Received: 17 January 2019 | Revised: 18 June 2019 | Accepted: 20 June 2019

DOI: 10.1002/ece3.5448

Ecology and Evolution Open Access WILEY

ORIGINAL RESEARCH

Analysis of interval-grouped data in weed science: The binnednp Rcpp package

Daniel Barreiro-Ures[1] | Mario Francisco-Fernández[1] | Ricardo Cao[1] | Basilio B. Fraguela[2] | Ramón Doallo[2] | José Luis González-Andújar[3] | Miguel Reyes[4]

[1]Research Group MODES, CITIC, Departamento de Matemáticas, Facultade de Informática, Universidade da Coruña, A Coruña, Spain

[2]Research Group GAC, CITIC, Departamento de Ingeniería de Computadores, Facultade de Informática, Universidade da Coruña, A Coruña, Spain

[3]Instituto de Agricultura Sostenible (CSIC), Córdoba, Spain

[4]Departamento de Actuaría, Física y Matemáticas, Universidad de las Américas-Puebla, Puebla, México

Correspondence
José Luis González-Andújar, Instituto de Agricultura Sostenible (CSIC), Apartado 4084, Córdoba 14080, Spain.
Email: andujar@ias.csic.es

Funding information
Ministerio de Economía y Competitividad, Grant/Award Number: AGL2015-64130-R, MTM2014-52876-R, MTM2017-82724-R and AGL2012-33736; Consellería de Cultura, Educación e Ordenación Universitaria, Xunta de Galicia, Grant/Award Number: ED431C-2016-015 and ED431G/01; European Regional Development Fund

Abstract

1. Weed scientists are usually interested in the study of the distribution and density functions of the random variable that relates weed emergence with environmental indices like the hydrothermal time (HTT). However, in many situations, experimental data are presented in a grouped way and, therefore, the standard nonparametric kernel estimators cannot be computed.
2. Kernel estimators for the density and distribution functions for interval-grouped data, as well as bootstrap confidence bands for these functions, have been proposed and implemented in the binnednp package. Analysis with different treatments can also be performed using a bootstrap approach and a Cramér-von Mises type distance. Several bandwidth selection procedures were also implemented. This package also allows to estimate different emergence indices that measure the shape of the data distribution. The values of these indices are useful for the selection of the soil depth at which HTT should be measured which, in turn, would maximize the predictive power of the proposed methods.
3. This paper presents the functions of the package and provides an example using an emergence data set of *Avena sterilis* (wild oat).
4. The binnednp package provides investigators with a unique set of tools allowing the weed science research community to analyze interval-grouped data.

KEYWORDS
bandwidth selection, hydrothermal time, nonparametric kernel estimation, weed emergence model

1 | INTRODUCTION

The knowledge of the factors affecting the emergence patterns of weeds is not only interesting from a plant ecology perspective, but also in applied research, where the emergence of weeds is an important phase of the population dynamics (González-Andújar, 2008). This critical phase has important implications, either because of its effects on the determination of competition with the crop or because of the type and timing of the control tactics that must be used (Forcella, Benech-Arnold, Sánchez, & Ghersa, 2000). Temperature and water potential have been identified as essential factors that control weed emergence (Forcella et al., 2000). Some indices (Hunter, Glasbey, &

Naylor, 1984; Naylor, 1981) and modelling techniques (González-Andújar, Chantre, Morvillo, Blanco, & Forcella, 2016) are often used to predict weed emergence. In this context, thermal time (TT) models and hydrothermal time (HTT) models are useful tools to describe weed emergence (Bradford, 2002; Grundy, 2003; Zambrano-Navea, Bastida, & González-Andújar, 2013). Parametric regression models for emergence are usually employed in this framework. They may offer the simplicity and flexibility required for practical decision support (Grundy, 2003). However, due to the limitations of this approach, different modeling approaches have been proposed, including techniques that account for censoring (Onofri, Gresta, & Tei, 2010; Onofri, Mesgaran, Tei, & Cousens, 2011; Onofri, Piepho, & Kozak, 2019), genetic algorithms (Blanco et al., 2014; Haj Seyed-Hadi & Gonzalez-Andujar, 2009), and artificial neural networks (Chantre et al., 2012). Alternatively, the problem of studying the relation between HTT and weed emergence has been dealt with through nonparametric estimation of the distribution and density functions of cumulative HTT (CHTT) at emergence (Cao, Francisco-Fernández, Anand, Bastida, & González-Andújar, 2013; Reyes, Francisco-Fernández, & Cao, 2016). These nonparametric methods have been recently proven to outperform the usual regression approaches in terms of prediction error (González-Andújar, Francisco-Fernández, et al., 2016).

In addition, when gathering experimental data, a different problem arises due to the fact that seedlings are generally buried at different depths and, therefore, the best depth at which HTT should be measured has to be selected. For this task, emergence indices have been defined and nonparametric estimators for them have been constructed (Cao, Francisco-Fernández, Anand, Bastida, & González-Andújar, 2011).

The techniques required for both, the nonparametric estimation of the density and distribution functions and the emergence indices, have been implemented in the `binnednp` Rcpp package (Barreiro et al., 2019).

2 | METHODS

2.1 | Density and distribution estimation

Let us suppose that modeling the emergence of a certain weed seedling, based on CHTT at emergence, is being investigated. Denote by n the number of seedlings that have emerged at the end of the monitoring process, and by X the random variable measuring the CHTT at emergence (with density function f and distribution function F). Since the inspections to count the number of emerged seedlings are performed at a limited number of instants, say k, the values $X_1, X_2, \ldots, X_n$, measuring the CHTT at emergence of every single seedling, cannot be observed. However, what is observed is the total number of seedlings that have emerged in the intervals between consecutive inspection times, $n_1, n_2, \ldots, n_k$, or the corresponding sample proportions, $w_1, w_2, \ldots, w_k$, with $w_i = n_i/n$. In this sense, this type of data is called interval-grouped data.

In this interval-group framework, if the interest is to estimate the density function f, the standard kernel density estimator (Parzen, 1962; Rosenblatt, 1956),

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\tfrac{x-X_i}{h}),$$

cannot be computed. An appropriate version of this estimator for interval-grouped data has been proposed in Cao et al. (2011),

$$\hat{f}_h^g(x) = \frac{1}{h}\sum_{i=1}^{k} w_i K(\tfrac{x-t_i}{h}), \tag{1}$$

where t_i, $i = 1, \ldots, k$, are the central values between every pair of consecutive observed CHTT. In Equation (1), the function $K(\cdot)$ is the kernel function, and h is the bandwidth or smoothing parameter, controlling the amount of smoothing.

Similarly, to estimate the distribution function F, a kernel distribution estimator adapted for interval-grouped data, derived from (1), was proposed in Cao et al. (2013),

$$\hat{F}_h^g(x) = \sum_{i=1}^{k} w_i \mathbb{K}\left(\frac{x-t_i}{h}\right), \tag{2}$$

where $\mathbb{K}(u) = \int_{-\infty}^{u} K(t)\mathrm{d}t$.

It has been proven that the selection of the kernel function, $K(\cdot)$, is of secondary importance in terms of efficiency. However, the selection of the bandwidth, h, is crucial in the behavior of estimators (1) and (2).

As pointed out in the Introduction, nonparametric estimators (1) and (2) are novel approaches to model weed emergence, presenting some advantages over the parametric regression techniques traditionally used in this framework (Cao et al., 2013; González-Andújar, Francisco-Fernández, et al., 2016). These estimators, jointly with two types of bandwidth selectors, plug-in and bootstrap, have been implemented in the `binnednp` package. Moreover, a new and successful method to select the pilot bandwidth for the bootstrap bandwidths has been proposed and also implemented. Plug-in bandwidth selectors estimate the unknown terms in the expression of the asymptotically optimal bandwidth, whereas bootstrap bandwidths try to directly estimate the optimal bandwidth by mimicking the sampling process through resampling.

The `binnednp` package also allows to compute bootstrap confidence bands for the density and distribution functions that can be used to assess the uncertainty of the corresponding estimates. Additionally, the `binnednp` package includes a function to evaluate the effect of a specific factor on weed emergence, for example, when considering different treatments. This procedure is based on a bootstrap approach properly designed to address the multiple testing problem (Westfall & Young, 1993). The idea of this approach is the following. (a) Split the data set into subsets according to the different levels of the factor under study. (b) Compute the nonparametric estimator of the emergence curve considering the pooled sample and, for each level, the nonparametric estimators of the emergence curves using the corresponding data subsets. (c) A reasonable statistic, D, to test the null hypothesis that the factor effect is not significant is defined based on a Cramér-von Mises distance between the nonparametric estimator with the pooled sample and the nonparametric estimators in the different groups. This distance is inspired in the generalization of the Cramér-von Mises statistic to the problem of comparing k independent samples, proposed by Kiefer (1959). (d) To calibrate the test, a bootstrap procedure is used. For this, under the null hypothesis, B resamples for each one of the factor levels are generated, and the corresponding B bootstrap statistics, D_i^*, i = 1, ..., B, are computed. (e) Finally, given a significance level α, the null hypothesis is rejected if D is larger than the 1 − α quantile of {D_1^*, ..., D_B^* }. The p-value of this test can also be approximated by Monte Carlo. A more detailed description of this algorithm is given in Appendix 3.

