Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Modelling Pre- and Protohistorical Iconographic Compositions. The R package decorr

Thomas Huet
UMR 5140

Abstract

By definition, Prehistorical societies are characterised by the absence of a writing system. Writing is one of the most rational semiographical system with a clear distinction between signified and signifier – specially in alphabetic and binary writings – and the development of the signified on a horizontal, vertical or boustrophedon axis. Prehistorical times cover more than 99% of the human living. Even if it is being discussed, first symbolic manifestations start around 200,000 BC (d'Errico and Nowell 2000). The duration from first symbolic expressions to start of writing represents 97% of the human living. In illiterate societies, testimonies of symbolic systems mostly come from iconography (ceramic decorations, rock-art, statuary, etc.) and signs are displayed mostlty a discontinuous figures which can have different relationships one with another. An iconographical composition can be "read" as a spatial distribution of features having intrinsic values possibily having meaningful relationships one with another depending on their pairwise spatial proximities.

To understand meaningful associations of signs, geometric tools, graph analysis and statistical analysis offer great tools to recognize iconographical patterns and to infer collective conventions. We present the **decorr** R package which ground concepts, methods and tools to analyse ancient iconographical systems.

Keywords: Iconography, Prehistory, Graph Theory, Graph Drawing, Spatial Analysis, R.

1. Introduction: Count data regression in R

The introduction is in principle “as usual”. However, it should usually embed both the implemented *methods* and the *software* into the respective relevant literature. For the latter both competing and complementary software should be discussed (within the same software environment and beyond), bringing out relative (dis)advantages. All software mentioned should be properly `\cite{}`d. (See also Appendix B for more details on BibTeX.)

> For writing about software JSS requires authors to use the markup `\proglang{}` (programming languages and large programmable systems), `\pkg{}` (software packages), `\code{}` (functions, commands, arguments, etc.). If there is such markup in (sub)section titles (as above), a plain text version has to be provided in the LaTeX command as well. Below we also illustrate how abbrevations should be introduced and citation commands can be employed. See the LaTeX code for more details.

For decades, study of ancient iconography was linked to history of religion because closely linked to symbolism, believes and religions. Since the *New Archaeology* developpement during the 60's (Clarke 2014), symbolic expressions start to be studied with the same formal methods (statistics, seriations, distribution maps, etc.) as any another aspect of social organisation: settlement patterns, tools *chaine opÃl'ratoire*, susbsitence strategies, etc. (Renfrew and Bahn 1991), (Leroi-Gourhan 1992). But unlike many aspects of the material culture – a flint blade for cutting, a pottery for containing, a house for living –, the function of an iconographic composition cannot be drawn directly from itself. Whether study of ancient iconography had undergone significative improvements at the site scale – with GIS, database, paleoclimatic restitutions, etc. – and at the sign scale with the development of archaeological sciences – radiocarbon dating, use-wear analysis, elemental analysis, etc. –, these improvement do not necessary help to understand the semantic content of the iconography.

Semantics or semiotics can be defined as a system of conventional and repetead signs organised also in conventional manners . As any formal system, iconography can be seen as spatial features related one with the other depending on rules of proximities.

Until our days, formal methods to study ancient iconography Semantics, has been mostlty been grounded (explicitly or not) on the prime principle of Saussurian linguistic: the 'linearity of the signifier' (De Saussure 1989). This principle states that the meaning of a lingusitic or writing pattern is linear and directed. Let us take the example of the word `"art"` which contains three vertices (`a`, `r`, `t`) and two edges (one between `a` and `r`, the other between `r` and `t`).

Figure 1: `a`, `r` and `t` graphical units is `art`.

Applying this principle to any other graphical content than a writing, allows to considered as the organisation of graphical as a relationship of figures grouping graphical units, patterns grouping figures, motives grouping patterns, etc., until the entire decorated support is de-

scribed and can be compared to another decoration (XXX). But during this *decomposition* process, the groups and relationships are often defined empirically, their level of significance are often implicit and the iconographical and spatial proximities between graphical units and categories of graphical units are not quantified. Furthermore, due to the inherent variability of iconography, most of the studies developp proper descriptive vocabularies, singular relationships of categories, idosyncratic methods in a site-dependend or period-dependend scales. This limits drastically the possibility to conduct cross-cultural comparisons and to draw a synthesis of humankind's symbolism at a large scale and over the long-term.
In this article we present the `R` package **decorr**. Its purpose is to formalise a method based on geometric graphs to analyse any graphical content. The idea is that a graphical system can be represented by vertices connected (or not) to each other with edges. This package has been grounded on the seminal work of C. Alexander (Alexander 2008) and its first IT implementation by T. Huet (Huet 2018).

