Weighted Poisson and Semiparametric Kernel Models Applied to Parasite Growth

Tristan Senga Kiessé, Dominique Mizere

WEIGHTED POISSON AND SEMIPARAMETRIC KERNEL MODELS APPLIED TO PARASITE GROWTH

TRISTAN SENGA KIESSÉ[1,*] AND DOMINIQUE MIZÈRE[2]

Office National des Forêts and Université Marien Ngouabi

Summary

This work deals with some parametric and semiparametric modeling approaches for count data distributions related to development of spiraling whitefly which is an insect pest collected in Brazzaville, Republic of Congo. In this study, the count data distributions are assumed to be modified Poisson probability mass functions. For the discrete semiparametric associated kernel estimator investigated, its almost sure consistency and asymptotic normality are shown under some asumptions. Some weighted Poisson models (WPD) are applied in comparison with the semiparametric approach for finite samples characterizing the growth of spiraling whitefly. Finally, the discrete semiparametric estimation is simple and effective for estimating any count distribution while WPD are practically more meaningful.

Key words: count data; discrete associated kernel; semiparametric estimation; weighted Poisson distribution.

1. Introduction

The spiraling whitefly (*Aleurodicus dispersus Russell*) is an insect pest which causes damage to plants by sucking the sap, decreasing photosynthesis activity and drying up the leaves. This insect comes originally from Central America and the Caribbean islands, and is now present in the Congo-Brazzaville. Congolese biologists are searching for a suitable method for modeling data related to the growth of this insect. Thus, some experimental populations were raised on plantations of several host plants, among them some fruit trees well-known in the Congo, such as safou (*Dacryodes edulis*), mango (*Mangifera indica*) or citrus (*Citrus paradisi*); see Kiyindou *et al.* (1999), Mizère *et al.* (2008) and Mizère (2007). These plantations consisted of young trees (5 to 6 months old) under varying conditions of temperature and humidity, and the observations were made using a binocular loupe. The development of the insect parasite studied is described by the following count explanatory variables observed in days: the preimarginal development time from egg to adult stage, the total number of days of egg laying and the longevity of the adult insect. These count data deviate from the equidispersion assumption; thus it becomes necessary to use suitable count estimation models for under- or overdispersed data and the standard framework provided by the Poisson model is not sufficient. In order to express the deviation from classical Poisson

*Author to whom correspondence should be addressed.
[1]Département Recherche et Développement, Centre ONF de Nancy, Velaine en Haye, F-54840, France. e-mail: tristan.senga-kiesse@onf.fr
[2]Faculté des Sciences, B.P. 69 Brazzaville, Congo. e-mail: dmizere@yahoo.fr
Acknowledgements. The authors are grateful to the Associate Editor and two anonymous referees whose comments improved this paper.

models, any count data distribution f, on the set $\mathbb{N}$ of non-negative integers, can be formulated as a *weighted Poisson distribution* (WPD) such that

$$f(x) = p(x;\mu) \times \omega(x), \quad x \in \mathbb{N}, \tag{1}$$

where $p(x;\mu) = \mu^x \exp(-\mu)/x!$ is the Poisson probability mass function (p.m.f.) with mean parameter $\mu > 0$ and $\omega(x)$ is the nonnegative normalized Poisson weight function. When the discrete function ω does not represent the real recording mechanism, or is not well-specified, it is better to allow the count data to yield an estimate of this weight function by a nonparametric method. This opens the way for semiparametric modeling which consists of the construction of an estimate $\widehat{p}$ of the standard Poisson p.m.f. p multiplied by a nonparametric kernel estimate of the function $\omega = f/\widehat{p}$. The nonparametric estimate plays the role of a correction factor of the parametric estimate and intrinsically takes into account special features of the counting phenomenon such as overdispersion (or underdispersion) and zero-inflation (or zero-deflation); see Kokonendji *et al.* (2009). For comparison, several WPD are investigated as alternatives to the parametric Poisson model classically applied for count data by specifying different discrete Poisson weight functions ω. Indeed, these weighted versions of the standard Poisson distribution allow us to take into account the counting phenomena mentioned previously. More precisely, some truncated and translated Poisson distributions, are investigated. Finally, the semiparametric estimation procedure and WPD are applied to count datasets related to the growth of spiraling whitefly in plantations of citrus trees. The advantages provided by each method are investigated with respect to the goodness-of-fit, the new information on insect growth and the meaningfulness of the results in these applications.

The rest of the paper is organized as follows. Section 2 presents the discrete semiparametric kernel estimator using the Poisson p.m.f. as the start function, then WPD are also presented. Basic properties of discrete kernel estimator studied are shown; in particular, mathematical results on the strong consistency and asymptotic normality of the estimator are formulated. Section 3 contains the results of applications of the parametric and semiparametric methods. Concluding remarks are given in Section 4.

2. Semiparametric estimation models and weighted Poisson distributions

Let us recall some notions about discrete semiparametric kernel estimation and WPD.

2.1. Semiparametric kernel estimation

For the semiparametric procedure, the discrete Poisson weight function $\omega(\cdot)$ in (1) is not specified; thus, a discrete nonparametric kernel estimator of $\omega(\cdot)$ is used in addition to a parametric estimate of $p(\cdot;\mu)$.