Both bandwidth selection methods, plug-in and bootstrap, for the nonparametric estimators (1) and (2) implemented in the corresponding functions of the `binnednp` package are described in Appendix 1. Moreover, Appendix 2 contains the steps of the bootstrap algorithm used in the `binnednp` package to compute confidence bands for the distribution function. Finally, Appendix 3 describes the statistical procedure implemented in the `binnednp` package to test whether a factor can be statistically significant.

2.2 | Emergence indices estimation

In this context, another interesting problem is that of finding the best soil depth at which to measure the HTT. For this, moment-based indices and probability density-based indices were proposed in Cao et al. (2011), and estimates of them are also implemented in the `binnednp` package. Some of these index estimators are based on nonparametric methods and require the selection of a bandwidth. Different techniques to automatically obtain approximately optimal bandwidths have also been included in the package.

In order to maximize the predictive power of the weed emergence models considered, one should choose the depth such that the density function of X, measuring the CHTT at emergence, is as flatter as possible (or the distribution of X has as much spread as possible). Taking this into account, two indices based on the moments of X, the coefficient of variation and the kurtosis of X, have been considered:

$$I_1 = \frac{\sigma}{\mu},$$

$$I_2 = \frac{m_4}{\sigma^4},$$

where $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathbb{V}[X]$, and $m_4 = \mathbb{E}[(X-\mu)^4]$ are the mean, the variance, and the fourth central moment of X, respectively. Large values for I_1 and small values for I_2 would be associated with a highly spread and light-tailed distribution and, therefore, are desirable for good weed emergence prediction properties. Indices based on the density of X, namely.

$$J_1 = \sigma^3 \int f'(x)^2 dx,$$

$$J_2 = \sigma^5 \int f''(x)^2 dx,$$

have been also considered.

The indices J_1 and J_2 measure the curvature of the distribution and density functions of X, respectively. Therefore, small values for both J_1 and J_2 are desirable.

3 | PACKAGE ARCHITECTURE

`binnednp` is an R package (R Development Core Team, 2019) designed for nonparametric estimation of both density and distribution functions of interval-grouped data. Although `binnednp` can be used for the analysis of any variable presented as grouped in intervals, the package and its structure were designed for its use by the weed science research community.

The package was developed using the Rcpp API (Eddelbuettel & François, 2011) which allows the integration of C++ code in R. The parts of the package code relative to bandwidth selection are quite consuming in terms of computation time, especially those that make use of bootstrapping. For this reason, writing portions of the package in C++ was crucial since it allows obtaining numerical results in a very short time. Moreover, the runtime of some of the functions can be further reduced (up to 60%) by means of parallelism with sockets. This was observed in a simulation study (not shown here for the sake of brevity) performed to analyze the CPU time of the functions of the package, when the sample size increases, and the effect of using (or not) parallel computing.

Regarding the structure of the package, it consists of the four functions described below (Sections 3.1–3.4). A more complete description and additional examples can be found in the reference manual of the package (Barreiro et al., 2019). Next, the following notation is considered in the arguments of those functions:

1. `n`: Number of seedlings that have emerged at the end of the experiment. In general, it is the size of the unknown complete sample.
2. `y`: Vector with the measurements of the CHTT at each inspection time. In general, this vector contains the endpoints of the intervals where the data are grouped.
3. `w (ni)`: Vector with the proportion (number) of seedlings that have emerged between each pair of consecutive CHTT. In general, each element of this vector indicates the proportion (number) of observations lying within each of the intervals where the data are grouped.

3.1 | Density estimation

```
bw.dens.binned(n, y, w, ni, gboot, pilot.type=3, hn=100,
plugin.type="N", confband=FALSE, alpha=0.05, B=1000,
plot = TRUE, print = TRUE, model, parallel = FALSE,
pars = new.env())
```

This function computes the plug-in and bootstrap bandwidths for the density estimator (1). Regarding the plug-in bandwidth, with the parameter plugin.type, the iterative process to estimate the bandwidth can be chosen. As for the bootstrap bandwidth, the parameter pilot.type allows the user to select the method to automatically compute the pilot bandwidth needed for the calculation of the bootstrap bandwidth, whereas the parameter gboot allows to manually select that pilot bandwidth. In most situations, it is recommended to employ the default values of these parameters. Additionally, the estimation process can be further personalized using parameters like hn, that determines the number of iterations done during the optimization stage, or B, that indicates the number of bootstrap replicates used for the construction of confidence bands, in case that confband = TRUE. Furthermore, if parallel = TRUE, confidence bands are estimated using parallel computing. Finally, for the sake of comparison with (1), the parameter model allows to fit different parametric families of distributions to the grouped sample. The parameters of these distributions are estimated by maximum likelihood.

3.2 | Distribution estimation

```
bw.dist.binned(n, y, w, ni, gplugin, type = "N",
confband = FALSE,
B = 1000, alpha = 0.05, plot = TRUE, print = TRUE,
model,
parallel = FALSE, pars = new.env())
```

This function computes the plug-in bandwidth for the distribution estimator (2). The parameter type allows the user to choose the iterative process to be used to estimate the bandwidth, whereas with the parameter `gplugin`, the bandwidth used in the last iteration, when `type = "A"`, can be manually selected. Due to the erratic behavior of the bandwidth selector with `type = "A"`, it is strongly recommended to compute the plug-in bandwidth using `type = "N"`. Anyway, the bootstrap bandwidth selector, computed with the function described below, has shown a better performance than any of the plug-in bandwidths in most scenarios. If `confband = TRUE`, bootstrap confidence bands are calculated considering `B` replicates and using parallel computing, in the case that `parallel = TRUE`.

Parameter `model` plays a similar role as in `bw.dens.binned`.

```
bw.dist.binned.boot(n, y, w, ni, g, pilot.type = 2,
nit = 10,
confband = FALSE, B = 1000, alpha = 0.05, print =
TRUE, plot = TRUE,
parallel = FALSE, pars = new.env())
```

This function computes the bootstrap bandwidth selector for the distribution estimator (2). The parameter `pilot.type` defines the method to select the pilot bandwidth used for the estimation of the final bandwidth, whereas the parameter `g` allows the user to manually select that pilot bandwidth. The parameter `nit` fixes the number of iterations to be done in the optimization stage. If `confband = TRUE`, bootstrap confidence bands are estimated considering `B` resamples and using parallel computing, in the case that `parallel = TRUE`.

3.3 | Emergence indices estimation

```
emergence.indices(n, y, w, ni, hseq, hn = 200, nmix =
4, B = 500,
method = "np", last.iter.np = F, confint = FALSE,
B.conf = 1000,
alpha = 0.05, print = TRUE, parallel = FALSE, pars =
new.env())
```

This function computes estimates for grouped data of the moment-based and density-based emergence indices presented in Section 2.2. In the case of the density-based indices, with the parameter `method`, the method to select the bandwidth used for their estimation can be chosen: if `method = "plugin"`, a plug-in approach is considered, whereas if `method = "mix"` or `method = "np"`, a parametric or nonparametric bootstrap approach is considered, respectively. If `confint = TRUE`, bootstrap confidence intervals are constructed considering `B` resamples and using parallel computing, in the case that `parallel = TRUE`.

3.4 | Analysis with different treatments

```
anv.binned(n, y, trt.w, abs.values = FALSE, B = 500)
```

This function allows to analyze whether a factor has a significant effect on the emergence curve. The idea behind this approach was briefly explained in Section 2.1, and it is described in detail in Appendix 3. It consists in using a bootstrap approach and a Cramér-von-Mises type distance.

In this case, `n` is a vector composed of the sizes of the complete samples corresponding to each treatment, and `trt.w` is a matrix each of whose columns contains the proportion of observations lying within each of the intervals for the corresponding treatment. If instead of proportions the user wants to provide absolute values, the parameter `abs.values` must be set to TRUE. Furthermore, the user can choose the number of bootstrap resamples through the parameter `B`. The function `anv.binned` returns the *p*-value of the test.