Just like in `R`, the features `a`, `r` and `t` concatenated in this order with a `paste0()`, mean `"art"`, and not `"rat"`. In illetrate societies, spatial relations between graphical features are not necessarly linear and directed but multi-directional and undirected: the direction of the interactions of pairwise graphical units can be in any order.

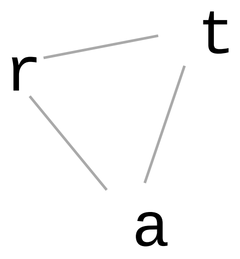


Figure 2: Potential spatial relations between `a`, `r` and `t` graphical units.

2. Concepts

Graph theory offers a conceptual framework and indices (global at the entire graph scale, local at the vertex scale) to deal with notions of networks, relationships and neighbourhoods. The spatial levels of the graphical units can be retrieve by a planar graph (Graph Theory) and a spatial (GIS) analysis.

Nodes and edges – repectively for graphical units and their connexion – are created on a GIS interface. In the GIS, the decoration figure is open in the first place in a new project with no projection. The decoration image will be considerated as the basemap of the project and will cover the region of interest of the analysis. The decoration image can be binarized where graphical units are considerated active and the undecorated parts of the support, or background, are considerated inactive. After what, the decoration image is tiled. A simplier solution will be to create directly centroids over the graphical units. The x and y coordinates of the nodes are relative to the decoration and measured in pixels. Exist a link between a couple of graphical units when these graphical units share a border. A planar graph is constructed from grahical units (nodes) and their proximity links (edges). This model is a Voronoi diagram of the support where the Voronoi seeds are the graphical units. Its geographical equivalent is a Thiessen polygon.