2.1.1. Estimator

Let $X_1, X_2, \ldots, X_n$ be a sample of independent observations with an unknown count distribution f as in (1). A discrete semiparametric estimator of f is proposed by Kokonendji *et al.* (2009) as the combination of a parametric estimator $\widehat{p}(x) = p(x;\widehat{\mu})$ of p followed by a nonparametric kernel estimator $\widehat{\omega}_n(x)$ of $\omega(x) = f(x)/\widehat{p}(x)$, such that we have

$$\widehat{f}_n(x) = \widehat{p}(x) \times \widehat{\omega}_n(x) = \frac{1}{n} p(x;\widehat{\mu}) \sum_{i=1}^{n} \frac{K_{x,h}(X_i)}{p(X_i;\widehat{\mu})}, \quad x \in \mathbb{N}. \tag{2}$$

The estimator $\widehat{\mu} = n^{-1}\sum_{i=1}^{n} X_i$ is the sample mean, the bandwidth $h = h(n) > 0$ is an arbitrary sequence of smoothing parameters that fulfills $\lim_{n\to\infty} h(n) = 0$, and the discrete associated kernel $K_{x,h}(\cdot)$ of the random variable $\mathcal{K}_{x,h}$ is a p.m.f. with support $\mathcal{S}_x$ (included in $\mathbb{N}$) satisfying the following hypotheses:

H1 $\lim_{h\to 0} \mathrm{E}(\mathcal{K}_{x,h}) = x$ and

H2 $\lim_{h\to 0} \mathrm{Var}(\mathcal{K}_{x,h}) = 0$.

The two previous quite general assumptions can be replaced by

H1′ $\Pr\ (\mathcal{K}_{x,h} = x) = 1 - h\mathrm{A}(\mathcal{K}_{x,h}) + O(h^2)$ and

H2′ $\mathrm{Var}\ (\mathcal{K}_{x,h}) = h\mathrm{V}(\mathcal{K}_{x,h}) + O(h^2)$,

with $\sum_{y\in\mathcal{S}_x\setminus\{x\}} \Pr(\mathcal{K}_{x,h} = y) = h\mathrm{A}(\mathcal{K}_{x,h}) + O(h^2) \to 0$ as $h \to 0$. Indeed, one can verify that the hypotheses H1′–H2′ lead to H1–H2 and are also less general. Note that the expressions for A and V are related to the chosen discrete kernel $\mathrm{K}_{x,h}$ but do not depend on x and h as we will see in the following example.

Example of a discrete kernel. For $(x,a) \in \mathbb{N} \times \mathbb{N}$ and $h > 0$, Kokonendji *et al.* (2007) present the symmetric discrete triangular kernel which is associated with random variable (r.v.) $\mathcal{K}_{a;x,h}$ on support $\mathcal{S}_{a,x} = \{x, x \pm 1, ..., x \pm a\}$ and whose p.m.f. is given by

$$\Pr(\mathcal{K}_{a;x,h} = z) = \frac{(a+1)^h - |z-x|^h}{P(a,h)}, \quad \forall z \in \mathcal{S}_{a;x},$$

with $P(a,h) = (2a+1)(a+1)^h - 2\sum_{k=1}^{a} k^h$ the normalizing constant. This associated kernel satisfies the assumptions H1′ and H2′, and consequently H1 and H2, since we have

$$\Pr(\mathcal{K}_{a;x,h} = x) = 1 - 2h\mathrm{A}(a) + O(h^2) \text{ and } \mathrm{Var}(\mathcal{K}_{a;x,h}) = 2h\mathrm{V}(a) + O(h^2),$$

with $\mathrm{A}(a) = a\log(a+1) - \sum_{k=1}^{a}\log(k)$ and $\mathrm{V}(a) = (a(2a^2+3a+1)/6)\log(a+1) - \sum_{k=1}^{a} k^2\log(k)$. One can see that A and V depend only on the integer parameter a but not on x and h as stated previously. An R package for symmetric and asymmetric discrete triangular distributions is provided by Senga Kiessé *et al.* (2010).

Remark 1. Other examples of discrete associated kernels satisfying H1′–H2′ are the Dirac and Aitchison-Aitken kernels given as examples by Kokonendji & Senga Kiessé (2011). For the Dirac kernel, which is a particular case of an associated kernel without smoothing parameter, *i.e.* $h = 0$, the modal probability at x is equal to 1 and thus $\mathrm{A} = 0$. For the Aitchison-Aitken kernel, the modal probability at x is equal to $1 - h$ and thus $\mathrm{A} = 1$.

Now we propose a data-driven bandwidth selection procedure for the estimator $\widehat{f}_n$.

Bandwidth choice. The bandwidth is generally chosen to minimize the mean integrated squared error (MISE) of $\widehat{f}_n$ such that an ideal parameter value is $h_{id} = \arg\min_{h>0} \mathrm{MISE}(n,h,K,f)$ with, successively,

$$\begin{aligned}\mathrm{MISE}(\widehat{f}_n(x)) &= \sum_{x\in\mathbb{N}} \mathrm{E}(\widehat{f}_n(x) - f(x))^2 \\ &= \mathrm{E}\Big(\sum_{x\in\mathbb{N}} \widehat{f}_n^2(x)\Big) - 2\mathrm{E}\Big(\sum_{x\in\mathbb{N}} \widehat{f}_n(x) f(x)\Big) + \sum_{x\in\mathbb{N}} f^2(x) \\ &= \mathrm{MISE}_{cv}(h) + \sum_{x\in\mathbb{N}} f^2(x).\end{aligned}$$