4 | EXAMPLE

In this section, an unpublished data set of wild oat (*Avena sterilis* L.) emergence is considered to illustrate the use of the `binnednp` Rcpp package. These data were taken from an experiment performed during Winter–Spring 2006–2007 in Gibraleon (37°C 22′N, 6°C 54′W; altitude 26 m), located in the province of Huelva (Andalucia, South of Spain).

Briefly, the experiment consisted in four polyvinylchloride cylinders (250 mm diameter 50 mm height) placed 1 m apart. For each sample, 200 seeds of *A. sterilis* were mixed thoroughly with the soil and distributed over the 0–100 mm depth. Numbers of emerged weed seedlings were recorded once or twice a week and then removed by cutting seedling stems at ground level with minimum disturbance of the substrate. All the data for the cumulative numbers of seedling emergence from the field were converted to a square meter

TABLE 1 Seedling emergence data of *Avena sterilis*

	CHTTT			No. seedlings				
	Depth			Cylinder				
Date	10 mm	20 mm	50 mm	1	2	3	4	Pooled
27 November 2006	100	92	67	0	0	0	0	0
4 December 2006	160	146	105	0	0	0	0	0
12 December 2006	218	199	143	2	6	8	3	19
14 December 2006	218	217	155	1	0	0	1	2
19 December 2006	218	217	185	2	1	1	3	7
22 December 2006	218	217	199	2	1	1	0	4
26 December 2006	218	217	204	1	1	0	0	2
28 December 2006	218	217	204	0	0	0	0	0
2 January 2007	218	217	204	0	0	0	0	0
5 January 2007	218	217	204	0	2	0	0	2
9 January 2007	218	217	204	2	2	9	2	15
12 January 2007	218	217	204	3	7	18	11	39
18 January 2007	218	217	204	12	7	19	22	60
25 January 2007	218	217	204	6	5	8	13	32
1 February 2007	265	261	232	2	5	7	7	21
9 February 2007	352	340	287	13	12	5	8	38
15 February 2007	405	421	343	7	12	13	4	36
23 February 2007	459	505	421	0	0	1	0	1
5 March 2007	509	571	538	0	0	0	0	0
19 March 2007	509	571	538	0	0	0	0	0
Num emerged seedlings				53	61	90	74	*n* = 278

basis. Additionally, following the same procedure as that described in Cao et al. (2011), the CHTT at emergence in the different inspection days, at three depths (10, 20, and 50 mm), was calculated.

The observed emergence data are shown in Table 1. As it can be seen, the cumulative hydrothermal time at emergence cannot be observed for every individual seed, but just in an aggregated way.

In the first part of the study, we use the function `anv.binned` to evaluate whether a factor (in this case, “the cylinder factor") has a significant effect on the emergence curve, for any of the three depths. As pointed out in Section 2.1, the test implemented in the function `anv.binned` is based on a Cramér-von Mises distance between the nonparametric estimator with the pooled sample and the nonparametric estimators in the different levels of the factor. For example, in the case of depth 10 mm, the distance between all these curves could be visually observed comparing Figures 1 and 3. Figure 1 contains the nonparametric emergence curve estimates with bootstrap bandwidths (jointly with the corresponding 95% bootstrap confidence bands) for each one of the four cylinders. Figure 3 depicts the emergence curve estimates using the nonparametric approach with the pooled sample. However, it is clear that a reliable and formal solution for this problem requires the application of a statistical test, such as that implemented in the function `anv.binned`.

Note that, in this case, the identical experimentation conditions carried out in the four cylinders seem to support the idea that the null hypothesis could be true and, for this, only a very strong evidence against it will lead us to reject the null hypothesis of “nonsignificant cylinder effect."

Denoting by `y1`, `y2`, and `y3` the CHTT was calculated for 10, 20, and 50 mm, respectively, and after applying the function `anv.binned` for the three depths, using the following code:

```
# Observed values of CHTT for each depth
y1 = c(100,160,218,265,352,405,459,509)
y2 = c(92,146,199,217,261,340,421,505,571)
y3 = c(67,105,143,155,185,199,204,232,287,343,421,538)

# size of the complete sample for each treatment
n = c(53,61,90,74)

# nij: number of emerged seedlings for treatment
i and depth j
n11 = c(0,0,31,2,13,7,0)
n21 = c(0,0,32,5,12,12,0)
n31 = c(0,0,64,7,5,13,1)
n41 = c(0,0,55,7,8,4,0)

# wij: proportion of emerged seedlings for treat-
ment i and depth j
w11 = n11/n[1]
```

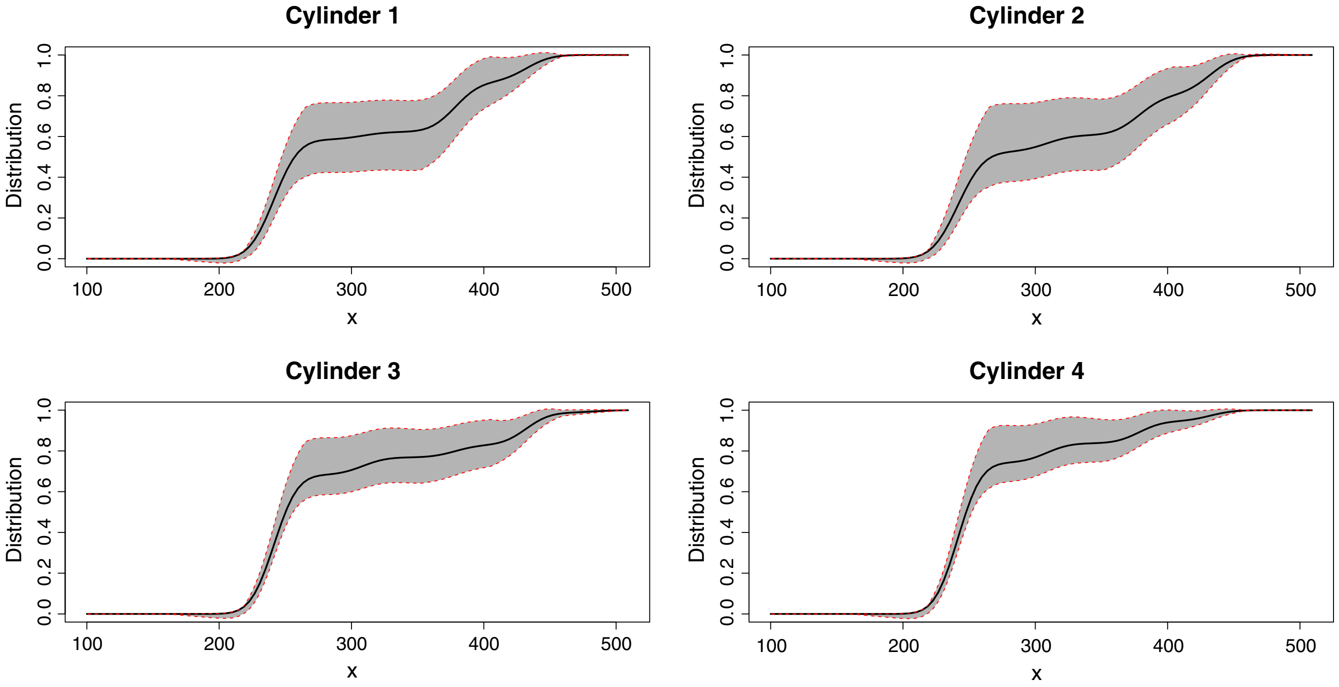

FIGURE 1 Nonparametric estimates of the emergence curves using the samples in each of the four cylinders, jointly with the corresponding 95% bootstrap confidence bands

```
w21 = n21/n[2]
w31 = n31/n[3]
w41 = n41/n[4]

# Analysis with different treatments for 10 mm
res1 = anv.binned(n,y1,cbind(w11,w21,w31,w41),B=1000)
n12 = c(0,0,2,29,2,13,7,0)
n22 = c(0,0,6,33,5,12,12,0)
n32 = c(0,0,8,56,7,5,13,1)
n42 = c(0,0,3,42,7,8,4,0)

w12 = n12/n[1]
w22 = n22/n[2]
w32 = n32/n[3]
w42 = n42/n[4]

# Analysis with different treatments for 20 mm
res2 = anv.binned(n,y2,cbind(w12,w22,w32,w42),B=1000)

n13 = c(0,0,2,1,2,2,24,2,13,7,0)
n23 = c(0,0,6,0,1,1,24,5,12,12,0)
n33 = c(0,0,8,0,1,1,54,7,5,13,1)
n43 = c(0,0,3,1,3,0,48,7,8,4,0)

w13 = n13/n[1]
w23 = n23/n[2]
w33 = n33/n[3]
w43 = n43/n[4]
# Analysis with different treatments for 50 mm
res3 = anv.binned(n,y3,cbind(w13,w23,w33,w43),B=1000)
```

the results obtained are:

----- Result of the application of anv.binned to the samples of the four cylinders -----

```
> res1
[1] 0.016

> res2
[1] 0.144

> res3
[1] 0.011
```

This indicates that, with a significance level of 0.01, the effect of the cylinders is not significant for any of the three depths. The strong certainty of the null hypothesis justifies the use of this significance level (see, e.g. Cramer & Howit, 2004, p. 151, for a comment on the election of the significance level). Therefore, it makes sense to analyze the samples jointly and not separately.