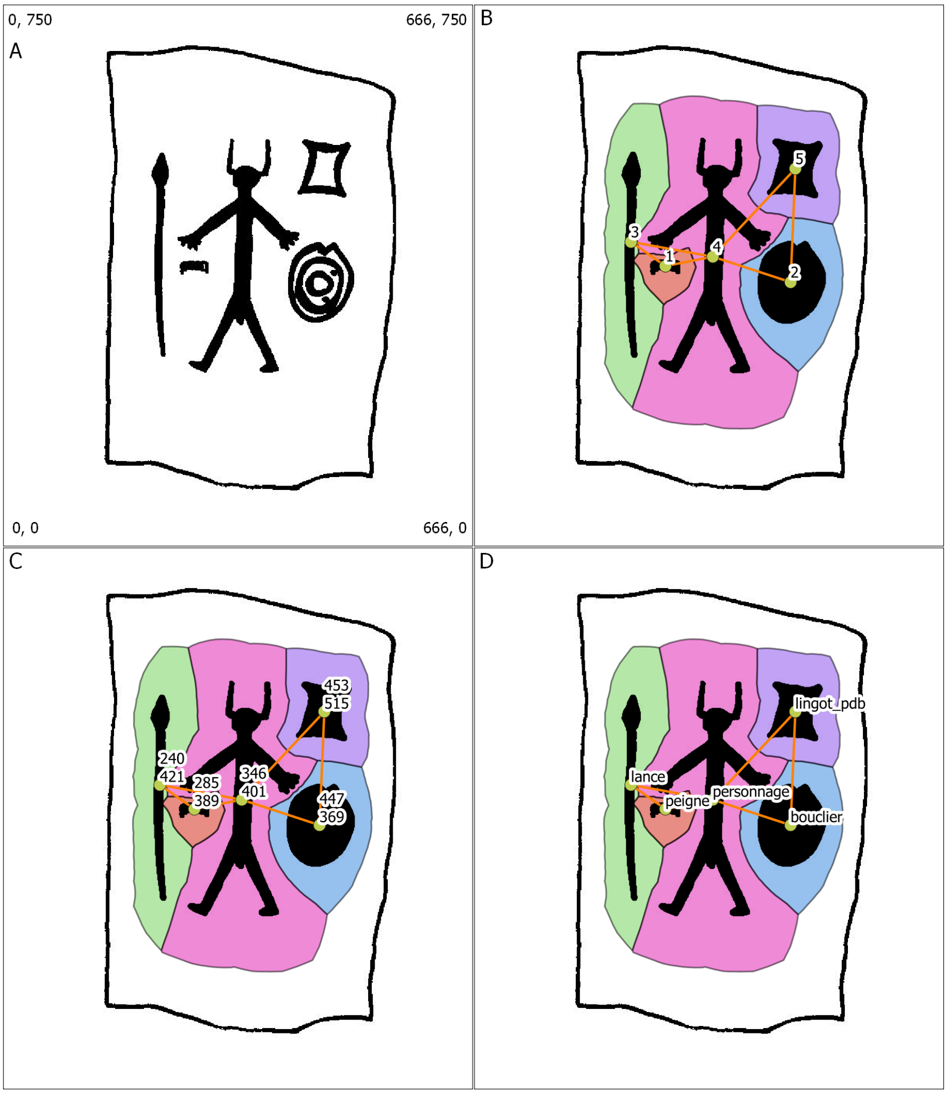


Figure 3: GIS interface. A) Original decoration of the Late Bronze Age Cerro Muriano 1 stelae (drawing: Díaz-Guardamino Uribe (2010)) with its extent (`xmin, xmax, ymin, ymax`); B) After the polygonisation of the graphical units, including the border of the stelae, the Voronoi cells, the centroid of graphical units and the links between graphical units having adjacent cells (ie, sharing a border) are calculated; C) For each graphical units, x and y are calculated; D) At least one variable, like the `type` of the graphical units is defined in order to compute composition analysis.

This model has a minimal of *a priori* definitions. Those only concern the graphical units (type, technology, color, orientation, size, etc.). Between two graphical units the links are conventionaly represented with a plain line. But sometimes a graphical unit can be divided into a main unit (eg, a man) and attribute units (eg, a helmet, a sword). So, the links between the main unit and the attribute units are directed and displayed with a dashed line.