Thus, for a given discrete kernel $\mathrm{K}_{x,h}$ with $x \in \mathbb{N}$ and $h > 0$, an optimal bandwidth parameter $h_{cv} = \arg\min_{h>0} \mathrm{CV}(h)$ is obtained by minimizing the cross-validation estimator

$$\begin{aligned}\mathrm{CV}(h) &= \sum_{x\in\mathbb{N}} \widehat{f}_n^2(x) - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{n,-i}(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{p(X_i;\widehat{\mu})p(X_j;\widehat{\mu})} \sum_{x\in\mathbb{N}} p^2(x;\widehat{\mu}) K_{x,h}(X_i) K_{x,h}(X_j) \\ &\quad - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} K_{X_i,h}(X_j) \frac{p(X_i;\widehat{\mu}_{-i})}{p(X_j;\widehat{\mu}_{-i})},\end{aligned}$$

where $\widehat{f}_{n,-i}(x) = (n-1)^{-1}\sum_{j\neq i}^{n} K_{x,h}(X_j)$ is the leave-one-out kernel estimator of $\widehat{f}_n(x)$ and $\widehat{\mu}_{-i}$ is computed as $\widehat{\mu}$ by excluding X_i. This estimator is asymptotically unbiased for $\mathrm{MISE}_{cv}(h)$.

2.1.2. Asymptotic properties

First, the basic properties of the estimator $\widehat{f}_n$, such as its bias and variance have been established already by Kokonendji *et al.* (2009); here, we take into account the novel assumptions H1′–H2′ which ensure that we have

$$\begin{aligned}\mathrm{Bias}(\widehat{f}_n(x)) &= \mathrm{E}(\widehat{f}_n(x)) - f(x) \\ &= \frac{h}{2}\mathrm{V}(\mathcal{K}_{x,h}) p_0(x)\omega^{(2)}(x) + o(h^2) + O(h^2), \quad x \in \mathbb{N}, \\ \mathrm{Var}(\widehat{f}_n(x)) &= \frac{1}{n}\sum_{y\in\mathcal{S}_x} f(y)\big(\Pr(\mathcal{K}_{x,h}=y)\big)^2 - \frac{1}{n}\left(\sum_{y\in\mathcal{S}_x} f(y)\Pr(\mathcal{K}_{x,h}=y)\right)^2 \\ &= \frac{1}{n} f(x)(1 - h\mathrm{A}(\mathcal{K}_{x,h}))^2 - \frac{1}{n} f^2(x) + o\left(\frac{1}{n}\right) + O(h^2),\end{aligned}$$

where $p_0 = p(x;\mu_0)$ is the Poisson p.m.f with mean μ_0, $\widehat{\mu}$ converges to μ_0 and $\omega^{(2)}$ is the finite difference of second order of ω. It ensues $\mathrm{Bias}(\widehat{f}_n(x)) \to 0$ and $\mathrm{Var}(\widehat{f}_n(x)) \to 0$ when $h = h(n) \to 0$ and $n \to \infty$. Therefore, the pointwise and global consistencies of $\widehat{f}_n$

can be deduced easily by showing, respectively, that the mean squared error MSE and the integrated MISE both tend to 0 as $h \to 0$ and $n \to \infty$ since we have:

$$\mathrm{MISE}(\widehat{f}_n(x)) = \sum_{x\in\mathbb{N}} \mathrm{MSE}(x) = \sum_{x\in\mathbb{N}} \mathrm{Bias}^2(\widehat{f}_n(x)) + \sum_{x\in\mathbb{N}} \mathrm{Var}(\widehat{f}_n(x)).$$

Next a mathematical result on the almost sure consistency of the estimator $\widehat{f}_n$ is formulated, followed by another result concerning its asymptotic normality. The proofs of the two results are postponed to the Appendix.

Theorem 1. *For any fixed $x \in \mathbb{N}$, the semiparametric estimator $\widehat{f}_n(x)$ converges almost surely to p.m.f. $f(x)$ as follows:*

$$\widehat{f}_n(x) \to f(x) \text{ as } n \to \infty.$$

Theorem 2. *For any fixed $x \in \mathbb{N}$, the semiparametric estimator $\widehat{f}_n(x)$ converges in distribution to the normal law as follows:*

$$\frac{\widehat{f}_n(x) - \mathrm{E}(\widehat{f}_n(x))}{\left(\mathrm{Var}(\widehat{f}_n(x))\right)^{1/2}} \to \mathrm{N}(0,1) \text{ as } n \to \infty.$$

In the following section, we are interested in WPD when the discrete Poisson weight function $\omega(\cdot)$ in (1) is well-specified. Thus, the modeling approach developed is completely parametric.

2.2. Weighted Poisson distributions

Let X be a r.v. having a Poisson p.m.f. $p(x;\mu)$ with mean parameter $\mu > 0$. The r.v. X^ϕ said to be the *weighted version of X* has a p.m.f. given by

$$\Pr(X^\phi = x) = \frac{\phi(x)p(x;\mu)}{\mathrm{E}(\phi_\mu(X))} = p_\phi(x;\mu), \quad x \in \mathbb{N}, \tag{3}$$

where $\phi(x)$ is a nonnegative weight function on $\mathbb{N}$ and the denominator is the normalizing constant depending on μ such that $0 < \mathrm{E}(\phi_\mu(X)) < \infty$. The discrete weight function $\phi(x) = \phi(x,\lambda)$ can depend both on the parameters λ and μ, where λ represents the recording mechanism. Clearly, the standard Poisson distribution is a WPD with unit weight function $\omega(x) = 1, \forall x \in \mathbb{N}$. In addition, the weighted variable X^ϕ is said to be overdispersed (underdispersed) when Fisher dispersion indicator $I(X^\phi) = \mathrm{Var}(X^\phi)/\mathrm{E}(X^\phi)$ is greater (smaller) than 1, while the Poisson variable is equidispersed when $I(X) = 1$. Let us finally remark that by comparing the equation (3) to equation (1) we have $\omega(\cdot) = \phi(\cdot)/\mathrm{E}(\phi_\mu(X))$. In the following we give some examples of WPD.