Taking into account the result obtained after applying function `anv.binned` to the samples of the four cylinders, in the second part of the study, the data with the pooled number of emerged seedlings, given in the last column of Table 1, are considered. Note that the total sample size of emerged seedlings at the end of the experiment is n =278.

Given these weed emergence data, a first interesting issue is to find out what is the best depth, among the three possibilities available in this case, 10, 20, and 50 mm, to measure the CHTT in order to have more prediction power. Denoting, as before, by `y1`, `y2`, and `y3` the CHTT calculated for 10, 20, and 50 mm, respectively, and by `ni` the vector with pooled number of emerged seedlings, the function `emergence.indices` can be applied to (`y1`, `ni`), (`y2`, `ni`), and (`y3`, `ni`) to obtain estimates of the indices presented in Section 2.2.

```
ind1 <- emergence.indices(n, y1, ni) # emergence
indices for 10 mm
ind2 <- emergence.indices(n, y2, ni) # emergence
indices for 20 mm
ind3 <- emergence.indices(n, y3, ni) # emergence
indices for 50 mm

indices <- data.frame(I1=c(ind1$I1,ind2$I
1,ind3$I1), I2=c(ind1$I2,ind2$I2,ind3$I2),
J1=c(ind1$J1,ind2$J1,ind3$J1),
J2=c(ind1$J2,ind2$J2,ind3$J2))
```

Obtaining the results:

```
Estimates of emergence indices for each depth

          10 mm         20 mm         50 mm
I1   0.2995597     0.2632137     0.2225033
I2   2.4839678     2.8994871     3.0647913
J1   3.3301301     9.2641671    21.0453859
J2  81.0092447   541.0724742  2057.1117305
```

To maximize the predictive power a specific weed emergence model, the HTT should be measured at a depth producing a density function as flat as possible or, equivalently, a distribution function as dispersed as possible. Therefore, small values of J_1 and J_2 are preferable. On the other hand, CHTT samples with higher coefficient of variation (higher value of I_1) and a lower kurtosis (lower value of I_2) will improve weed emergence prediction. Consequently, 10 mm seems to be the best soil depth to predict weed emergence in terms of indices I_1, I_2, J_1, and J_2 and, therefore, only observations at 10 mm are considered in what follows.

Now, the function `bw.dens.binned`, described in Section 3.1, can be applied using the pooled sample and the CHTT at 10 mm to compute the plug-in and bootstrap bandwidths for the kernel density estimator (1).

```
# computing bandwidths for kernel density estimator
dens <- bw.dens.binned (n, y, ni, plot = FALSE) #
Bandwidths for density
```

Obtaining the results:

```
-----Plug-in and bootstrap bandwidths for the density
estimator-----
> dens
$h _ plugin
[1] 14.93051
$h _ boot
[1] 19.68239
```

Figure 2 shows, in the left panel, the kernel density estimates computed using (1) with the obtained plug-in (green lines) and bootstrap (red lines) bandwidths. Moreover, the parameter `model` was used in `bw.dens.binned` to fit parametric logistic and Weibull densities to the emergence data. The right panel of Figure 2 shows the corresponding density estimators (green and blue lines for the Weibull and the logistic densities, respectively). Additionally, the nonparametric estimator computed with the bootstrap bandwidth is also included in this figure (red line) for the sake of comparison.

Next, using the functions `bw.dist.binned` and `bw.dist.binned.boot`, described in Section 3.2, the plug-in and bootstrap bandwidths for the kernel distribution estimator (2) are calculated.

```
# computing bandwidths for kernel distribution
estimator

# Plug-in bandwidth
dist _ pi <- bw. dist. binned (n, y, ni, plot =
FALSE) $h # Plug-in

# Bootstrap bandwidth
dist _ boot <- bw. dist. binned. boot (n, y, ni,
plot = FALSE)$h # Bootstrap
```

Obtaining the results:

```
-----Plug-in and bootstrap bandwidths for the distribution
estimator-----
> dist _ pi$h
[1] 9.833762
> dist _ boot$h
[1] 13.73533
```

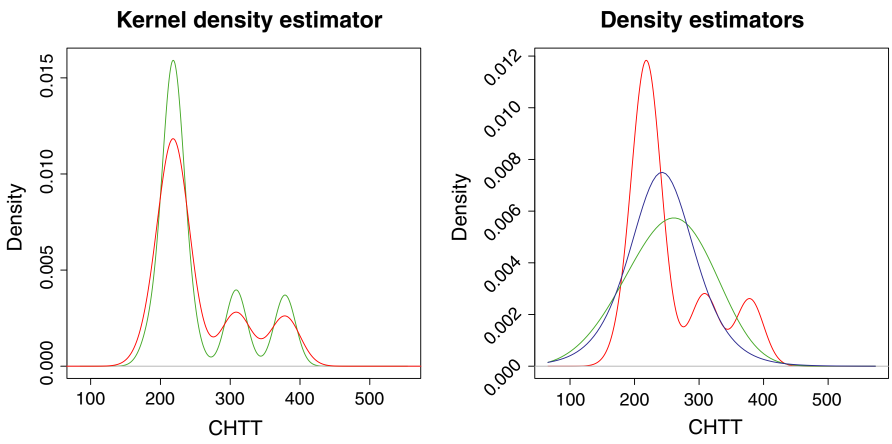

FIGURE 2 Left panel: kernel density estimates considering plug-in (green line) and bootstrap (red line) bandwidths. Right panel: parametric Weibull (green line) and logistic (blue line) density estimates, and nonparametric kernel density estimate using the bootstrap bandwidth (red line)

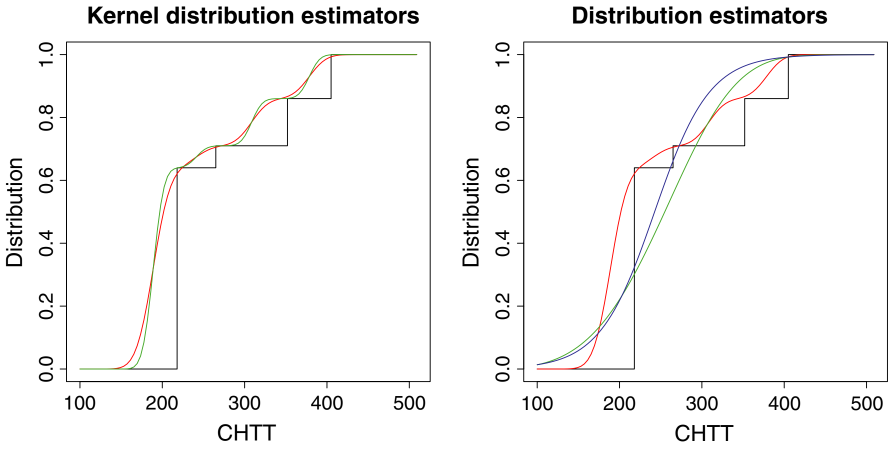

FIGURE 3 Kernel distribution estimates considering plug-in (green line) and bootstrap (red line) bandwidths. Right panel: parametric regression fits, Weibull (green line) and logistic (blue line), and nonparametric kernel distribution estimate using the bootstrap bandwidth (red line). The empirical distribution of the grouped sample (black lines) is also shown

Figure 3 shows, in the left panel, the kernel distribution estimates computed using (2) with the obtained plug-in (green line) and bootstrap (red line) bandwidths, that is, the estimates of the emergence curves using the nonparametric approach. In contrast to the density case, the effect that the bandwidth has on the behavior of the distribution estimator is less evident, since slightly different bandwidths produce very similar distribution estimates. As in the density case, the parameter `model` in the function `bw.dist.binned` was set to `weibull` and `logistic` to fit parametric regression functions following these models to describe seedling emergence. The corresponding fits are shown in the right panel of Figure 3, using a green line for the Weibull and a blue line for the logistic. The nonparametric distribution estimator (2) is also included in this plot (red line). In both figures, the empirical distribution of the grouped sample data is represented with black lines.

5 | CONCLUSION

The `binnednp` R package gives the weed science research community a simple tool to analyze interval-grouped data. This is useful to study, for example, the CHTT at seedling emergence. Using nonparametric density and distribution estimation, the researcher can both visualize the underlying nature of the data and make predictions without loosing flexibility or making inadequate assumptions about the data. Moreover, estimation of emergence indices measures the adequacy of the depth chosen to register the values of CHTT. Additionally, analyses to test the effect of a specific factor on weed emergence can be also performed.