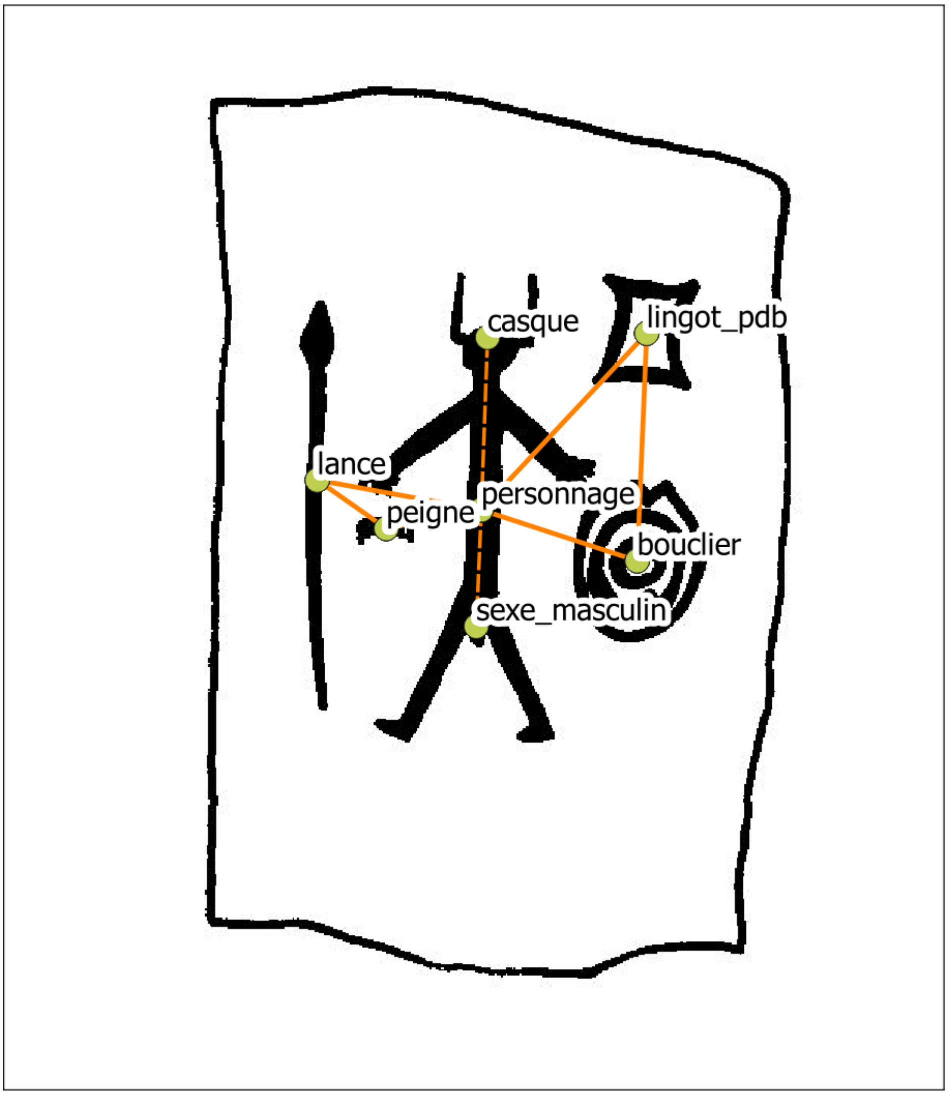


Figure 4: The graphical unit 'casque' (helmet) is an attribute of the graphical unit 'personnage'. (character).

The `r` graphical unit receives two edges, so its centrality degree is 2, `r` is also central in the graph so its betweeness degree is 1.

3. The R package decorr

The **decorr** package can be downloaded from GitHub

```
R>   devtools::install_github("zoometh/iconr")
```

3.1. External package

The **decorr** package imports the following packages:

- **magick** for image manipulation (Ooms 2018)

- **igraph** for graph and network analysis (Csardi and Nepusz 2006)

- **rgdal** to read shapefiles of nodes and/or edges (Bivand, Keitt, and Rowlingson 2019)

- **grDevices** for colors and font plotting, **graphics** for graphics, **utils** and **methods** for formally defined methods and *varia* methods (all combinations, etc.) (R Core Team 2019)

3.2. External data

The training dataset is a selection of four decorated Late Bronze age stelae of the SW Iberian peninsula.

idf	site	decor	img
1	Cerro Muriano	Cerro Muriano 1	Cerro_Muriano.Cerro_Muriano_1.jpg
2	Torrejon Rubio	Torrejon Rubio 1	Torrejon_Rubio.Torrejon_Rubio_1.jpg
3	Brozas	Brozas	Brozas.Brozas.jpg
4	Zarza de Montanchez	Zarza De Montanchez	Zarza_de_Montanchez.Zarza_De_Montanchez.jpg

Table 1: studied corpus

The default dataframe is `imgs`. It list the decorations. The field `imgs$idf` is the short name of the decoration. The concatened of `imgs$site` and `imgs$decor` is the key of the decoration. Decoration images are All the drawings belongs to (Díaz-Guardamino Uribe 2010).