- We first present the modified Poisson model $WPD_2(\mu,\lambda)$ with discrete probability distribution given by

$$p_\phi(x;\mu) = \frac{1 + x/(\lambda+1)}{1 - \exp(-\mu) + \mu/(\lambda+1)} p(x;\mu), \quad x \in \mathbb{N}\setminus\{0\}.$$

Note that $WPD_2(\mu,\lambda) \to P_0(\mu)$ as $\lambda \to \infty$, where $P_0(\mu)$ corresponds to the Poisson p.m.f. $P(\mu)$ truncated at 0. This WPD is underdispersed.

- The second model considered is $WPD_3(\mu,k,\lambda)$ with

$$p_\phi(x;\mu) = \frac{1 + (x-k)/(\lambda+1)}{1 + \mu/(\lambda+1)} p(x-k;\mu), \quad x \geq k.$$

We have $WPD_3(\mu,k,\lambda) \to PT(\mu,k)$ as $\lambda \to \infty$, where PT is the translated Poisson p.m.f. with parameters μ and k. This WPD is also underdispersed.

- The third model is the zero-modified weighted distribution $ZMW(\mu,\lambda,\pi_0)$ with the following p.m.f.:

$$p_\phi(x;\mu,\lambda,\pi_0) = \begin{cases} \pi_0 & \text{if } x = 0 \\ (1-\pi_0)\frac{1+x/(\lambda+1)}{1-\exp(-\mu)+\mu/(\lambda+1)} p(x;\mu) & \text{if } x \in \mathbb{N}\setminus\{0\}, \end{cases}$$

with $0 < \pi_0 < 1$. This distribution can be under- or overdispersed depending on the parameter π_0.

Let us give some details about the possible interpretation of the parameters and the method for their estimation. The integer parameter λ serves to construct a family of distributions $p_\phi(x;\mu,\lambda)$ which converge to $p_\phi(x;\mu)$ and has no particular biological interpretation; the parameter μ, is the mean of Poisson p.m.f.. These two parameters can be estimated by maximum likelihood. The parameter π_0 is the theoretical zero proportion and can be estimated by the empirical zero proportion. Finally, the parameter k is the absolute minimum time it takes for an insect to become an adult parasite; thus the host plant with the lowest k-value is more favorable to the development of the parasite. This last parameter is estimated using the method of moments. In the applications, our main concern is the estimated value of k because it is useful for controlling reproduction of this specific insect species. For more details on modeling count data phenomena and WPD, see Kokonendji *et al.* (2008) and Mizère (2006).

3. Applications

In this section, some diagnostic checks are used to choose between the parametric and semiparametric models. Then, the results are given for the application of each method (classical Poisson, WPD and semiparametric $\widehat{f}_n$) on count datasets related to the growth of Congolese spiraling whitefly.

Note that, for the discrete triangular kernel semiparametric estimator, the parameter $a \in \mathbb{N}$ is equal to $1, 2$ or 3 in practice. We propose here to fix $a = 1$ since the global error MISE increases with $a \in \mathbb{N}$ (Kokonendji *et al.*, 2007). For example, Figure 1 illustrates the comparative behaviors of function $a \mapsto \mathrm{MISE}(a;n,h,f)$ of $\widehat{f}_n$ with a discrete triangular kernel $K_{a;x,h}$ for the simulated p.m.f. $f(x) = 0.4\mathrm{Pn}(x;0.5) + 0.6\mathrm{Pn}(x;10)$, $x \in \mathbb{N}$, which is a mixture of two Poisson distributions $\mathrm{Pn}(x;\mu)$ with respective means $\mu_1 = 0.5$ and $\mu_2 = 10$. For fixed $h > 0$ and sample sizes n, the optimal value $a_{opt} = \arg\min_{a\in\mathbb{N}} \mathrm{MISE}(a)$ is less

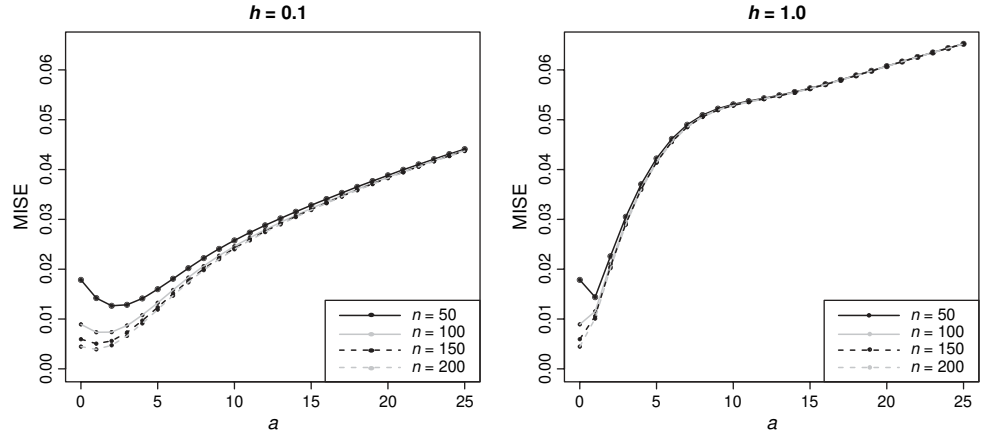

Figure 1. Simulated function $a \mapsto \mathrm{MISE}(a; n, h, f)$ of the semiparametric estimator using the discrete triangular kernel for $f = 0.4\mathrm{Pn}(0.5) + 0.6\mathrm{Pn}(10)$.

than or equal to 3; note that the case $a = 0$ for the discrete triangular kernel results in a naive kernel of Dirac type.