ACKNOWLEDGEMENTS

This research has been partially supported by MINECO grants MTM2014-52876-R and MTM2017-82724-R, for the first three authors, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), for the first fifth authors, all of them through the European Regional Development Funds (ERDF). The research of sixth author has been partially funded by FEDER (ERDF) and MINECO grants AGL2012-33736 and AGL2015-64130-R. The authors thank Dr. Fernando Bastida for permission to use the *Avena sterilis* L. emergence data employed in Section 4. The authors also thank two anonymous referees for numerous useful comments that significantly improved this article.

CONFLICT OF INTEREST

None declared.

AUTHOR'S CONTRIBUTIONS

RC, MF-F, and JLG-A conceived the ideas and designed the methodology; DB-U, BBF, RD, and MR wrote the code; JLG-A provided the data; DB-U, RC, MF-F, MR, and JLG-A analyzed the data. All authors led the writing of the manuscript and contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

The current stable version of the `binnednp` package requires R (≥2.10) and is distributed under the GPL-3 license. It is publicly available on the Comprehensive R Archive Network at https://cran.r-project.org/package=binnednp

ORCID

Daniel Barreiro-Ures https://orcid.org/0000-0003-4930-3603
Mario Francisco-Fernández https://orcid.org/0000-0002-9201-5423
Ricardo Cao https://orcid.org/0000-0001-8304-687X
Basilio B. Fraguela https://orcid.org/0000-0002-3438-5960
Ramón Doallo https://orcid.org/0000-0002-6011-3387
José Luis González-Andújar https://orcid.org/0000-0003-2356-4098
Miguel Reyes https://orcid.org/0000-0002-1460-1814

REFERENCES

Barreiro, D., Fraguela, B., Doallo, R., Cao, R., Francisco-Fernández, M., & Reyes, M.. (2019). *binnednp: nonparametric estimation for interval-grouped data*. Retrieved from https://cran.r-project.org/package=binnednp, R package version 0.4.0.

Blanco, A. M., Chantre, G. R., Lodovichi, M. V., Bandoni, J. A., López, R. L., Vigna, M. R., … Sabbatini, M. R. (2014). Modelling seed dormancy release and germination for predicting *Avena fatua* L. field emergence: A genetic algorithm approach. *Ecological Modelling*, *272*, 293–300.

Bradford, K. J. (2002). Applications of hydrothermal time to quantifying and modeling seed germination and dormancy. *Weed Science*, *50*, 248–260. https://doi.org/10.1614/0043-1745(2002)050[0248:AOHTTQ]2.0.CO;2

Cao, R., Francisco-Fernández, M., Anand, A., Bastida, F., & González-Andújar, J. L. (2011). Computing statistical indices for hydrothermal times using weed emergence data. *Journal of Agricultural Science*, *149*, 701–712. https://doi.org/10.1017/S002185961100030X

Cao, R., Francisco-Fernández, M., Anand, A., Bastida, F., & González-Andújar, J. L. (2013). Modeling *Bromus diandrus* seedling emergence using nonparametric estimation. *Journal of Agricultural, Biological and Environmental Statistics*, *18*, 64–86. https://doi.org/10.1007/s13253-012-0122-x

Chantre, G. R., Blanco, A. M., Lodovichi, M. V., Bandoni, J. A., Sabbatini, M. R., López, R. L., … Gigón, R. (2012). Modeling Avena fatua seedling emergence dynamics: An artificial neural network approach. *Computers and Electronics in Agriculture*, *88*, 95–102. https://doi.org/10.1016/j.compag.2012.07.005

Cramer, D., & Howit, D. (2004). *The SAGE dictionary of statistics*. London: SAGE Publications.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18.

Forcella, F., Benech-Arnold, R., Sánchez, R., & Ghersa, C. (2000). Modelling seedling emergence. *Field Crops Research*, *67*, 123–139.

González-Andújar, J. L. (2008). Weed control models. In S. E. Jørgensen, & B. D. Fath (Eds.), *Population dynamics. Vol. [5] of encyclopedia of ecology* (pp. 3776–3780). Oxford: Elsevier.

González-Andújar, J. L., Chantre, G. R., Morvillo, C., Blanco, A., & Forcella, F. (2016). Predicting field weed emergence with empirical models and soft computing techniques. *Weed Research*, *56*, 415–423. https://doi.org/10.1111/wre.12223

González-Andújar, J. L., Francisco-Fernández, M., Cao, R., Reyes, M. A., Urbano, J. M., Forcella, F., & Bastida, F. (2016). A comparative study between nonlinear regression and nonparametric approaches for modeling *Phalaris paradoxa* seedling emergence. *Weed Research*, *56*, 367–376.

Grundy, A. C. (2003). Predicting weed emergence: A review of approaches and future challenges. *Weed Research*, *43*, 1–11. https://doi.org/10.1046/j.1365-3180.2003.00317.x

Haj Seyed-Hadi, M. R., & Gonzalez-Andujar, J. L. (2009). Comparison of fitting weed seedling emergence models with nonlinear regression and genetic algorithm. *Computers and Electronics in Agriculture*, *65*, 19–25.

Hunter, E. A., Glasbey, C. A., & Naylor, R. E. L. (1984). The analysis of data from germination tests. *Journal of Agricultural Science, Cambridge*, *102*, 207–213. https://doi.org/10.1017/S0021859600041642

Kiefer, J. (1959). *k*-Sample analogues of the Kolmogorov-Smirnov, Cramér-von Mises tests. *Annals of Mathematical Statistics*, *30*, 420–447.

Naylor, R. E. L. (1981). An evaluation of various germination indices for predicting differences in seed vigour in Italian ryegrass. *Seed Science and Technology*, *9*, 593–600.

Onofri, A., Gresta, F., & Tei, F. (2010). A new method for the analysis of germination and emergence data of weed species. *Weed Research*, *50*, 187–198. https://doi.org/10.1111/j.1365-3180.2010.00776.x

Onofri, A., Mesgaran, M. B., Tei, F., & Cousens, R. D. (2011). The cure model: An improvement way to describe seed germination. *Weed Research*, *51*, 516–524.

Onofri, A., Piepho, H.-P., & Kozak, M. (2019). Analysing censored data in agricultural research: A review with examples and software tips. *Annals of Applied Biology*, *174*, 3–13. https://doi.org/10.1111/aab.12477

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, *32*, 1065–1076. https://doi.org/10.1214/aoms/1177704472

R Development Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Reyes, M., Francisco-Fernández, M., & Cao, R. (2016). Nonparametric kernel density estimation for general grouped data. *Journal of Nonparametric Statistics*, *2*, 235–249.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*, 832–837. https://doi.org/10.1214/aoms/1177728190

Westfall, P. H., & Young, S. S. (1993). *Resamplig-based multiple testing: Examples and methods for p-value adjustment*. New York: John Wiley & Sons Inc.

Zambrano-Navea, C., Bastida, F., & González-Andújar, J. L. (2013). A hydrothermal seedling emergence model for *Conyza bonariensis*. *Weed Research*, *53*, 213–220.

How to cite this article: Barreiro-Ures D, Francisco-Fernández M, Cao R, et al. Analysis of interval-grouped data in weed science: The `binnednp` Rcpp package. *Ecol Evol*. 2019;9:10903–10915. https://doi.org/10.1002/ece3.5448

APPENDIX 1

In this appendix, the plug-in and bootstrap bandwidth selection methods implemented in the functions `bw.dens.binned`, bw.dist.binned, and bw.dist.binned.boot are briefly described.

BANDWIDTH SELECTION FOR THE DENSITY FUNCTION

The function `bw.dens.binned` of the `binnednp` package computes plug-in and bootstrap bandwidths for the density estimator (1). Plug-in bandwidths are obtained minimizing in h the expression of the asymptotic mean squared error (AMISE) of (1) and estimating the unknown quantities in the minimizer of the AMISE. Under suitable assumptions, asymptotic properties of (1) were obtained in Reyes et al. (2016). In that paper, it is obtained that the AMISE of (1) is:

$$\mathrm{AMISE}(\hat{f}_h^g) = \frac{1}{4}\mu_2(K)^2 h^4 A(f'') + \frac{1}{nh}A(K), \tag{A1}$$

where $\mu_2(K) = \int x^2 K(x)\mathrm{d}x > 0$, $A(K) = \int K^2(x)\mathrm{d}x$ and $A(f\prime\prime) = \int f''(x)^2\mathrm{d}x$.

