Extended Abstract

An R Package Implementation for Statistical Modeling of Emergence Curves in Weed Science [†]

Daniel Barreiro-Ures ^{*,‡}, Ricardo Cao [‡] and Mario Francisco-Fernández [‡]

Department of Mathematics, Faculty of Computer Science, University of A Coruña, 15008 A Coruña, Spain; ricardo.cao@udc.es (R.C.); mario.franco@udc.es (M.F.-F.)

* Correspondence: daniel.barreiro.ures@udc.es

† Presented at the XoveTIC Congress, A Coruña, Spain, 27–28 September 2018.

‡ These authors contributed equally to this work.

Published: 18 September 2018

Abstract: Over the last few years, the research group MODES has carried out a research line (in collaboration with researchers from the Sustainable Agriculture Institute of the CSIC in Córdoba) on statistical modeling in weed science. One of the aspects dealt with in this line is that of the estimation of the so-called emergence curves from data obtained from field studies. In this context, new indices have been developed for hydrothermal times, new nonparametric methods have been proposed, which have been compared with other existing parametric methods and applied to relevant pests. In this context, the objective pursued was the development of an R package that can be useful for the statistical analysis of weed science data and, in particular, for the estimation of emergence curves. Currently, the package is available in the CRAN and it is intended to become a standard of use among the research community in weed science.

Keywords: binned data; nonparametric; kernel density; kernel distribution; bandwidth selection

1. Scenario

Let $X_1, \dots, X_n$ be a random sample of our random variable of interest, X , with density function f and distribution function F . Let us assume that a set of k intervals, $[y_{j-1}, y_j)$, $j = 1, \dots, k$, whose midpoints will be denoted by $t_j = \frac{1}{2}(y_{j-1} + y_j)$, $j = 1, \dots, k$. Let us also assume that we do not know the values of each X_i but only the interval to which each of them belongs. Therefore, in this scenario we cannot directly observe the sample $X_1, \dots, X_n$ but only the proportion of observations that fall into each of the intervals, that is, our random sample turns to be $(w_1, \dots, w_k)$, where $w_j = F_n(y_j^-) - F_n(y_{j-1}^-)$, $j = 1, \dots, k$, and F_n denotes the empirical distribution function of $X_1, \dots, X_n$. In this context, the classical kernel density and distribution estimators cannot be computed and must be adapted to this interval-grouping scenario. Kernel density and distribution estimators for interval-grouped data were proposed in [1,2], respectively.

2. Bandwidth Selection for Interval-Grouped Data

The bandwidth selectors proposed in [3] were analyzed through several simulation studies and some modifications were proposed. For instance, in the case of kernel density estimation, a new method for the selection of a pilot bandwidth for the bootstrap selector was proposed and shown to outperform the previously used one in most scenarios. In the case of kernel distribution estimation, a bootstrap bandwidth selector was developed along with a method to select a pilot bandwidth similar to the one proposed for the density case. For both kernel density and distribution estimation, the bootstrap bandwidth selectors were shown to outperform the plug-in selectors in most cases.

3. Application to the Estimation of Seedling Emergence Curves

Weed scientists are usually interested in the prediction of seedling emergence using environmental variables such as the cumulative hydrothermal time (CHTT). However, due to several factors such as budget constraints, the value of the CHTT cannot be measured continuously and so we end up facing an interval-grouped scenario. Weed scientists have traditionally modeled the relationship between seedling emergence and CHTT through parametric regression. To overcome the limitations of this approach, we decided to face the task of seedling emergence estimation from a nonparametric distribution estimation viewpoint. Namely, we want to estimate the distribution (or density) of the random variable CHTT. Furthermore, due to the nature of the measuring process and the fact that the seedlings under study are at different soil depths, we also face the problem of selecting the depth at which to measure the CHTT. Since the depth at which the CHTT is measured will affect the shape of the distribution of our random variable, our objective is to find a depth such that it maximizes the flatness of the distribution of the CHTT. For this task, emergence indices were proposed in [4].

4. Implementation

The methods were coded in C++ to minimize the execution time and integrated into an R package, *binnednp* [5], through the Rcpp API (see [6]). The package is composed of four functions focused on different tasks: kernel density estimation (*bw.dens.binned*), plug-in bandwidth selection for kernel distribution estimation (*bw.dist.binned*), bootstrap bandwidth selection for kernel distribution estimation (*bw.dist.binned.boot*) and nonparametric estimation of emergence indices (*emergence.indices*).

Author Contributions: Conceptualization, D.B., R.C. and M.F.; Methodology, D.B., R.C. and M.F.; Software, D.B., R.C. and M.F.; Validation, D.B., R.C. and M.F.; Formal Analysis, D.B., R.C. and M.F.; Investigation, D.B., R.C. and M.F.; Resources, D.B., R.C. and M.F.; Data Curation, D.B., R.C. and M.F.; Writing—Original Draft Preparation, D.B., R.C. and M.F.; Writing—Review & Editing, D.B., R.C. and M.F.; Visualization, D.B., R.C. and M.F.; Supervision, D.B., R.C. and M.F.; Project Administration, D.B., R.C. and M.F.; Funding Acquisition, D.B., R.C. and M.F.

Funding: This research received no external funding.

Acknowledgments: This research has been supported by MINECO grant MTM-2014-52876-R and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Reyes, M.; Francisco-Fernández, M.; Cao, R. Nonparametric kernel density estimation for general grouped data. *J. Nonparametr. Stat.* **2016**, *2*, 235–249.
2. Cao, R.; Francisco-Fernández, M.; Anand, A.; Bastida, F.; González-Andújar, J.L. Modeling *bromus diandrus* seedling emergence using nonparametric estimation. *J. Agric. Biol. Environ. Stat.* **2013**, *18*, 64–86.
3. Reyes, M.A. Statistical Methods for Studying Emergence Curves in Weed Science. Ph.D. Thesis, Universidade da Coruña, A Coruña, Spain, 2015.
4. Cao, R.; Francisco-Fernández, M.; Anand, A.; Bastida, F.; González-Andújar, J.L. Computing statistical indices for hydrothermal times using weed emergence data. *J. Agric. Biol. Environ. Stat.* **2011**, *149*, 701–712.
5. Barreiro-Ures, D.; Fraguera, B.; Doallo, R.; Cao, R.; Francisco-Fernández, M.; Reyes, M. *binnednp*: Nonparametric Estimation for Interval-Grouped Data. CRAN **2018**. Available online: <https://cran.r-project.org/package=binnednp>, Rpackageversion0.1.0 (accessed on 9 September 2018).
6. Eddelbuettel, D.; Francois, R. Rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **2011**, *40*, 1–18.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).