3.1. Model diagnostics for semiparametric estimation

The estimated discrete Poisson weight function $\widehat{\omega}_n$ in (2) provides useful information for model diagnostics. This Poisson weight function should equal one if the Poisson p.m.f. is indeed the true p.m.f. Therefore, the adequacy of the model can be checked by examining a plot of the weight function: we are interested here in plotting the log weight function $\log \widehat{\omega}_n(x) = \log\{\widehat{f}_n(x)/p(x;\widehat{\mu})\}$ to see how far away it is from zero. Thus, a simple graphical goodness-of-fit diagnostic emerges by plotting x against

$$Z(x) = \frac{\log \widehat{\omega}_n(x) + (2n)^{-1}(p(x;\widehat{\mu}))^{-1} \Pr(\mathcal{K}_{x,h} = x)}{\left(n^{-1}\big(p(x;\widehat{\mu})\big)^{-1} \Pr(\mathcal{K}_{x,h} = x)\right)^{1/2}}.$$

When the Poisson p.m.f. is indeed the true p.m.f., $Z(x)$ is approximately distributed as standard normal for each target x, meaning that the $Z(x)$-values should lie within ± 1.96 about 95% of the time; see Hjort & Glad (1995, section 8.2).

Concerning the preimarginal development time and egg laying datasets, the $Z(x)$-values stay within ± 1.96 all the time; see Figures 2 and 3, respectively. This suggests that it would be of interest to consider parametric Poisson (or also WPD) models for these data. For the longevity data, only 44.4% of $Z(x)$-values belong to the confidence band ± 1.96 (Figure 4). This means semiparametric methods should work better than parametric methods (see also Table 3 later).

The following section provides the detailed results about the performances of standard and weighted Poisson models in comparison with the semiparametric kernel estimator.

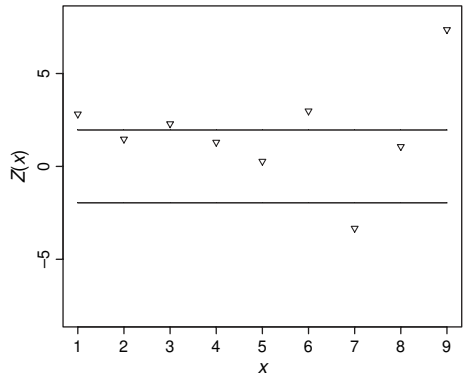

Figure 2. The $Z(x)$-values associated with the results on data of preimarginal development time.

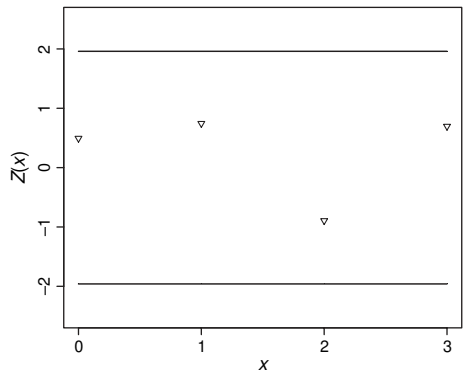

Figure 3. The $Z(x)$-values associated with the results on data of egg laying.

3.2. Parametric and semiparametric results

In this section, the performance of each model applied is evaluated by using the practical integrated squared error

$$\mathrm{ISE} = \sum_{x \in \mathbb{N}} \left(\widehat{f}(x) - f_0(x) \right)^2,$$

which is a descriptive measure-of-fit, where f_0 is the observed frequency and $\widehat{f}$ represents the estimated frequency from the application of WPD or $\widehat{f}_n$. For count data, we can also measure performance through the following chi-squared (χ^2) distance:

$$\chi^2_0 = \sum_{x' \in \{0,1,\ldots,N_0\}} \frac{n \left(\widehat{f}(x') - f_0(x') \right)^2}{\widehat{f}(x')},$$

where $N_0 + 1$ represents the number of valid classes in the χ^2-test. Thus, the statistic χ^2_0 can be suitably approximated by the χ^2-distribution with $N_0 - r$ degrees of freedom (d.f.), where r is the number of estimated parameters (*e.g.*, Greenwood & Nikulin 1996; Saporta

TABLE 1

Estimation of preimarginal development time in days using weighted Poisson distributions and semiparametric triangular kernel estimation.

Days	Observed frequencies	Expected $PT(\mu, 20)$ frequencies	Expected $WPD_3(\mu, 20, 30)$ frequencies	Expected semiparametric $\widehat{f}_n$ with $a = 1$ frequencies
25	10	8.174	8.120	8.044
26	7	6.783	6.775	7.807
27	7	8.993	9.000	7.716
28	10	10.432	10.454	11.127
29	18	10.757	10.787	14.970
30	4	9.983	10.011	6.191
31	3	8.423	8.441	4.148
32	8	6.514	6.521	7.174
33	6	4.650	4.647	6.131
34	4	3.082	3.074	4.628
35	5	4.204	4.165	4.061
$\widehat{\mu}$		9.280	9.054	–
h_{cv}		–	–	0.38
ISE		0.019	0.019	0.035×10^{-1}
df		7	7	6
χ^2_0		13.954	13.995	2.540
p-value		0.052	0.051	0.864
AIC		−28.673	−28.655	−30.523

TABLE 2

Estimation of total number of days of egg laying using weighted Poisson distributions and semiparametric triangular kernel estimation.