From (A1), it follows that the asymptotically optimal global bandwidth for (1) is

$$h_g^{\mathrm{dens}} = \left[\frac{A(K)}{\mu_2(K)^2 A(f'')\, n}\right]^{\frac{1}{5}}. \tag{A2}$$

Using (A2) and estimating the unknown quantity $A(f'')$, a plug-in bandwidth for estimator (1) can be derived. In Cao et al. (2011), a nonparametric estimator for $A(f'')$, denoted by $\hat{A}_\eta^g$, depending on an auxiliary smoothing parameter η was proposed. Plugging $\hat{A}_\eta^g$ in (A2) gives the plug-in bandwidth selector for $\hat{f}_h^g$:

$$\hat{h}_g^{\text{dens}} = \left[\frac{A(K)}{\mu_2(K)^2 \hat{A}_\eta^g n} \right]^{\frac{1}{5}}.$$

In practice, to calculate $\hat{A}_\eta^g$ a new bandwidth η has to be selected. This pilot smoothing parameter can be selected using again a plug-in procedure, appearing new auxiliary bandwidths, and so on. The usual strategy is to stop this iterative process after two steps and estimate the unknown quantities assuming that f is Gaussian.

As for the bootstrap bandwidth selector for (1) implemented in `bw.dens.binned`, this is obtained minimizing the bootstrap version of the mean integrated squared error (MISE) of $\hat{f}_h^g$.

Using standard calculations, a closed form expression for the MISE of estimator (1) is given by:

$$\begin{aligned} \text{MISE}\left(\hat{f}_h^g\right) &= \mathbb{E}\left[\int \left(\hat{f}_h^g(x) - f(x)\right)^2 dx\right] \\ &= \int \left[\sum_{i=1}^{k} p_i K_h(x - t_i) - f(x)\right]^2 dx \qquad \text{(A3)} \\ &+ \frac{A(K)}{nh} - \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{k} p_i p_j (K * K)_h (t_i - t_j), \end{aligned}$$

where $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$, $p_j = F(y_j) - F(y_{j-1})$, for $j = 1, \ldots, k$, and symbol * stands for convolution.

The bootstrap version MISE* of the MISE is obtained as follows. Let

$$\hat{f}_\zeta^g(x) = \sum_{i=1}^{k} w_i K_\zeta (x - t_i)$$

be the estimator (1) based on a pilot bandwidth ζ. Draw a bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{f}_\zeta^g$ and, given a bandwidth h, consider the analogue of the kernel density estimator, $\hat{f}_h^{g*}(x) = \sum_{i=1}^{k} w_i^* K_h (x - t_i)$, where $w_i^* = F_n^*(y_i -) - F_n^*(y_{i-1} -)$, with $F_n^*(y) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i^* \le y\}}$. Then

$$\text{MISE}^*\left(\hat{f}_h^{g*}\right) = \mathbb{E}^*\left[\int \left(\hat{f}_h^{g*}(x) - \hat{f}_\zeta^g(x)\right)^2 dx\right], \qquad \text{(A4)}$$

where $\mathbb{E}^*$ denotes the bootstrap expectation (with respect to $\hat{f}_\zeta^g$).

Using a parallel process to that followed to obtain (3), it is possible to derive a closed representation for (4). This expression is given by

$$\begin{aligned} \text{MISE}^*\left(\hat{f}_h^{g*}\right) &= \frac{n-1}{n} \sum_{i=1}^{k} \sum_{j=1}^{k} w_i^\zeta w_j^\zeta (K * K)_h (t_i - t_j) \\ &-2 \sum_{i=1}^{k} \sum_{j=1}^{k} w_i^\zeta w_j (K_h * K_\zeta)(t_i - t_j) \qquad \text{(A5)} \\ &+ \sum_{i=1}^{k} \sum_{j=1}^{k} w_i w_j (K * K)_\zeta (t_i - t_j) + \frac{A(K)}{nh}. \end{aligned}$$

Note that expression (A5) allows us to directly evaluate MISE* over a grid of values of h, without using Monte Carlo. The bootstrap bandwidth h_{MISE}^* is obtained minimizing (A5), that is,

$$h_{\text{dens}}^* = \underset{h>0}{\text{argmin}}\, \text{MISE}^*\left(\hat{f}_h^{g*}\right),$$

To select the pilot bandwidth ζ, the method implemented in the `binnednp` package is inspired by the idea of smoothing splines, based on selecting the pilot parameter that minimizes the squared distance between the nonparametric density estimator, $\hat{f}_h^g$, and the histogram of the grouped data, $\hat{H}$, plus a penalty term to avoid obtaining very small bandwidths. The idea consists in finding the parameter denoted by $\zeta_{\text{hist}}^\lambda$, such that,

$$\zeta_{\text{hist}}^\lambda = \underset{h>0}{\text{argmin}} \sum_{i=1}^{k} (\hat{H}(t_i) - \hat{f}_h^g(t_i))^2 + \lambda \int \hat{f}_h^{g''}(x)^2 dx,$$

where $\lambda \ge 0$ determines the penalty degree over the global curvature of the nonparametric density estimator, defined in (1). To select an ``optimal'' penalty degree, λ_{opt}, we have used the rule of finding the penalty allowing to obtain a pilot bandwidth that best approximate the overall curvature of the population density, that is,

$$\lambda_{\text{opt}} = \underset{\lambda \ge 0}{\text{argmin}} \left| A(\hat{f}_{\zeta_{\text{hist}}^\lambda}^{g''}) - A(f'') \right|.$$

In practice, λ_{opt} can be estimated by

$$\hat{\lambda}_{\text{opt}} = \underset{\lambda \ge 0}{\text{argmin}} \left| A(\hat{f}_{\zeta_{\text{hist}}^\lambda}^{g''}) - A(\hat{f}_{\text{mix}}'') \right|,$$

where $\hat{f}_{\text{mix}}$ is a normal mixture model with a maximum number of $r = 5$ components fitted with the grouped-data sample. In practice, the expectation–maximization (EM) method was used to estimate the parameters of the mixture model, using the BIC criterion to select the best fit.

BANDWIDTH SELECTION FOR THE DISTRIBUTION FUNCTION

Following analogous arguments to those described for the case of the density function, plug-in and bootstrap bandwidth selectors can be proposed for the nonparametric distribution estimator given in (2). They are implemented in the functions bw.dist.binned and bw.dist.binned.boot, respectively.

Regarding the plug-in bandwidth, under some assumptions, it can be obtained that the AMISE of (2) is:

$$\text{AMISE}\left(\hat{F}_h^g\right) = \frac{h^4}{4} \mu_2(K)^2 A(f') + \frac{1}{n} \int F(x)\left[1 - F(x)\right] dx - \frac{h}{n} C_0 \qquad \text{(A6)}$$

with $A(f') = \int f'(x)^2 dx$, and

$$C_0 = 2 \int z K(z) \mathbb{K}(z) dz > 0.$$

From equation (A6), it is immediate to get an asymptotically optimal global bandwidth for $\hat{F}_h^g$. Taking the first derivative of (A6), equating to zero and solving for h, it is obtained

$$h_g^{\text{dist}} = \left[\frac{C_0}{\mu_2(K)^2 A(f')\, n} \right]^{\frac{1}{3}}. \qquad \text{(A7)}$$

In equation (A7), an estimate of $A(f')$ is required to have a practical bandwidth. Using a similar process to that described for the case of the density estimator, a practical plug-in bandwidth selector for $\hat{F}_h^g(x)$ is obtained:

$$\hat{h}_g^{\text{dist}} = \left[\frac{C_0}{\mu_2(K)^2 \tilde{A}_\eta^g\, n} \right]^{\frac{1}{3}},$$

where $\tilde{A}_\eta^g$ denotes an estimator of $A(f')$ depending on a pilot bandwidthη. A similar iterative procedure to that described in the case of the density function is used here.

On the other hand, the bootstrap bandwidth implemented in the function bw.dist.binned.boot is based on minimizing the bootstrap version of the MISE of $\hat{F}_h^g$.

In this case, using standard calculations and assuming that $F(y_k) = 1$ and $F(y_0) = 0$, it is easy to prove that the expectation and the variance of $\hat{F}_h^g(x)$ are, respectively,

$$\mathbb{E}\left[\hat{F}_h^g(x)\right] = \sum_{i=1}^{k} \mathbb{K}\left(\frac{x - t_i}{h}\right) p_i \qquad \text{(A8)}$$

and

$$\mathbb{V}\left[\hat{F}_h^g(x)\right] = \frac{1}{n} \sum_{i=1}^{k} \mathbb{K}^2\left(\frac{x - t_i}{h}\right) p_i (1 - p_i) - \frac{2}{n} \sum_{i<j} \mathbb{K}\left(\frac{x - t_i}{h}\right) \mathbb{K}\left(\frac{x - t_j}{h}\right) p_i p_j. \qquad \text{(A9)}$$

From (A8) and (A9), it is straightforward to obtain a closed expression for the MISE of the estimator defined in (2):

$$\text{MISE}\left(\hat{F}_h^g\right) = \mathbb{E}\left\{ \int \left[\hat{F}_h^g(x) - F(x)\right]^2 dx \right\} = B + V,$$

where,

$$B = \int \left\{ \mathbb{E}\left[\hat{F}_h^g(x)\right] - F(x) \right\}^2 dx$$

denotes the integrated squared bias and

$$V = \int \mathbb{V}\left[\hat{F}_h^g(x)\right] dx$$

is the integrated variance.

