRAN.R-project.org/package=plotly.

Smouse P, Long J, Sokal R (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology*, **35**, 627–632.

Smouse P, Peakall R (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, **82**, 561–573.

Sokal R, Oden N (1978a) Spatial autocorrelation in biology: 1. Methodology. *Biological Journal of the Linnean Society*, **10**, 199–228.

Sokal R, Oden N (1978b) Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society*, **10**, 229–249.

Sokal R (1986) Spatial data analysis and historical processes. In: *Data analysis and informatics* (eds. Diday E, Escoufier Y, Lebart L, Pages J, Scherktman Y, Tomassone R), Volume IV, pp. 29–43. Elsevier Publishers, North-Holland, Amsterdam.

Sokal R, Oden N, Thomson B (1998) Local spatial autocorrelation in a biological model. *Geographical Analysis*, **30**, 331–354.

Sokal R, Thomson B (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. *American Journal of Physical Anthropology*, **129**, 121–131.

Storfer A, Murphy M, Evans J, Goldberg C, Robinson S, Spear S, Dezzani R, Delmelle E, Vierling L, Waits L (2007) Putting the landscape in landscape genetics. *Heredity*, **98**, 128–142.

Vekemans X, Hardy O. (2004) New insights from fine- scale spatial genetic structure analyses in plant populations. *Molecular Ecology*, **13**, 921–935.

Wagner H (2004) Direct multi-scale ordination with canonical correspondence analysis. *Ecology*, **85**, 342–351.

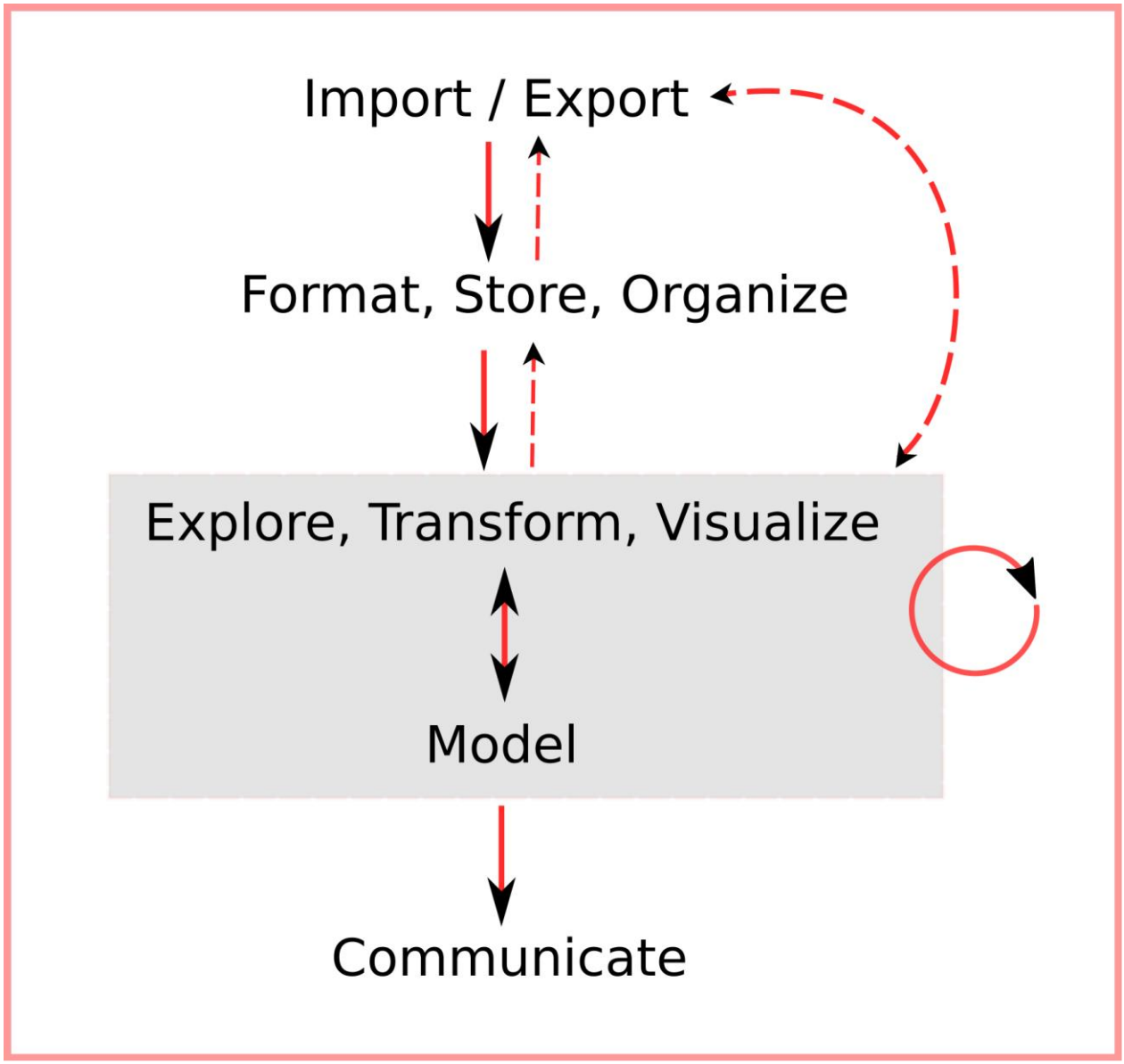
Wickham H (2009) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Wickham H (2014) Tidy Data. *Journal of Statistical Software*, **59**.

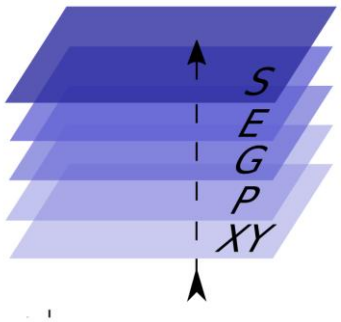
Wickham H, Grolemund G (2016) R for Data Science. O'Reilly Media, Sebastopol, California.

Author Contributions

L.G.R designed and wrote the package code. J.C.V. supervised the design and tested the functions. L.I.F. and B.O.S. revised the final design of output display. All authors contributed to writing the manuscript.



(A)



(B)

```
                || ECOGEN CLASS OBJECT ||
-------------------------------------------------------------------------
Access to slots: <ecoslot.> + <name of the slot> + <(name of the object)>
See help("EcoGenetics accessors")
-------------------------------------------------------------------------

| slot XY:    | --> 225 x 2        coordinates
| slot P:     | --> 225 x 8        phenotypic variables
| slot G:     | --> 225 x 10       loci         >> ploidy: 2 || codominant
| slot A:     | --> 225 x 40       alleles
| slot E:     | --> 225 x 6        environmental variables
| slot S:     | --> 225 x 1        structure    >> 1 structure found
| slot C:     | --> 0 x 0          variables
| slot OUT:   | --> 0              results
-------------------------------------------------------------------------
```