Days	Observed frequencies	Expected $P(\mu)$ frequencies	Expected $ZMW(\mu, 30, \widehat{\pi}_0)$ frequencies	Expected semiparametric $\widehat{f}_n$ with $a = 1$ frequencies
0	48	52.862	48	48.345
1	33	23.207	32.075	32.240
2	0	5.094	1.851	0.458
3	1	0.835	0.073	0.956
$\widehat{\mu}$		0.439	0.111	–
h_{cv}		–	–	0.05
ISE		0.021	0.0765×10^{-2}	0.013×10^{-2}
df		1	1	0
χ^2_0		8.677	0	0.006
p-value		0.003	1	0
AIC		−15.120	−8.675	−13.051

1978). In particular, in semiparametric models, one definition of the degrees of freedom for a kernel density estimation fit could be derived from theories of local regression modeling which give

$$\sum_{i=1}^{n} \log(\widehat{f}_{n,-i}(X_i)) \approx \sum_{i=1}^{n} \log(\widehat{f}_n(X_i)) - \mathrm{df} + 1;$$

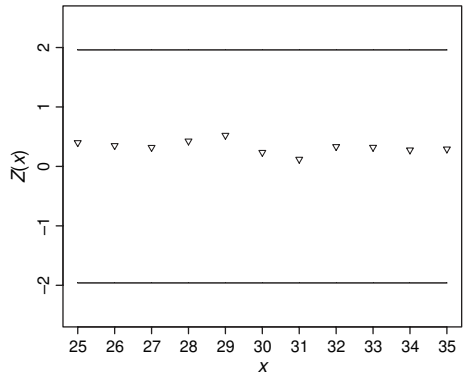

Figure 4. The $Z(x)$-values associated with the results on data of longevity of adult insects.

TABLE 3

Estimation of longevity of adult insects observed in days using weighted Poisson distributions and semiparametric triangular kernel estimation.

Days	Observed frequencies	Expected $P_0(\mu)$ frequencies	Expected $WPD_2(\mu, 30)$ frequencies	Expected semiparametric $\widehat{f}_n$ with $a = 1$ frequencies
1	29	22.678	22.659	27.417
2	16	24.797	24.809	17.553
3	22	18.075	18.092	21.512
4	8	9.882	9.887	8.178
5	2	4.322	4.319	2.452
6	4	1.575	1.571	3.739
7	0	0.421	0.489	0.061
8	0	0.134	0.133	0.160
9	1	0.041	0.410	0.926
$\widehat{\mu}$		2.187	2.123	–
h_{cv}		–	–	0.07
ISE		0.022	0.022	0.082×10^{-2}
df		3	3	2
χ^2_0		6.131	6.137	0.259
p-value		0.105	0.105	0.878
AIC		−8.732	−13.617	−13.712

see Loader (1999, page 92). Using this equation, the values of df are not integers but real numbers. Here, these computed values of df are rounded to integers; they are often called effective number of parameters. Thus, model comparison using the computed df is made through the Akaike information criterion (AIC) and the χ^2 goodness-of-fit test.

Preimarginal duration. Table 1 presents the preimarginal development time data (in days) of the insect studied. The observations have mean 29.280 and variance 8.870 so the Fisher dispersion indicator is $I = 0.303 < 1$ (*i.e.* underdispersion). Concerning the modified Poisson models, $WPD_3(\mu, k, \lambda)$ and $PT(\mu, k)$ are applied since they are left censored and underdispersed. We have the estimated values $\widehat{k} = 20$ and $\widehat{\lambda} = 30$ obtained by the methods of moments and maximum likelihood, respectively. Looking at the discrete semipara-

metric triangular kernel estimator $\widehat{f}_n$ with $a = 1$, the cross-validation procedure provides an optimal h-value $h_{cv} = 0.38$. The descriptive measure-of-fit ISE indicates that the semiparametric model is better than the models $WPD_3(\mu, 20, 30)$ and $PT(\mu, k)$ which both have closed performances. In particular, let us note that the modal frequency at $x = 29$ is equal to 18 for observations while it is equal to 14.970 and around 10.8 for the semiparametric and Poisson models, respectively. Looking at the χ^2 goodness-of-fit test, for the models PT and WPD_3 we have χ^2_0-values equal to 13.954 and 13.995 (df = 7) with the p-values equal to 0.052 and 0.051, respectively; in comparison, for the semiparametric model we have $\chi^2_0 = 2.540$ but a smaller df equal to 6 with p-value $= 0.864$. Finally, looking at the Akaike information criterion, for PT and WPD_3 we have AIC values around -28.7; the value is -30.523 for $\widehat{f}_n$.

Total number of days of egg laying. The observations of these data have mean 0.439 and variance 0.323 so the Fisher dispersion indicator is $I = 0.736 < 1$; $ZMW(\mu, \lambda, \pi_0)$ and standard $P(\mu)$ are applied on these data (see Table 2). Concerning $ZMW(\mu, 30, \widehat{\pi}_0)$, the zero proportion observed $\widehat{\pi}_0 = 0.585$ is smaller than the zero proportion $\exp(-0.439) = 0.644$ expected under the Poisson model $P(\mu)$. For the discrete semiparametric triangular kernel estimator $\widehat{f}_n$ with $a = 1$, we have $h_{cv} = 0.05$. Finally, the quality of fit of $\widehat{f}_n$ is better (in terms of ISE) than that of $ZMW(\mu, 30, \widehat{\pi}_0)$ which is itself better than $P(\mu)$. However, $\widehat{f}_n$ and $ZMW(\mu, 30, \widehat{\pi}_0)$ have some ISE-values of the same order. Looking at the χ^2-test, for the models $ZMW(\mu, \lambda, \pi_0)$ and $P(\mu)$ we have χ^2_0 values equal to 0 and 8.677 with p-values equal to 1 and 0.003, respectively, for the same df equal to 1; in comparison, for the semiparametric estimator $\widehat{f}_n$, we have $\chi^2_0 = 0$ and p-value $= 0$ with a null value of df. In particular, let us consider the AIC since zero degrees of freedom looks unusual: ZMW, P and $\widehat{f}_n$ have AIC values of -8.675, -15.120 and -13.051, respectively. AIC and χ^2_0 values do not result in the same conclusion concerning the performance of the different models applied, but, a parametric model (ZMW or P) is better than the semiparametric model, depending on the criterion used.

Longevity. Table 3 presents the longevity (in days) of the adult insect studied; the data have mean 2.463 and variance 2.425 with a Fisher dispersion indicator of $I = 0.985 < 1$ which is almost identical to one. Here $WPD_2(\mu, \lambda)$ and truncated $P_0(\mu)$ are used since there is no observed value at day 0. The discrete semiparametric triangular kernel estimator $\widehat{f}_n$ with $a = 1$ and $h_{cv} = 0.07$ outperforms $WPD_2(\mu, 30)$ and $P_0(\mu)$ models in the measure ISE. In particular, using $\widehat{f}_n$ reduces boundary bias since it gives a good adjustment at the left boundary $x = 1$. Looking at the χ^2-test, $\widehat{f}_n$ has a χ^2_0 value equal to 0.259 (with df = 2 and p-value $= 0.878$) while WPD_2 and P_0 have χ^2_0 values around 6.13 (with df = 3 and p-value=0.105). Finally, WPD_2, P_0 and $\widehat{f}_n$ have AIC values of -13.617, -8.732 and -13.712, respectively.

4. Concluding remarks

The discrete semiparametric kernel approach, in addition to being simple and effective for estimating any unknown count distribution, was intended to work well even if the unknown p.m.f. cannot be approximated well by the Poisson distribution. This semiparametric modeling intrinsically took into account the special features of count data via

the discrete nonparametric weight function ω, and, it provided some interesting measure-of-fit in diagnostics. However, WPD opened the way for more practical interpretation and discussion of counting phenomena observed in the data. Thus, concerning the examples treated in this work, the biologist had to focus on stopping the reproduction of *Aleurodicus* to fight against its spread because the minimum development time of this insect pest was $\widehat{k} = 20$ days for citrus trees. Finally, WPD were more meaningful in these applications than semiparametric modeling since they provided new information on insect growth.

Appendix A: Proofs

The proof of Theorem 1 requires the use of the following lemma (see Hoeffding, 1963).

Lemma. *Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. random variables with finite second moments. If there exist constants a and b such that $\Pr(Z_i \in [a,b]) = 1$, then given $\epsilon > 0$ we have*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{\epsilon(b-a) + 2\mathrm{Var}(Z_1)}\right).$$

Proof of Theorem 1. The demonstration is based on the following decomposition:

$$\widehat{f}_n(x) - f(x) = \widehat{f}_n(x) - \mathrm{E}(\widehat{f}_n(x)) + \mathrm{E}(\widehat{f}_n(x)) - f(x).$$

First, the term $\mathrm{E}(\widehat{f}_n(x)) - f(x)$ tends to 0 since $\mathrm{MSE}(x) = \mathrm{E}(\widehat{f}_n(x) - f(x))^2 \to 0$ as $n \to \infty$. Second, let us write $\widehat{f}_n(x) - \mathrm{E}(\widehat{f}_n(x)) = (1/n)\sum_{i=1}^{n} Z_i$ with $Z_i = p_0(x)/p_0(X_i)$ $K_{x,h}(X_i) - \mathrm{E}\{p_0(x)/p_0(X_i)K_{x,h}(X_i)\}$. For any $x \in \mathbb{N}$, there exists $0 < M_1 < \infty$ and $0 < M_2 < \infty$ such that we have

$$|Z_i| \leq M_1 \text{ and } \mathrm{Var}(Z_i) \leq \mathrm{E}(p_0(x)/p_0(X_i)K_{x,h}(X_i))^2 < M_2,$$

since $p_0(\cdot)$ and $K_{x,h}(\cdot)$ are p.m.f. Therefore, according to the Hoeffding lemma, one has

$$\Pr\left(\left|\widehat{f}_n(x) - \mathrm{E}(\widehat{f}_n(x))\right| \geq \epsilon\right) = \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-n\epsilon^2}{2\epsilon M_1 + 2M_2}\right),$$

for any $\epsilon > 0$. Consequently, the Borel-Cantelli lemma leads to get $\widehat{f}_n(x) - \mathrm{E}(\widehat{f}_n(x)) \overset{a.s.}{\to} 0$ since $\sum_{n\geq 1} \Pr(|\widehat{f}_n(x) - \mathrm{E}(\widehat{f}_n(x))| \geq \epsilon) < \infty$.