To build a bootstrap version of MISE, we consider a pilot bandwidth, ζ, and construct the grouped-data smooth estimator of F as defined in (2), but replacing h by ζ. The idea is to draw resamples from $\hat{F}_\zeta^g$, to group the data and to compute the estimator $\hat{F}_h^g$ with those bootstrap samples. The bootstrap resampling plan proceeds as follows.

1. Fix some pilot bandwidth, ζ, and consider the grouped-data smooth cdf estimator, $\hat{F}_\zeta^g$.

2. Draw $(n_1^*, \dots, n_k^*)$ from a multinomial distribution $M_k(n; \tilde{p}_1^\zeta, \dots, \tilde{p}_k^\zeta)$ with $\tilde{p}_i^\zeta = \hat{F}_\zeta^g(y_i) - \hat{F}_\zeta^g(y_{i-1}), i = 1, \cdots, k$, and define $w_i^* = n_i^*/n$.

3. Compute the grouped-data smooth cdf estimator based on this bootstrap resample:

$$\hat{F}_h^{g*}(x) = \sum_{i=1}^{k} w_i^* \mathbb{K}\left(\frac{x - t_i}{h}\right).$$

4. Define the bootstrap version of MISE

$$\text{MISE}^*\left(\hat{F}_h^{g*}\right) = \mathbb{E}^*\left\{ \int \left[\hat{F}_h^{g*}(x) - \hat{F}_\zeta^g(x)\right]^2 dx \right\},$$

where, $\mathbb{E}^*$ denotes the bootstrap expectation (with respect to $\hat{F}_\zeta^g$). Therefore, defining

$$\hat{p}_i^\zeta = \frac{\tilde{p}_i^\zeta}{\sum_{j=1}^{k} \tilde{p}_j^\zeta}, i = 1, 2, \dots, k,$$

and substituting p_i by $\hat{p}_i^\zeta$ in (A8) and (A9), the bootstrap version of the MISE admits the following closed expression:

$$\text{MISE}^*\left(\hat{F}_h^{g*}\right) = \mathbb{E}^*\left\{ \int \left[\hat{F}_h^{g*}(x) - \hat{F}_\zeta^g(x)\right]^2 dx \right\} = B^* + V^*,$$

where

$$B^* = \int \left\{ \mathbb{E}^*\left[\hat{F}_h^{g*}(x)\right] - \hat{F}_\zeta^g(x) \right\}^2 dx$$

and

$$V^* = \int \mathbb{V}^*\left[\hat{F}_h^{g*}(x)\right] dx,$$

with

$$\mathbb{E}^*\left[\hat{F}_h^{g*}(x)\right] = \sum_{i=1}^{k} \mathbb{K}\left(\frac{x - t_i}{h}\right) \hat{p}_i^\zeta$$

and

$$\mathbb{V}^*\left[\hat{F}_h^{g*}(x)\right] = \frac{1}{n} \sum_{i=1}^{k} \mathbb{K}^2\left(\frac{x - t_i}{h}\right) \hat{p}_i^\zeta \left(1 - \hat{p}_i^\zeta\right) - \frac{2}{n} \sum_{i<j} \mathbb{K}\left(\frac{x - t_i}{h}\right) \mathbb{K}\left(\frac{x - t_j}{h}\right) \hat{p}_i^\zeta \hat{p}_j^\zeta.$$

Finally, the bootstrap bandwidth is defined as the minimizer of $\text{MISE}^*\left(\hat{F}_h^{g*}\right)$, in the smoothing parameter, h:

$$h_{\text{dist}}^* = \underset{h>0}{\operatorname{argmin}}\, \text{MISE}^*\left(\hat{F}_h^{g*}\right).$$

As in the case of the bootstrap bandwidth selector for the density function, an important step in this bootstrap procedure is that of selecting the pilot bandwidth ζ. The same type of procedure inspired by the idea of smoothing splines, but adapted for the distribution function, was employed. In this case, the idea consists in finding the parameter, denoted by $\zeta^{\lambda}_{\text{emp}}$, such that,

$$\zeta^{\lambda}_{\text{emp}} = \underset{h>0}{\text{argmin}} \sum_{i=0}^{k} \left[F_n(y_i) - \hat{F}^{g}_{h}(y_i)\right]^2 + \lambda \int \hat{f}^{g\prime}_{h}(x)^2 \, dx,$$

where $\lambda \geq 0$ determines the penalty degree over the global slope of the nonparametric density estimator, defined in (2). To select an “optimal” penalty degree, λ_{opt}, we have used the rule of finding the penalty allowing to obtain a pilot bandwidth that best approximate the overall slope of the population density, that is,

$$\lambda_{\text{opt}} = \underset{\lambda \geq 0}{\text{argmin}} \left| A\left(\hat{f}^{g\prime}_{\zeta^{\lambda}_{\text{emp}}}\right) - A(f^{\prime}) \right|.$$

In practice, λ_{opt} can be estimated by

$$\hat{\lambda}_{\text{opt}} = \underset{\lambda \geq 0}{\text{argmin}} \left| A\left(\hat{f}^{g\prime}_{\zeta^{\lambda}_{\text{emp}}}\right) - A(\hat{f}^{\prime}_{\theta}) \right|,$$

where $\hat{f}^{\prime}_{\theta}$ represents a parametric estimator of the first derivative of the density function, fitted with the grouped-data sample and flexible enough to capture, at least partially, the global slope of f. It was checked that fitting normal mixture models with a maximum number of r = 5 components provided, in general, very good results. In practice, the expectation--maximization (EM) method was used to estimate the parameters of these models, using the BIC criterion to select the best fit.

APPENDIX 2

As an option, the functions bw.dens.binneC, bw.dist.binned, and bw.dist.binned.boot allow to compute bootstrap confidence bands for the density and distribution functions, using the nonparametric estimators (1) and (2). This bands are designed to contain the whole function with a prescribed high probability, typically 95%, and can be employed, for example, to assess the uncertainty of the corresponding estimates or to perform analyses with different treatments. The procedures used for the density and the distribution functions follow the same steps, and therefore, for the sake of brevity, in this appendix, only the algorithm used in bw.dist.binned and bw.dist.binned.boot to compute the bootstrap confidence bands for the distribution function, $F(t)$, is detailed.

First, given an initial confidence level, 1–α, for a small α (α = .01 or .05, typically), we start by constructing individual (1–α)-confidence intervals, (ℓ_j, u_j), for F, in the cumulative observed HTT at inspections, $y_1, y_2, \ldots, y_k$. The process is the following:

1. Fix a pilot bandwidth, g, and compute $\hat{F}_g(x)$, the smooth cumulative distribution function estimation given in (2).
2. Draw B complete bootstrap resamples of size n from $\hat{F}_g(x)$:$X^{*(j)}_1, X^{*(j)}_2, \ldots, X^{*(j)}_n$, j = 1, 2, ..., B. To do this, we first draw n observations from a discrete random variable, I, that takes the values 1, 2, ..., k, with probabilities w_1, w_2, ..., w_k. Let us denote these observations by $I^{(j)}_1, I^{(j)}_2, \ldots, I^{(j)}_n$. Now, for every i = 1, 2, ..., n, we generate $V^{(j)}_i$ with cdf $\mathbb{K}$ and define the bootstrap observations as $X^{*(j)}_i = t_{I^{(j)}_j} + g \cdot V^{(j)}_i$.
3. Consider the incomplete version of these bootstrap resamples: $F^{*(j)}_n(y_0) \leq F^{*(j)}_n(y_1) \leq \cdots \leq F^{*(j)}_n(y_k)$ for j = 1, 2, ..., B.
4. Compute the estimation given in (2) using the incomplete bootstrap resamples: $\hat{F}^{*(j)}_h(x)$, j = 1, 2, ..., B.
5. Aproximate the sampling distribution of the stochastic process $D_n(x) = \hat{F}_h(x) - F(x)$ by the bootstrap distribution of $D^{*}_n(x) = \hat{F}^{*}_h(x) - \hat{F}_g(x)$.
6. Now, the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the bootstrap distribution are computed: $D^{*\left(\lceil \frac{\alpha}{2} B \rceil\right)}_n(x)$ and $D^{*\left(\left\lceil \left(1-\frac{\alpha}{2}\right) B \right\rceil\right)}_n(x)$ ($\lceil \zeta \rceil$ denotes the integer part of ζ).
7. The final bootstrap confidence interval for $F(x)$ is

$$\left[\hat{F}_h(x) - D^{*\left(\left\lceil \left(1-\frac{\alpha}{2}\right) B \right\rceil\right)}_n(x), \hat{F}_h(x) - D^{*\left(\lceil \frac{\alpha}{2} B \rceil\right)}_n(x)\right].$$

Steps 4–7 are repeated for every $x = y_j$, j = 1, ..., k, and, therefore, k pointwise intervals (ℓ_j, u_j), for each value $F(y_j)$, j = 1, ..., k, are calculated. Individual confidence intervals have approximately the nominal coverage probability, 1–α, when they are considered separately (for a particular grid point). However, the probability that the whole curve is included in the band depicted by the set of intervals is much smaller. This is known as the multiple range testing problem or the false discovery rate in high dimensional statistical problems.