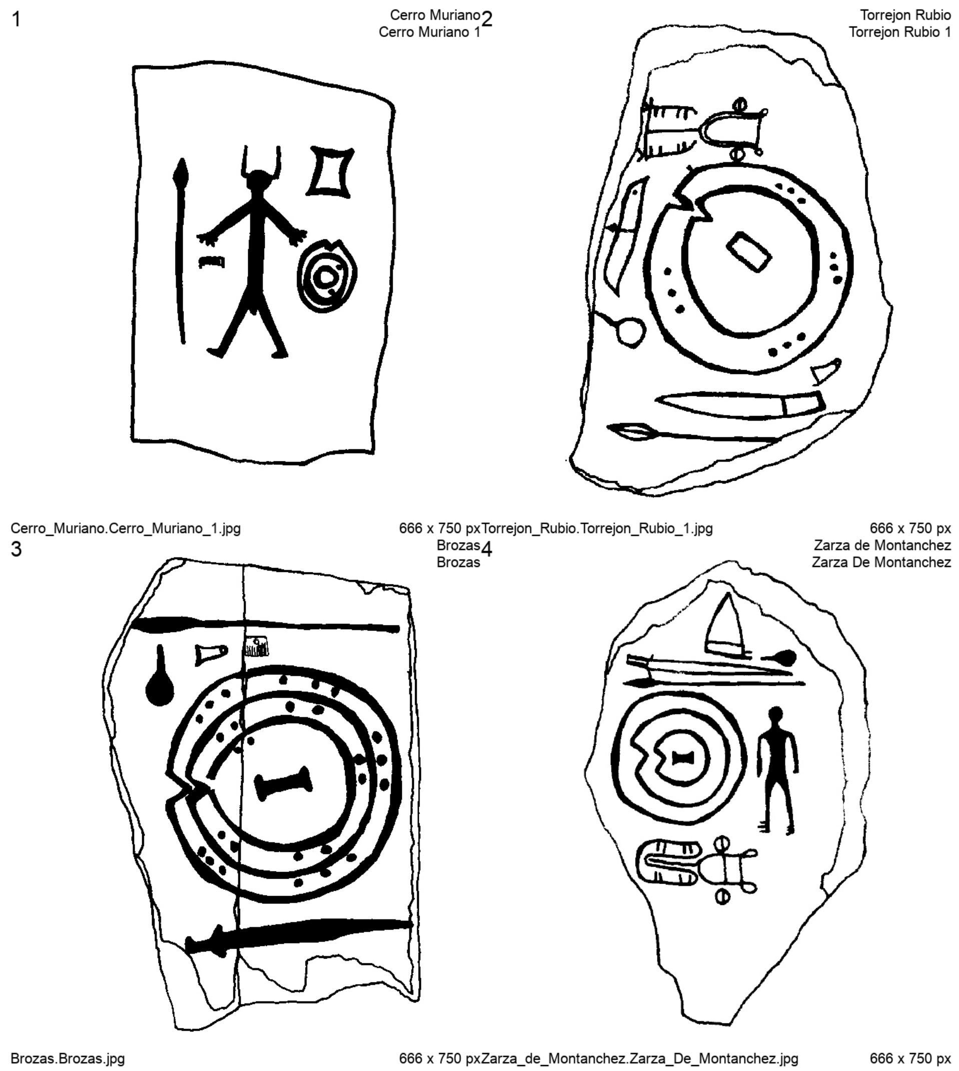
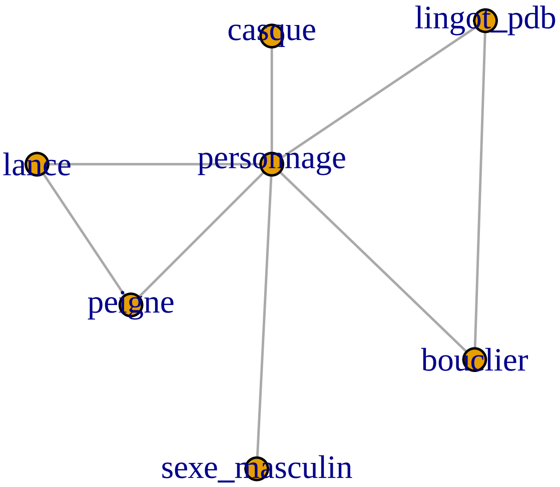


Figure 5: Dataset of four stelae showing the information on the decoration (`site`, `decor` and `img`) and the dimensions in pixel of the decoration image

3.3. Functions

The `list_dec()` function allows to store undirected graphs for each decorations stored into `nodes`, `edges` and `images` dataframes and store the graphs in a list. The join between these dataframes is done on the two fields `site` and `decor`.

```
> # read imgs, nodes and edges dataframes
> imgs <- read.table(system.file("extdata", "imgs.csv", package = "decorr"),
+                     sep="\t",stringsAsFactors = FALSE)
> nodes <- read.table(system.file("extdata", "nodes.csv", package = "decorr"),
+                      sep="\t",stringsAsFactors = FALSE)
> edges <- read.table(system.file("extdata", "edges.csv", package = "decorr"),
+                       sep="\t",stringsAsFactors = FALSE)
> # create the list of graphs
> lgrph <- decorr::list_dec(imgs,nodes,edges,var="type")
> # plot the first graph of the list
> g <- lgrph[[1]]
> plot(g,
+      vertex.size = 10,
+      vertex.label.cex =.8)
```



The `labels_shadow()` is a re-use of the `shadowtext()` function from the **TeachingDemos** package (Snow 2020).
The others **decorr** package functions can be divided into:

1. functions for a single decoration

2. functions for comparisons between different decorations

Single decoration

Functions allowing to model a single decoration with a geometric graph are:

- `read_nds()` and `read_eds()` allow to read respectively a file of nodes and a file of edges (`.tsv` or `.shp` files)
- `plot_dec_grph ()` allows to plot a geometric graph over a decoration image

Decoration comparisons

Functions allowing to compare different decorations with geometric graphs are:

- `list_nds_compar()` and `list_eds_compar()` allow to compare respectively the common nodes and the common edges between two decorations
- `plot_nds_compar()` and `plot_eds_compar()` allow to plot and save two figures side-by-side for a decorations pairwise with, respectively, common nodes and common edges identified
- `same_nds()` and `same_eds()` allow to repectively count matching nodes and matching edges between decoration pairwises

3.4. Function `listdec()`

This function store graphs in a list.