Proof of Theorem 2. In order to get the desired convergence, a sufficient condition is

$$\lim_{n\to\infty} \frac{\mathrm{E}|p_0^{-1}(X_1)K_{x,h}(X_1) - \mathrm{E}(p_0^{-1}(X_1)K_{x,h}(X_1))|^3}{n^{1/2}(\mathrm{Var}(p_0^{-1}(X_1)K_{x,h}(X_1)))^{3/2}} = 0,$$

since $\widehat{f}_n$ is a sum of i.i.d. random variables (see Breiman 1968, theorem 9.2). For this, note that there exists $0 < M < \infty$ such that

$$\mathrm{E}|p_0^{-1}(X_1)K_{x,h}(X_1) - \mathrm{E}(p_0^{-1}(X_1)K_{x,h}(X_1))|^3 < 2M^3$$

and

$$\lim_{n\to\infty} \mathrm{Var}(p_0(x)p_0^{-1}(X_1)K_{x,h}(X_1)) = f(x) - f^2(x).$$

Then, we have the following expressions successively,

$$\begin{aligned}
&\mathrm{Var}(p_0(x)p_0^{-1}(X_1)K_{x,h}(X_1)) - (f(x) - f^2(x)) \\
&= \sum_{y\in\mathbb{N}\cap\mathcal{S}_x} (p_0(x)p_0^{-1}(y)K_{x,h}(y))^2 f(y) - f(x) \\
&\quad - \Big(\sum_{y\in\mathbb{N}\cap\mathcal{S}_x} p_0(x)p_0^{-1}(y)K_{x,h}(y)f(y)\Big)^2 + f^2(x) \\
&= \sum_{y\in\mathbb{N}\cap\mathcal{S}_x} ((p_0(x)p_0^{-1}(y))^2 K_{x,h}(y)f(y) - f(x))K_{x,h}(y) - f(x)\sum_{y\in\overline{\mathbb{N}}\cap\mathcal{S}_x} K_{x,h}(y) \\
&\quad + \Big(f^2(x) - \mathrm{E}^2(\widehat{f}_n(x))\Big) \le (1 - K_{x,h}(x))K_{x,h}(x)f(x) \\
&\quad + \sum_{y\in\mathbb{N}\cap\mathcal{S}_x\setminus\{x\}} |(p_0(x)p_0^{-1}(y))^2 K_{x,h}(y)f(y) - f(x)|K_{x,h}(y) + 2|\mathrm{E}(\widehat{f}_n(x)) - f(x)| \\
&\quad \le (1 - K_{x,h}(x)) + 2|\mathrm{E}(\widehat{f}_n(x)) - f(x)| + O(h) \to 0 \text{ as } n \to \infty. \qquad \square
\end{aligned}$$

References

BREIMAN, L. (1968). *Probability*. Reading, Mass.-London-Don Mills, Ont: Addison-Wesley Publishing Company.

GREENWOOD, P.E. & NIKULIN, M.S. (1996). *A Guide to Chi-Squared Testing*. New York: John Wiley and Sons.

HJORT N.L. & GLAD, I.K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.* **23**, 882–904.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.

KIYINDOU, A., ADOUMBAYE, I.P., MIZÈRE, D. & MOUSSA, J.B. (1999). Influence de la plante hôte sur le développement et la reproduction de l'aleurode Aleurodicus dispersus Russell (Hom: Aleyrodidae) en République du Congo. *Fruit* **54**, 115–122.

KOKONENDJI, C.C. & SENGA KIESSÉ, T. (2011). Discrete associated kernel method and extensions. *Stat. Methodol.* **8**, 497–516.

KOKONENDJI, C.C., MIZÈRE, D. & BALAKRISHNAN, N. (2008). Connection of the Poisson weight function to overdispersion and underdispersion. *J. Statist. Plann. Inference* **138**, 1287–1296.

KOKONENDJI, C.C., SENGA KIESSÉ, T. & BALAKRISHNAN, N. (2009). Semiparametric estimation for count data through weighted distributions. *J. Statist. Plann. Inference* **139**, 3625–3638.

KOKONENDJI, C.C., SENGA KIESSÉ, T. & ZOCCHI, S.S. (2007). Discrete triangular distributions and non-parametric estimation for probability mass function. *J. Nonparam. Statist.* **19**, 241-254.

LOADER, C. (1999). *Local regression and likelihood*. New York, Springer.

MIZÈRE, D. (2006). Contribution à la modélisation et à l'analyse statistique des données de dénombrement (Ph.D. thesis). University of Pau, France.

MIZÈRE, D. (2007). Modélisation statistique des données Aleurodicus sur agrume par des lois de Poisson pondérées. *Afr. Statist.* **2**, 59–72.

MIZÈRE, D., KISSITA, G. & KIYINDOU, A. (2008). Etude statistique de linfluence de la plante hôte sur le développement de laleurode Aleurodicus dispersus Russel via le modèle de Poisson translaté. *Annales de l'Université Marien NGOUABI Sciences et Techniques* **9**, 19–29.

SAPORTA, G. (1978). *Théorie et méthodes de la statistique*. Paris: Edition Technip.

SENGA KIESSÉ, T., LIBENGUÉ, F.G., ZOCCHI, S.S. & KOKONENDJI, C.C. (2010). The R package for general discrete triangular distributions. Available at http://cran.r-project.org/web/packages/TRIANGG/index.html.