A typical way to correct for multiple testing is the popular Bonferroni approach. In a hypothesis testing context, the idea behind this approach is to consider a new significance level, $\alpha_{\text{Bonf}} = \alpha/k$, and compute individual tests using this new level. The resulting multiple test has a multiple level which is much closer to the desired α. However, it is well known that the Bonferroni approach is a conservative procedure. In our context, this means that the joint coverage probability of the confidence "band," computed with the Bonferroni approach, would be larger than the desired 1–α.

Starting from the conservative Bonferroni approach and the anticonservative individual testing approach, the following algorithm finds an approximate (1–α) confidence interval, with a given approximation error δ (δ is typically small in comparison with the nominal α, for instance $\delta = \frac{\alpha}{10}$):

1. Fix $\alpha^{(0)}_{\text{low}} = \alpha_{\text{Bonf}} = \frac{\alpha}{k}$ and $\alpha^{(0)}_{\text{high}} = \alpha$. Fix the iteration number, i = 0.
2. Compute $\alpha^{(i)}_{\text{mean}} = \frac{\alpha^{(i)}_{\text{low}} + \alpha^{(i)}_{\text{high}}}{2}$.
3. Use the bootstrap resamples to compute individual confidence intervals with $1 - \alpha^{(i)}_{\text{low}}$, $1 - \alpha^{(i)}_{\text{mean}}$ and $1 - \alpha^{(i)}_{\text{high}}$ confidence levels.

4. Compute, with the same bootstrap resamples, the proportion of bootstrap curves that are included in each of these confidence bands. These proportions satisfy $p^{(i)}_{\text{low}} \geq p^{(i)}_{\text{mean}} \geq p^{(i)}_{\text{high}}$ and $p^{(i)}_{\text{low}} \geq 1-\alpha \geq p^{(i)}_{\text{high}}$.
5. If $p^{(i)}_{\text{mean}} \geq 1-\alpha$, then define $\alpha^{(i+1)}_{\text{low}} = \alpha^{(i)}_{\text{mean}}$ and $\alpha^{(i+1)}_{\text{high}} = \alpha^{(i)}_{\text{high}}$. Otherwise define $\alpha^{(i+1)}_{\text{low}} = \alpha^{(i)}_{\text{low}}$ and $\alpha^{(i+1)}_{\text{high}} = \alpha^{(i)}_{\text{mean}}$.
6. Stop at step *i* if $\left| p^{(i)}_{\text{mean}} - (1-\alpha) \right| < \delta$. Otherwise increase *i* in one unit and repeat Steps 2–5.

The final approximate (1–α) simultaneous confidence intervals are those obtained for level $1-\alpha^{(i)}_{\text{mean}}$ in the last iteration.

APPENDIX 3

In this appendix, the statistical procedure implemented in the function `anv.binned`, presented in Section 3.4, is described in detail.

Consider an interval-grouped sample of size *n*. This means that given a set of *k* intervals $[y_{j-1}, y_j)$, *j* = 1, ..., *k*, only the number of observations (n_1, ..., n_k) within each interval (instead of the value of every single observation) is known. Note that $n = n_1 + ... + n_k$. For example, as described in Section 2.1, in the weed emergence problem studied in this paper, the vector (y_0, y_1, ..., y_k) denotes the CHTT at emergence and the vector (n_1, ..., n_k) represents the number of seedlings that have emerged in each interval. Additionally, consider that there exists a factor of interest, with *G* levels or treatments, and we want to test whether this factor has a significant effect in the emergence. Let us denote by N^i_j the number of observations in the interval *j* considering the treatment *i*, with *j* = 1, ..., *k* and *i* = 1, ..., *G*. Then, $\sum_{j=1}^{k} N^i_j = N_i$, where N_i denotes the number of observations associated with treatment *i*, *i* = 1, ..., *G* ($n = N_1 + \cdots + N_G$), and $\sum_{i=1}^{G} N^i_j = n_j$, *j* = 1, ..., *k*.

Denoting by F_i the emergence curve considering treatment *i*, *i* = 1, ..., *G*, the problem of testing whether the factor has a significant effect in the emergence can be formulated as:

$$\begin{cases} H_0 : F_i = F_j, \forall i,j \in \{1, \ldots, G\} \\ H_1 : \exists i,j \in \{1, \ldots, G\} \mid F_i \neq F_j \end{cases}$$

A reasonable statistic to address this hypothesis testing is, for example, the following one, of Cramér-von Mises type, based on the statistic to compare *k* independent samples, proposed by Kiefer (1959):

$$D = \sum_{i=1}^{G} N_i \int \left(\hat{F}^g_i(t) - \hat{F}^g(t) \right)^2 d\hat{F}^g(t),$$

where $\hat{F}^g_i(\cdot)$ denotes the corresponding nonparametric estimator of $F_i(\cdot)$, using (2), computed using N_i and $(N^i_1, \ldots, N^i_k)$, for every *i* = 1, ..., *G*, and where $\hat{F}^g(\cdot)$ represents the nonparametric estimator (2) using the pooled sample *n* and (n_1, ..., n_k).

Intuitively, the null hypothesis H_0 will be accepted for small values of *D* and rejected for large values of *D*. To calibrate the test, a bootstrap procedure is employed to approximate the sampling distribution of *D*. The specific steps are the following:

1. Using a bandwidth *h* (e.g., the one provided by `bw.dist.binned.boot`), consider the grouped-data nonparametric estimator, $\hat{F}^g_h$, using the pooled data set with *n* and (n_1, ..., n_k).
2. For each treatment *i*, *i* = 1, ..., *G*, draw $(N^{i*}_1, \ldots, N^{i*}_k)$ from a multinomial distribution $M_k(N_i; \hat{p}^h_1, \ldots, \hat{p}^h_k)$, with $\hat{p}^h_j = \hat{F}^g_h(y_j) - \hat{F}^g_h(y_{j-1})$, *j* = 1, ..., *k*, and define $w^{i*}_j = N^{i*}_j / N_i$.
3. Using the weights $w^{i*}_1, \ldots, w^{i*}_k$, compute the grouped-data nonparametric estimator $\hat{F}^{g*}_i$, for each treatment *i* = 1, ..., *G*, and the nonparametric estimator $\hat{F}^{g*}$ using the pooled bootstrap resample *n* and $\left(\sum_{i=1}^{G} N^{i*}_1, \ldots, \sum_{i=1}^{G} N^{i*}_k \right)$. In all the cases, the bootstrap bandwidths obtained with `bw.dist.binned.boot` can be employed.
4. Define the bootstrap version of *D*:

$$D^* = \sum_{i=1}^{G} N_i \int \left(\hat{F}^{g*}_i(t) - \hat{F}^{g*}(t) \right)^2 d\hat{F}^{g*}(t),$$

5. Steps 2–4 are repeated a large number of times, *B* and, then, a sequence $\{D^*_1, \ldots, D^*_B\}$ is obtained. Given a significance level α, the null hypothesis is rejected if

$$D > D^*_{(\lceil (1-\alpha)\cdot B \rceil)},$$

where $\lceil \cdot \rceil$ represents the integer part, and $\{D^*_{(i)}\}_{i=1}^{B}$ is the sample $\{D^*_i\}_{i=1}^{B}$ arranged in increasing order of magnitude. Additionally, the *p*-value of this test can be approximated by:

$$\hat{p} = \frac{1}{B} \sum_{i=1}^{B} \mathbb{I}_{\{D^*_i > D\}},$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function.
It is important to note that with the resampling process described in Step 2, the bootstrap resamples for each treatment are generated under the null hypothesis.