3.5. Function `xxx`

xxx

The basic Poisson regression model for count data is a special case of the GLM framework ?. It describes the dependence of a count response variable y_i $(i = 1, \dots, n)$ by assuming a Poisson distribution $y_i \sim \text{Pois}(\mu_i)$. The dependence of the conditional mean $\mathsf{E}[y_i \,|\, x_i] = \mu_i$ on the regressors x_i is then specified via a log link and a linear predictor

$$\log(\mu_i) \quad = \quad x_i^\top \beta, \tag{1}$$

where the regression coefficients β are estimated by maximum likelihood (ML) using the iterative weighted least squares (IWLS) algorithm.

> Note that around the `{equation}` above there should be no spaces (avoided in the LaTeX code by `%` lines) so that "normal" spacing is used and not a new paragraph started.

R provides a very flexible implementation of the general GLM framework in the function `glm()` (?) in the **stats** package. Its most important arguments are

```
glm(formula, data, subset, na.action, weights, offset,
  family = gaussian, start = NULL, control = glm.control(...),
  model = TRUE, y = TRUE, x = FALSE, ...)
```

Type	Distribution	Method	Description
GLM	Poisson	ML	Poisson regression: classical GLM, estimated by maximum likelihood (ML)
		Quasi	"Quasi-Poisson regression": same mean function, estimated by quasi-ML (QML) or equivalently generalized estimating equations (GEE), inference adjustment via estimated dispersion parameter
		Adjusted	"Adjusted Poisson regression": same mean function, estimated by QML/GEE, inference adjustment via sandwich covariances
	NB	ML	NB regression: extended GLM, estimated by ML including additional shape parameter
Zero-augmented	Poisson	ML	Zero-inflated Poisson (ZIP), hurdle Poisson
	NB	ML	Zero-inflated NB (ZINB), hurdle NB

Table 2: Overview of various count regression models. The table is usually placed at the top of the page (`[t!]`), centered (`centering`), has a caption below the table, column headers and captions are in sentence style, and if possible vertical lines should be avoided.

where `formula` plus `data` is the now standard way of specifying regression relationships in R/S introduced in ?. The remaining arguments in the first line (`subset`, `na.action`, `weights`, and `offset`) are also standard for setting up formula-based regression models in R/S. The arguments in the second line control aspects specific to GLMs while the arguments in the last line specify which components are returned in the fitted model object (of class '`glm`' which inherits from '`lm`'). For further arguments to `glm()` (including alternative specifications of starting values) see `?glm`. For estimating a Poisson model `family = poisson` has to be specified.

> As the synopsis above is a code listing that is not meant to be executed, one can use either the dedicated `{Code}` environment or a simple `{verbatim}` environment for this. Again, spaces before and after should be avoided.
>
> Finally, there might be a reference to a `{table}` such as Table 2. Usually, these are placed at the top of the page (`[t!]`), centered (`\centering`), with a caption below the table, column headers and captions in sentence style, and if possible avoiding vertical lines.

4. Illustrations

For a simple illustration of basic Poisson and NB count regression the `quine` data from the **MASS** package is used. This provides the number of `Days` that children were absent from school in Australia in a particular year, along with several covariates that can be employed as regressors. The data can be loaded by

```
R> data("quine", package = "MASS")
```

and a basic frequency distribution of the response variable is displayed in Figure 6.

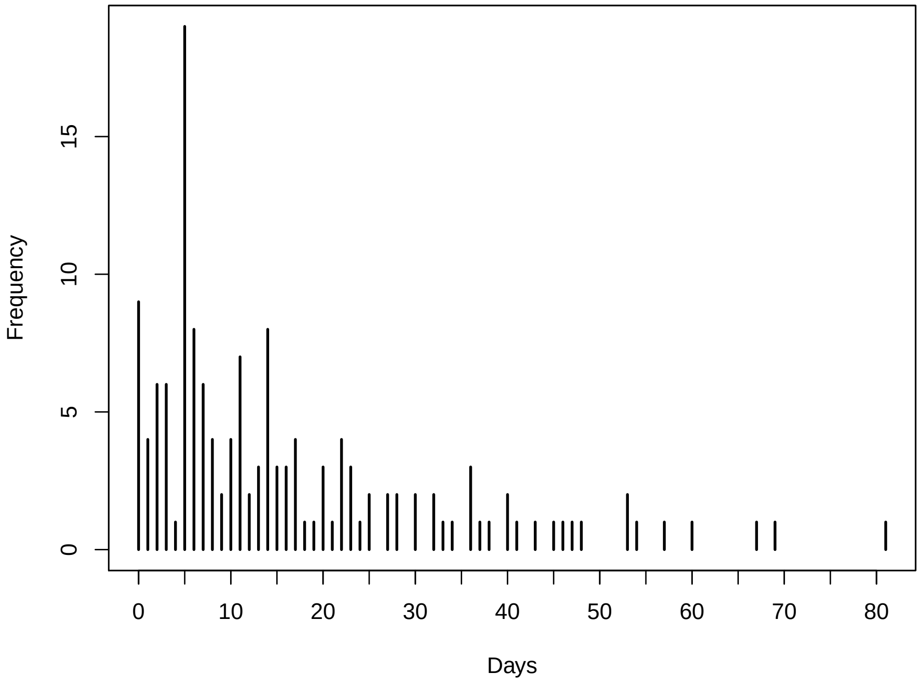


Figure 6: Frequency distribution for number of days absent from school.

> For code input and output, the style files provide dedicated environments. Either the "agnostic" {CodeInput} and {CodeOutput} can be used or, equivalently, the environments {Sinput} and {Soutput} as produced by Sweave() or **knitr** when using the render_sweave() hook. Please make sure that all code is properly spaced, e.g., using y = a + b * x and *not* y=a+b*x. Moreover, code input should use "the usual" command prompt in the respective software system. For R code, the prompt "R> " should be used with "+ " as the continuation prompt. Generally, comments within the code chunks should be avoided – and made in the regular LaTeX text instead. Finally, empty lines before and after code input/output should be avoided (see above).

As a first model for the quine data, we fit the basic Poisson regression model. (Note that JSS prefers when the second line of code is indented by two spaces.)

```
R> m_pois <- glm(Days ~ (Eth + Sex + Age + Lrn)^2, data = quine,
+    family = poisson)
```

To account for potential overdispersion we also consider a negative binomial GLM.

```
R> library("MASS")
R> m_nbin <- glm.nb(Days ~ (Eth + Sex + Age + Lrn)^2, data = quine)
```

In a comparison with the BIC the latter model is clearly preferred.

```
R> BIC(m_pois, m_nbin)
```

```
       df      BIC
m_pois 18 2046.851
m_nbin 19 1157.235
```

Hence, the full summary of that model is shown below.

```
R> summary(m_nbin)

Call:
glm.nb(formula = Days ~ (Eth + Sex + Age + Lrn)^2, data = quine,
    init.theta = 1.60364105, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0857  -0.8306  -0.2620   0.4282   2.0898

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.00155    0.33709   8.904  < 2e-16 ***
EthN        -0.24591    0.39135  -0.628  0.52977
SexM        -0.77181    0.38021  -2.030  0.04236 *
AgeF1       -0.02546    0.41615  -0.061  0.95121
AgeF2       -0.54884    0.54393  -1.009  0.31296
AgeF3       -0.25735    0.40558  -0.635  0.52574
LrnSL        0.38919    0.48421   0.804  0.42153
EthN:SexM    0.36240    0.29430   1.231  0.21818
EthN:AgeF1  -0.70000    0.43646  -1.604  0.10876
EthN:AgeF2  -1.23283    0.42962  -2.870  0.00411 **
EthN:AgeF3   0.04721    0.44883   0.105  0.91622
EthN:LrnSL   0.06847    0.34040   0.201  0.84059
SexM:AgeF1   0.02257    0.47360   0.048  0.96198
SexM:AgeF2   1.55330    0.51325   3.026  0.00247 **
SexM:AgeF3   1.25227    0.45539   2.750  0.00596 **
SexM:LrnSL   0.07187    0.40805   0.176  0.86019
AgeF1:LrnSL -0.43101    0.47948  -0.899  0.36870
AgeF2:LrnSL  0.52074    0.48567   1.072  0.28363
AgeF3:LrnSL       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.6036) family taken to be 1)

    Null deviance: 235.23  on 145  degrees of freedom
Residual deviance: 167.53  on 128  degrees of freedom
AIC: 1100.5

Number of Fisher Scoring iterations: 1
```

```
              Theta:  1.604
          Std. Err.:  0.214

 2 x log-likelihood:  -1062.546
```

5. Summary and discussion

> As usual . . .

Computational details

> If necessary or useful, information about certain computational details such as version numbers, operating systems, or compilers could be included in an unnumbered section. Also, auxiliary packages (say, for visualizations, maps, tables, . . .) that are not cited in the main text can be credited here.

The results in this paper were obtained using R 3.4.1 with the **MASS** 7.3.47 package. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

Acknowledgments

> All acknowledgments (note the AE spelling) should be collected in this unnumbered section before the references. It may contain the usual information about funding and feedback from colleagues/reviewers/etc. Furthermore, information such as relative contributions of the authors may be added here (if any).

References

Alexander C (2008). "The Bedolina map – an exploratory network analysis." In A Posluschny, K Lambers, I Herzog (eds.), *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Berlin, 2.-6. April 2007*, pp. 366–371. Koll. Vor- u. FrÃijhgesch. doi:https://doi.org/10.11588/propylaeumdok.00000512.

Bivand R, Keitt T, Rowlingson B (2019). *rgdal: Bindings for the 'Geospatial' Data Abstraction Library.* R package version 1.4-7, URL https://CRAN.R-project.org/package=rgdal.

Clarke DL (2014). *Analytical archaeology.* Routledge.

Csardi G, Nepusz T (2006). "The igraph software package for complex network research." *InterJournal*, **Complex Systems**, 1695. URL http://igraph.org.

De Saussure F (1989). *Cours de linguistique générale*, volume 1. Otto Harrassowitz Verlag.

d'Errico F, Nowell A (2000). "A new look at the Berekhat Ram figurine: implications for the origins of symbolism." *Cambridge Archaeological Journal*, **10**(1), 123–167.

Díaz-Guardamino Uribe M (2010). *Las estelas decoradas en la Prehistoria de la Península Ibérica*. Ph.D. thesis, Universidad Complutense de Madrid, Servicio de Publicaciones.

Huet T (2018). "Geometric graphs to study ceramic decoration." In M Matsumoto, E Uleberg (eds.), *Exploring Oceans of Data, proceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology, CAA 2016*, pp. 311–324. Archaeopress.

Leroi-Gourhan A (1992). *L'art pariétal: langage de la préhistoire*. Editions Jérôme Millon.

Ooms J (2018). "Magick: advanced graphics and image-processing in R." *CRAN. R package version*, **1**.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Renfrew C, Bahn PG (1991). *Archaeology: theories, methods and practice*, volume 2. Thames and Hudson London.

Snow G (2020). *TeachingDemos: Demonstrations for Teaching and Learning*. R package version 2.12, URL https://CRAN.R-project.org/package=TeachingDemos.

A. More technical details

Appendices can be included after the bibliography (with a page break). Each section within the appendix should have a proper section title (rather than just *Appendix*).

For more technical style details, please check out JSS's style FAQ at `https://www.jstatsoft.org/pages/view/style#frequently-asked-questions` which includes the following topics:

- Title vs. sentence case.
- Graphics formatting.
- Naming conventions.
- Turning JSS manuscripts into R package vignettes.
- Trouble shooting.
- Many other potentially helpful details…

B. Using BibTEX

References need to be provided in a BibTEX file (`.bib`). All references should be made with `\cite`, `\citet`, `\citep`, `\citealp` etc. (and never hard-coded). This commands yield different formats of author-year citations and allow to include additional details (e.g., pages, chapters, …) in brackets. In case you are not familiar with these commands see the JSS style FAQ for details.

Cleaning up BibTEX files is a somewhat tedious task – especially when acquiring the entries automatically from mixed online sources. However, it is important that informations are complete and presented in a consistent style to avoid confusions. JSS requires the following format.

- JSS-specific markup (`\proglang`, `\pkg`, `\code`) should be used in the references.
- Titles should be in title case.
- Journal titles should not be abbreviated and in title case.
- DOIs should be included where available.
- Software should be properly cited as well. For R packages `citation("pkgname")` typically provides a good starting point.

Affiliation:

Thomas Huet
UMR 5140
Archeologie des Societes Mediterraneennes
Universite Paul Valery
route de Mende
Montpellier 34199, France
E-mail: thomashuet7@gmail.com

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd