Journal of Insect Science OPEN ACCESS

RESEARCH

An Adaptive Kernel Smoothing Method for Classifying *Austrosimulium tillyardianum* (Diptera: Simuliidae) Larval Instars

Guanjun Cen,[1] Yonghao Yu,[2,3] Xianru Zeng,[2] Xiuzhen Long,[2] Dewei Wei,[2] Xuyuan Gao,[2] and Tao Zeng[2]

[1]Department of Applied Mathematics, College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China
[2]Guangxi Key Laboratory of Biology for Crop Diseases and Insect Pests/Institute of Plant Protection, Guangxi Academy of Agricultural Sciences, Nanning 530007, China
[3]Corresponding author, e-mail: yonghaoyucn@yeah.net

Subject Editor: Julie Urban

J. Insect Sci. *(2015)* 15(1): 159; DOI: 10.1093/jisesa/iev136

ABSTRACT. In insects, the frequency distribution of the measurements of sclerotized body parts is generally used to classify larval instars and is characterized by a multimodal overlap between instar stages. Nonparametric methods with fixed bandwidths, such as histograms, have significant limitations when used to fit this type of distribution, making it difficult to identify divisions between instars. Fixed bandwidths have also been chosen somewhat subjectively in the past, which is another problem. In this study, we describe an adaptive kernel smoothing method to differentiate instars based on discontinuities in the growth rates of sclerotized insect body parts. From Brooks' rule, we derived a new standard for assessing the quality of instar classification and a bandwidth selector that more accurately reflects the distributed character of specific variables. We used this method to classify the larvae of *Austrosimulium tillyardianum* (Diptera: Simuliidae) based on five different measurements. Based on head capsule width and head capsule length, the larvae were separated into nine instars. Based on head capsule postoccipital width and mandible length, the larvae were separated into 8 instars and 10 instars, respectively. No reasonable solution was found for antennal segment 3 length. Separation of the larvae into nine instars using head capsule width or head capsule length was most robust and agreed with Crosby's growth rule. By strengthening the distributed character of the separation variable through the use of variable bandwidths, the adaptive kernel smoothing method could identify divisions between instars more effectively and accurately than previous methods.

Key Words: *Austrosimulium tillyardianum*, instar determination method, adaptive kernel smoothing estimation, bandwidth selection

Determining larval instar stages is important for pest control and entomological studies. For example, to be effective, control applications must be timed to coincide with the period when instars are most vulnerable (Schmidt et al. 1977). In addition, the characterization of instar distributions is often required for life table analyses, key factor analyses, and other entomological investigations (Logan et al. 1998).

Larval instars are commonly separated based on the frequency distribution of insect body dimensions (Chen and Seybold 2013). The accuracy and reliability of frequency distribution methods depends on estimations of measurement density. In earlier studies, researchers generally used histograms to estimate density measurements. The number of instars was determined by the number of peaks in the histogram, and class limits for each instar were determined by visual inspection, followed by the application of Brooks' (Dyar's) rule and Crosby's growth rule to verify the results (Crosby 1973, Loerch and Cameron 1983). More commonly, however, there is an overlap in the measurement distribution between instar stages (McClellan and Logan 1994). In such situations, it is difficult to estimate the number of instar stages because the peaks of successive instars blend together in the histogram, and the boundaries between instars are indistinct. Indeed, some studies have demonstrated that the histogram method can yield misleading results (Kishi 1971, McNeil 1978). To analyze overlapping distribution data, various approaches that apply parametric models, such as the normal model, have been developed to estimate the density function of the measurements. Got (1988) fit a distribution model of head capsule widths (HCWs) of the European corn borer (Lepidoptera: Pyralidae) that composed of five normal distributions and proposed a discrimination method to determine classification boundaries based on these distributions. Beaver and Sanderson (1989) used this model to classify instars of the navel orange worm (Lepidoptera: Pyralidae), although these authors used an estimation-maximization algorithm to estimate the model parameters. McClellan and Logan (1994) assumed that the head capsule distribution for each instar was normally distributed and presented a method using a nonlinear least-squares parameter estimation to describe the distribution of gypsy moth instars. A generalized computer program for implementing this method was written by Logan et al. (1998) and was subsequently used in a number of studies (Panzavolta 2007, Dallara et al. 2012). Hammack et al. (2003) applied a multiple Gaussian model similar to the Logan model to determine instars of the corn rootworm. When using instar determination methods based on the normal models described above, the number of components in the models, which is determined by a visual inspection of the histogram, must be known beforehand (e.g., the number of instars). However, the shape of a histogram changes depending on the bin width, and therefore, the number and shape of the peaks in a histogram are subject to the influence of bin width and other factors (Schmidt et al. 1977, Fink 1984, Chen and Seybold 2013). In previous studies, the bin width for each histogram had to be manually specified by experienced researchers. Considering these limitations, it is necessary to improve methods to classify instar stages based on frequency distributions. First, an effective approach to selecting bandwidth is highly desired. Logan et al. (1998) noted that the selection of an appropriate frequency class width (bin width) is important because of the smoothing effect on the instar determination results. They also described an iterative method for the selection of a bandwidth, in which the bandwidth was calculated according to the number of frequency classes (bars in the histogram), which still required specification based on experience. Second, to better analyze the overlapping data, a density estimator that applies more smoothing than a histogram is necessary. Chen and Seybold (2013) used kernel density estimation with a fixed bandwidth to determine the number of peaks in the frequency distribution, but nonparametric methods that utilize fixed bandwidths exhibit significant limitations when these methods are used to fit measurement distributions with multiple peaks. Third, it is necessary to discuss in depth the method used to classify instars based on frequency distributions, and such discussions are infrequent in the literature.

The objectives of this study were to provide an in-depth discussion of the method used to determine instars from frequency distributions, to identify an approach to selecting the optimum bin width when applying nonparametric methods to estimate measurement distributions, and to develop an adaptive kernel smoothing technique for instar determination.

Materials and Methods

Data Source. Measurements from *Austrosimulium tillyardianum* larvae were obtained from Appendix 4 of Crosby's PhD thesis (1974). On the basis of these data, Crosby (1974) deduced the growth ratio rule that has since been widely used in related studies. The data included 16 types of measurements and morphological features from 343 *A. tillyardianum* larvae that were divided into two sets based on the year of data collection; the standardization set was used to build the models, and the test set was used to verify the models. There were 208 data entries in the standardization set and 135 in the test set. In this study, we used five types of measurements from this dataset: head capsule length (HCL), HCW, head capsule postoccipital width (HCPW), mandible length (ML), and antennal segment 3 length (AS3L).

Modeling

Discontinuity Intervals. All insects possess exoskeletons. Although the bodies of larvae regularly grow and continually increase in length, the sizes of certain sclerotized body parts, such as the head capsule, mouthparts, antennae, and mandible, display discontinuous growth rates because the growth of these sclerotized parts only occurs when an insect molts and a new, soft cuticle is produced and expanded (Chapman 1998). At each molt, the size of each sclerotized body part increases so significantly that gaps in measurements between successive instars can be observed. In this article, we refer to such gaps as discontinuity intervals. In other words, the appearance of a discontinuity interval marks a new growth stage for the larvae, and therefore, detecting discontinuity intervals facilitates instar classification.

Theoretically, discontinuity intervals are based on the density of measurements in the intervals being zero, as no sample larvae with measurements in the intervals can be collected from the field. However, in practice, the density is rarely zero; instead, it is a value lower than those of adjacent areas because the size and location of the intervals vary between individuals. The size of sclerotized parts will become stable when the larvae enter a new instar; hence, the density of measurements for this size quickly increases, and peaks emerge in the frequency distribution of the measurements. Therefore, we define the start point of a discontinuity interval as the densest measurement for the current instar and define the end as the densest measurement for the next instar. Many studies have shown that, within an instar, the measurements approximate a normal distribution, and the densest location is also the mean for a normally distributed measurement (McClellan and Logan 1994, Hammack et al. 2003, Panzavolta 2007, Dallara et al. 2012). The pattern of the frequency distribution alternates between peaks and valleys, as illustrated in Fig. 1. Indeed, the pattern in Fig. 1 appears in the frequency distributions of measurements of sclerotized parts (primarily the head capsule) for numerous species of insects that have been studied (Leibee et al. 1980, De Moor 1982, Loerch and Cameron 1983, Price and Craic 1984, Forsghler and Nordin 1991, Hammack et al. 2003, Panzavolta 2007, Chen and Seybold 2013).

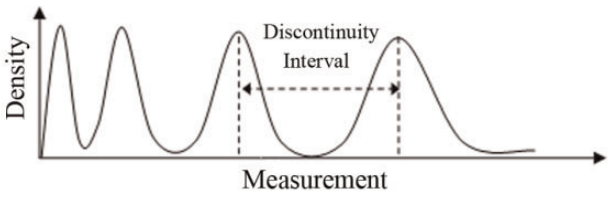

Fig. 1. Frequency distribution pattern of measurements for sclerotized body parts in insects.

Discontinuity Points. In this article, the point at which the measurement density is the lowest within a discontinuity interval is called the discontinuity point. In other words, the discontinuity point is a local minimum in the frequency distribution of measurements.

Instar classification using discontinuity points: Let X be a random variable for the measurements of a larval sclerotized body part and $x_j^*(1 \le j \le k)$ be the jth discontinuity point, with $x_0^* = 0$ and $x_{k+1}^* = \infty$. Then, the larvae will be classified into $k+1$ instars, and the boundary between the jth instar and the $j+1$ instar is x_j^*. In practice, the local minima of the frequency distributions of measurements have been used as the boundaries between successive instars in some studies, without a stated rationale (Leibee et al. 1980, De Moor 1982, McClellan and Logan 1994, Chen and Seybold 2013).

Estimation of Discontinuity Points. Let $f(x)$ be the density function of X and $x_j'(1 \le j \le k)$ be the jth local minimum of $f(x)$, with $x_0' = 0$ and $x_{j+1}' = \infty$. If there is a clear peak in the curve of $f(x)$ between x_j', x_{j-1}' and x_{j+1}', then x_j' will be a discontinuity point of X.

Adaptive kernel smoothing estimation for $f(x)$: Kernel density estimation is a kernel smoothing method used to estimate the density function of random variables. Compared with a histogram, kernel density estimation is a smoothing function and is also more efficient. Chen and Seybold (2013) previously used kernel density estimation to determine the number of peaks in a frequency distribution, as opposed to using histograms and visual determination as performed in other studies. The kernel density estimation for $f(x)$ is given by

$$\tilde{f}(x) = (nh)^{-1}\sum_{i=1}^{n} K\{(x - X_i)/h\}, \quad (1)$$

where K is a function satisfying $\int K(x) = 1$, h is a bandwidth smoothing parameter (often called the bin width in histogram methods), $h > 0$, and K has a continuous second derivative and is symmetric around zero. The majority of larvae go through multiple instar stages, which means there will be multiple peaks for $f(x)$, making $f(x)$ a multi-modal density function. Using a fixed bandwidth approach such as equation 1 to estimate $f(x)$ can result in under-smoothing in areas with sparse observations and over-smoothing in other areas. In contrast, varying the bandwidth to better reflect the sample data increases flexibility, reducing estimation variance in areas with few observations and estimation bias in areas with many observations. Kernel density estimation methods that rely on varying bandwidths are generally referred to as adaptive kernel density estimation methods. The adaptive kernel density estimation for $f(x)$ is given by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i} K\{(x - X_i)/h_i\}, \quad (2)$$

where $h_i = h \times \lambda_i$, $\lambda_i = \lambda(X_i)^i = \sqrt{G/\tilde{f}(X_i)}$, and G is the geometric mean of $\tilde{f}(X_i)(i = 1, 2, \cdots, n)$. According to equation 2, a larger bandwidth will be used in lower density areas, and a smaller bandwidth will be used in higher density areas in adaptive kernel smoothing methods. This smoothing exaggerates the peaks and valleys in the frequency distribution of body part measurements and will therefore be beneficial for instar classification.

Bandwidth Selection. It is crucial to choose an appropriate bandwidth for equations 1 and 2 to estimate $f(x)$. If h is too small, significant noise will remain in the data, and if h is too large, then it will not accurately reflect the distributed characteristics embedded in the data. The value of $f(x)$ at x estimated using equations 1 and 2 primarily depends on the number of observations between $[x - h, x + h]$, which corresponds to $2h$. To identify the lower boundary of the discontinuity interval, h must be half the size of the minimal discontinuity interval.

Brooks' (Dyar's) rule states that the growth of sclerotized parts proceeds in a regular geometric progression between successive instars (Crosby 1973). The exponential equation that has usually been fit to Brooks' rule in related studies is

$$y_j = ae^{bj}(a, b > 0, j = 0, 1, 2, \cdots, m), \tag{3}$$

where y_j is the mean of X for the jth instar, and m is the number of larval instars. Therefore, the minimal discontinuity interval lies between the zeroth instar and the first instar, with a size of $a(e^b - 1)$, and the optimal bandwidth is given by

$$h = \frac{1}{2}a(e^b - 1). \tag{4}$$

Let $\bar{y}$ and G_y be the mean and geometric mean of y, respectively. The following equations can be derived as the result of y being an exponentially increasing series:

$$bm = \sqrt{24\left\{\frac{\bar{y}}{G_y} - 1 - \sum_{q=2}^{\infty}\left(\frac{bm}{\sqrt[2q]{2^{2q}(2q+1)!}}\right)^{2q}\right\}} \tag{5}$$

$$a = G_y/e^{\frac{bm}{2}}. \tag{6}$$

In many related studies, the values of bm have been shown to be less than 2 (Panzavolta 2007, Dallara et al. 2012, Chen and Seybold 2013). Here, we assumed bm to be far less than $\sqrt[4]{2^4 5!} \approx 6.6195$, so $\sum_{q=2}^{\infty}\left(\frac{bm}{\sqrt[2q]{2^{2q}(2q+1)!}}\right)^{2q}$ is close to zero, and we can define

$$r = \sqrt{24\left(\frac{\bar{y}}{G_y} - 1\right)}$$

as the approximation of bm. We can also obtain an accurate approximation of a larger bm by solving equation 9 (see Appendix) with a higher order bm.

Equation 6 can be written as $c = G_y/e^{\frac{r}{2}}$, using r to approximate bm. Let $\overline{X}$ and G_X be the mean and geometric mean of X, respectively, which can be used to estimate $\bar{y}$ and G_y, respectively. Then, the estimation of r is given by

$$\hat{r} = \sqrt{24\left(\frac{\overline{X}}{G_X} - 1\right)}, \tag{7}$$

and the estimations of c and b are given by $\hat{c} = G_X/e^{\frac{\hat{r}}{2}}$ and $\hat{b} = \hat{r}/m$, respectively. The optimal bandwidth is estimated as

$$\hat{h} = \frac{1}{2}\hat{c}(e^{\hat{r}/m} - 1), \tag{8}$$

which is obtained by substituting $\hat{c}$ and $\hat{b}$ into equation 4.

Steps for Classifying Instars Using the Adaptive Kernel Smoothing Method. Equation 5 indicates that the estimated value of bm, which is obtained using equation 3, must be an accurate approximation of $\hat{r}$ with correctly classified instars, which can also be used as a criterion to verify the results of instar classification; another criterion is Crosby's growth rule. Using these two criteria to filter the instar classifications, the adaptive kernel smoothing method was performed according to the following five steps: 1) The range of m was specified by referring to the number of instars in other species of the same order or from related reports or experience. 2) For a given m, $\hat{r}$ and $\hat{h}$ were computed with equations 7 and 8, respectively, and then substituted into equation 2 to estimate the density function of the data. 3) The discontinuity points were identified using the density function estimated in step 2. These points corresponded to the separation between each instar. This resulted in one subset of data describing each instar and the estimated number of instars, which was recorded as $\hat{m}$. 4) The means were computed for each subset using equation 3 to obtain $\hat{b}$, and these means were also used to calculate Crosby's growth ratios. 5) A series of $\hat{b}\hat{m}$ values were calculated by repeating steps 2–4 for each m. The optimal solution required $|\hat{b}\hat{m} - \hat{r}|$ to be the minimum and Crosby's growth rule to be observed.

Algorithm. The adaptive kernel density estimation was implemented using the now-standard adaptive, two-stage estimator proposed by Abramson (1982). The iterative procedure was as follows: the initial values of $\tilde{f}(X_i)$, G, λ_i and $\hat{f}(X_i)$ were computed using equations 1 and 2, the pth iterative step was to set $\tilde{f}(X_i)^{(p)} = \hat{f}(X_i)^{(p-1)}$ to compute $G^{(p)}$, $\lambda_i^{(p)}$ and $\hat{f}(X_i)^{(1)}$, and the steps were repeated until the change in $G^{(\cdot)}$ was less than 0.0001. To find the local minimum of $\hat{f}(x)$, we found the first and second derivatives of $\hat{f}(x)$. We then identified zeroes in the first derivative and determined whether these zeroes were the local minima of $\hat{f}(x)$ by calculating the values of the second derivative at these zeroes. To calculate the local minima of the density function, the first and second derivatives of the density function were first solved. Then, the zeroes in the first derivative of the density function were solved, and the second derivatives of the density function were solved at these zeroes to judge whether the zeroes were local maxima or minima. There are a number of zeroes in the first derivative because $\hat{f}(x)$ is a function with multiple peaks and valleys. To rapidly find these zeroes, we expanded the first derivative as a Chebyshev polynomial series and truncated the series for sufficiently large N. Then, as solutions of the zeroes of the first derivative, the zeroes of the truncated Chebyshev series were identified by the Chebyshev–Frobenius companion matrix method, which was able to accurately and simultaneously find all of the zeroes within the target interval with less calculation (Boyd and Gally 2007). Fitting to equation 3 was conducted by regressing the natural log of y_j for each instar against the corresponding instar number. All the computations were performed in MATLAB (Math Works 2007, America), including the calculation of Crosby's growth ratios.

Table 1. Parameter (i.e.,$\log(\hat{a})$,$\hat{b}$, $\hat{r}$, $\hat{b}m$) estimates of equations 3 and 7 based on Crosby's (1974) instar determinations for the larvae of *A. tillyardianum*

Variable	The test set				The standardization set and the test set			
	$\log(\hat{a})$	$\hat{b}$	$\hat{b}m$	$\hat{r}$	$\log(\hat{a})$	$\hat{b}$	$\hat{b}m$	$\hat{r}$
HCW	4.4179	0.2181	1.9631	1.9947	4.3740	0.2264	2.0372	2.0132
HCL	4.6308	0.2302	2.0720	2.0914	4.5768	0.2334	2.1008	2.0463
HCPW	4.3458	0.2065	1.8588	1.8993	4.3377	0.2080	1.8724	1.9380
ML	3.7067	0.2316	2.0842	2.1203	3.6720	0.2417	2.1752	2.1477
AS3L	3.2616	0.2574	2.3163	2.3341	3.2762	0.2605	2.3448	2.2224

Results

The accuracy of the estimation of bm with r was assessed by comparing $\hat{bm}$ and $\hat{r}$ based on the test set and all the data for all five of the variables; the exponential growth rate between successive instars $\hat{b}$ was obtained using equation 3 with the data that had been split into nine subsets, which corresponded to Crosby's nine instars (1974). In this case, m is equal to 9, assuming that the separation of nine instars by Crosby (1974) is correct, and $\hat{bm}$ is the product of $\hat{b}$ and m. $\hat{r}$ is calculated using equation 7. Table 1 reveals that the values of $\hat{r}$ were very close to the value of $\hat{bm}$ for each variable, indicating that using r to estimate bm is reasonable and accurate. There were also only small discrepancies between $\hat{bm}$ and $\hat{r}$ when these values were estimated using the test set and all of the data, indicating that the estimation of bm with r is stable between different samples.

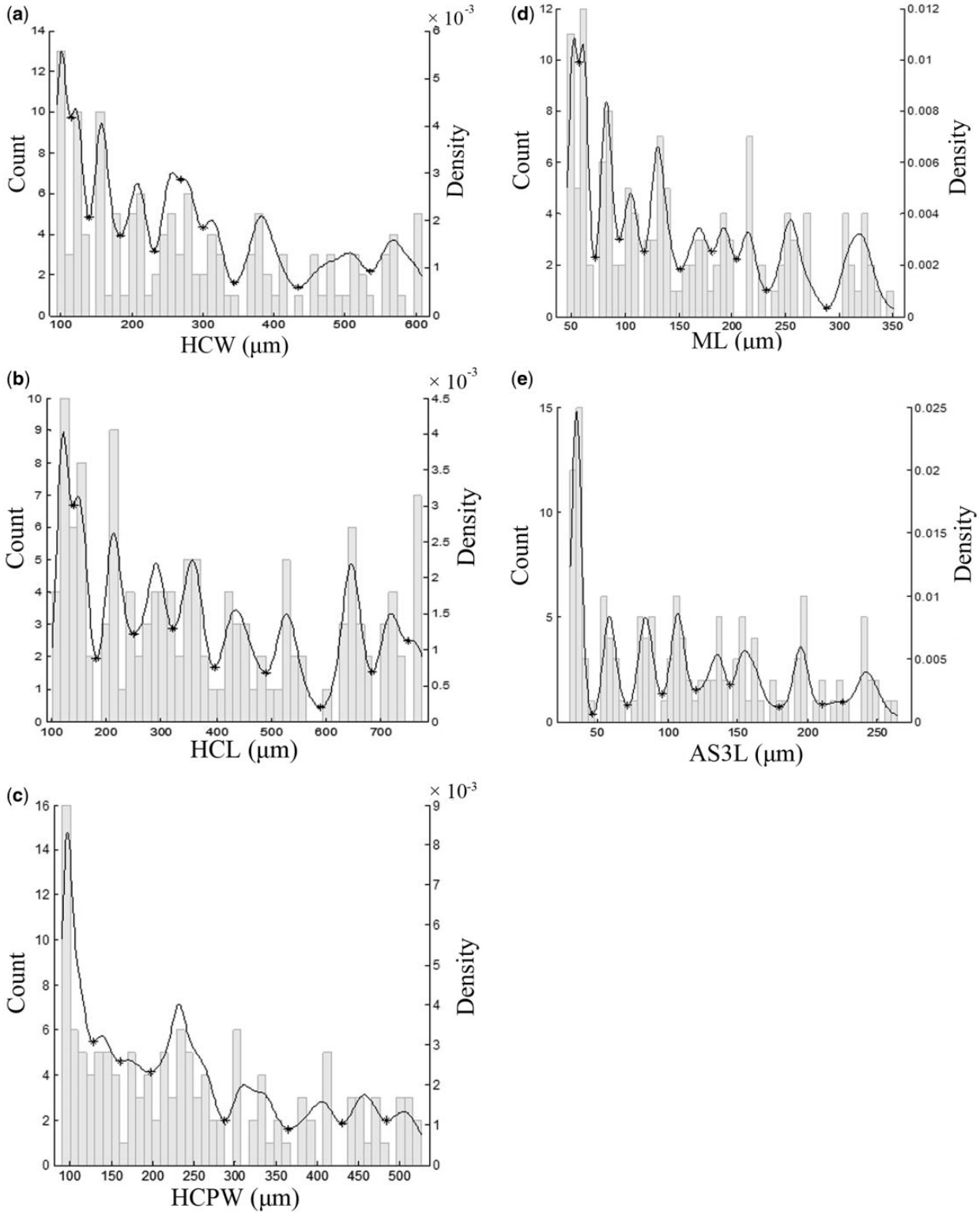

Fig. 2. Adaptive kernel density estimates of five variables with $m = 9$. *Local minimum of the density estimation curve based on the test set. (a) $h = 11.2713$; (b) $h = 14.8690$; (c) $h = 9.8230$; (d) $h = 5.9555$; (e) $h = 4.3462$.

Five variables were used to separate larval instars for *A. tillyardianum* using the adaptive kernel smoothing method based on the test set. The range of m was 4 to 10, which referred to the number of larval instars identified in another Simuliidae member species. Crosby's growth ratio rule, which states that some instars are likely grouped together when the difference between the growth ratios of two successive instars is more than 0.1, was applied as an additional criterion when choosing m.

The kernel density estimation of HCW produced nine local minima when $m = 9$, although the fifth local minimum was not a valid discontinuity point because no sharp peaks were found between the fifth and sixth local minima (Fig. 2a). Consequently, the larvae samples were classified into nine groups with eight discontinuity points. Within that classification, $\hat{b}\hat{m}$ (1.9783) was nearest to $\hat{r}$ (1.9631), and the max Crosby's ratio (0.0693) was less than 0.1 (Tables 1 and 2). The results suggest that the optimal separating solution is nine instars when using HCW as the separating variable.

The kernel density estimation for HCL produced nine local minima when $m = 9$, and the ninth local minimum was not a valid discontinuity point because no sharp peaks were found to its right (Fig. 2b). Consequently, the larval samples were classified into nine groups with eight discontinuity points. Within that classification, $\hat{b}\hat{m}$ (2.0660) was nearest to $\hat{r}$ (2.0720), and the max Crosby's ratio (0.1089) was very close to 0.1 (Tables 1 and 2). The results suggest that the optimal separating solution is nine instars when using HCL as the separating variable.

The kernel density estimation for HCPW produced seven local minima when $m = 9$, and all of the local minima were valid discontinuity points (Fig. 2c). Consequently, the larval samples were classified into eight groups with seven discontinuity points. Within that classification, $\hat{b}\hat{m}$ (1.8674) was nearest to $\hat{r}$ (1.8588), and the maximum Crosby's ratio (0.1082) was very close to 0.1 (Tables 1 and 2). These results suggest that separation into eight instars is the optimal solution when HCPW is used as the separating variable.

The kernel density estimation for ML produced nine local minima when $m = 9$, and all of the local minima were valid discontinuity points (Fig. 2d). Consequently, the larval samples were classified into 10 groups with nine discontinuity points. Within that classification, $\hat{b}\hat{m}$ (2.0846) was nearest to $\hat{r}$ (2.1203), and the maximum Crosby's ratio (0.1058) was very close to 0.1 (Tables 1 and 2). These results suggest that the separation into 10 instars is the optimal solution when ML is used as the separating variable.

No reasonable solution was found using the kernel smoothing method when AS3L was used as the separating variable. Although $\hat{b}\hat{m}$ (2.2450) was nearest to $\hat{r}$ (2.3341) when $m = 4$ (Tables 1 and 2), the maximum Crosby's ratio (0.3427) was far higher than 0.2, indicating that the Crosby's growth ratio rule was not observed (Table 2).

To validate the findings concluded from the test set, we applied the adaptive kernel smoothing method to analyze the standardization set, and the results of the density estimation when $m = 9$ are presented in Fig. 3. Figure 3 indicates that the number of valid discontinuity points on the estimated density curves is eight for HCW, HCL, and HCPW, nine for ML, and seven for AS3L, which is in accordance with the result obtained with the test set. However, the peaks and valleys of the estimated density curve in Fig. 3 are more obvious than those in Fig. 2, especially for early instars. The reason for this difference may be that the sample size of the standardization set is larger than that of the test set, and the adaptive kernel smoothing method performs better with samples of a larger size.

Figures 2 and 3 indicate that for each of the five variables, the discontinuity points between early instars can be detected by the adaptive kernel smoothing method. In particular, the peaks corresponding to early instars on the estimated density curve are quite sharp for each variable in Fig. 3, suggesting that there is a significant variation in the size of sclerotized

Table 2. Estimates of the parameters in each step of the kernel smoothing method used to classify instars of *A. tillyardianum* larvae based on the test set

Variable	Statistic	m						
		4	5	6	7	8	9	10
HCW	$\hat{m}$	4	4	7	7	7	9	10
	h	29.3708	22.2707	17.9158	14.9778	12.8640	11.2713	10.0284
	$\hat{b}$	0.4352	0.4229	0.2756	0.2757	0.2757	0.2198	0.1872
	$\hat{b}\hat{m}$	1.7409	1.6915	1.9293	1.9301	1.9301	1.9783	1.8716
	Crosby's ratio[a]	0.2930	0.2742	0.1263	0.1165	0.1165	0.0693	0.0752
	R^2	0.9559	0.9651	0.9912	0.9922	0.9922	0.9970	0.9904
HCL	$\hat{m}$	3	6	8	8	8	9	10
	h	39.0390	29.5194	23.7041	19.7916	16.9825	14.8690	13.2219
	$\hat{b}$	0.7376	0.3039	0.2299	0.2299	0.2299	0.2296	0.2125
	$\hat{b}\hat{m}$	2.2129	1.8234	1.8394	1.8394	1.8394	2.0660	2.1246
	Crosby's ratio[a]	0.2529	0.2012	0.1837	0.1837	0.1837	0.1089	0.1133
	R^2	0.9871	0.9869	0.9677	0.9677	0.9677	0.9848	0.9905
HCPW	$\hat{m}$	3	4	5	5	7	8	9
	h	25.4083	19.3188	15.5688	13.0319	11.2031	9.8230	8.7448
	$\hat{b}$	0.6599	0.4151	0.3146	0.3146	0.2606	0.2334	0.1959
	$\hat{b}\hat{m}$	1.9796	1.6606	1.5729	1.5729	1.8245	1.8674	1.7632
	Crosby's ratio[a]	0.2008	0.2570	0.2592	0.2592	0.1082	0.1082	0.1560
	R^2	0.9905	0.9694	0.9694	0.9272	0.9427	0.9782	0.9664
ML	$\hat{m}$	4	6	6	9	9	10	10
	h	15.6716	11.8402	9.5025	7.9311	6.8035	5.9555	5.2948
	$\hat{b}$	0.4937	0.3457	0.3471	0.2018	0.2025	0.2025	0.2026
	$\hat{b}\hat{m}$	1.9750	2.0744	2.0227	1.8161	1.8225	2.0846	2.0261
	Crosby's ratio[a]	0.1367	0.0887	0.1109	0.1648	0.1581	0.1058	0.1058
	R^2	0.9892	0.9847	0.9852	0.9771	0.9779	0.9893	0.9887
AS3L	$\hat{m}$	5	7	8	8	8	8	8
	h	11.6309	8.7328	6.9803	5.8095	4.9731	4.3462	3.8591
	$\hat{b}$	0.4490	0.3129	0.2574	0.2575	0.2575	0.2315	0.2307
	$\hat{b}\hat{m}$	2.2450	2.1903	2.0595	2.0596	2.0596	2.0886	2.0964
	Crosby's ratio[a]	0.3247	0.1611	0.1611	0.1611	0.1611	0.1611	0.1611
	R^2	0.9140	0.9814	0.9621	0.9631	0.9631	0.9470	0.9472

[a]The value of Crosby's ratio in Table 2 is the maximum of Crosby's ratio for a given m.

body parts among the early instars of *A*. *tillyardianum* larvae. Thus, as Crosby (1974) notes, it is not possible to group the early instars using a histogram or scatter plot. In contrast, adaptive kernel density estimation with a reasonable bandwidth can identify discontinuity points between early instars, making the separation of early instars possible.

Discussion

Crosby (1974) presented biometric evidence demonstrating the existence of nine instars for *A*. *tillyardianum*, consistent with the results obtained in this study using the adaptive kernel smoothing method with HCW and HCL as the separating variables. Although the instar

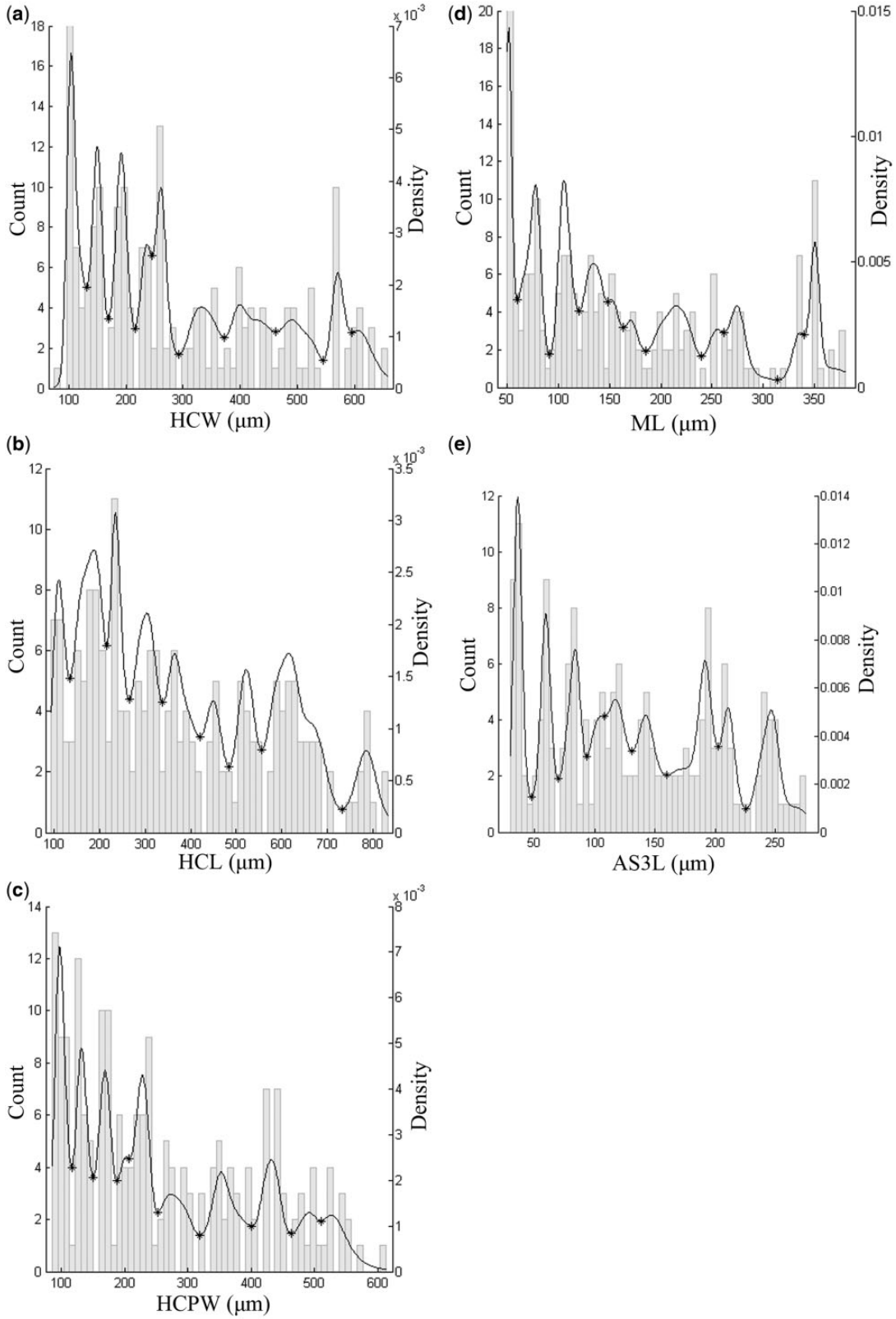

Fig. 3. Adaptive kernel density estimates of five variables with $m = 9$. *Local minimum of the density estimation curve based on the standard set. (a) $h = 10.5223$; (b) $h = 13.1736$; (c) $h = 9.1904$; (d) $h = 5.1915$; (e) $h = 4.6003$.

determinations using HCPW and ML did not completely agree with Crosby's result, these determinations were very close to the result obtained by Crosby. This discrepancy may be because the increases in size that occur between instars are not consistent across the four variables. The separation of larvae into nine instars using HCW and HCL was more reliable than separation using the other variables because the estimated number of instars ($\hat{m}$) equaled the assumed value (m) when these two variables were used to separate the larvae with the adaptive kernel smoothing method. Moreover, HCW is the morphological separating variable that has been most widely used in studies of this type. Crosby (1974) determined that the increase in the length of antennal segment 3 between instars did not follow Brooks' rule in *A. tillyardianum*, which was confirmed by the derivation of an unreasonable solution when AS3L was used to separate the larvae using the adaptive kernel smoothing method. For *A. tillyardianum*, Crosby (1974) noted that the separation of instars using frequency distribution methods is difficult for early instars. However, we were able to obtain results that agreed with Crosby's results, indicating that the adaptive kernel smoothing method may be the most effective frequency distribution analysis method reported to date.

The probability of misclassification when using discontinuity points as class limits to group the larvae was calculated using Got's (1988) and McClellan and Logan's (1994) method with the assumption that the measurement variables within each instar were normally distributed. The lower the probability at the discontinuity point between successive instars, the less the distribution between the two instars overlapped and the lower the misclassification probability was. The majority of individuals could be separated into the correct instar stages when the larvae were grouped based on discontinuity points because the majority of the larvae were found between neighboring discontinuity points (Figs. 2 and 3). In cases where there was no overlap between any of the instars in the distribution for a particular separating variable, the probability at the discontinuity points, and therefore the misclassification probability, could potentially be zero.

Table 1 reveals that r is an accurate approximation of bm, which could serve not only as a standard for assessing the quality of instar determination but also as a variable for selecting the optimal bandwidth for a nonparametric density estimation using kernel smoothing methods and histograms. The accuracy of the estimation of bm using r primarily depends on whether the growth of sclerotized body parts strictly proceeds along an exponential progression, such as in equation 3. The estimation of bm using r was still accurate for the variables, including AS3L, although AS3L does not follow Brooks' rule.

A good morphological separating variable ideally exhibits little variation within instars and a large variation between instars, yielding large discontinuity intervals and sharp peaks in the distribution for that variable. Without such a variable, as seen in *A. tillyardianum*, it can be difficult to assign clear-cut divisions because the frequency distribution overlaps significantly between instars. However, the adaptive kernel smoothing method was able to identify divisions between instars more effectively than previous methods by strengthening the pattern shown in Fig. 1 through the use of a variable bandwidth. Moreover, the kernel density estimation function is a continuous smooth curve that allows the distributed character of the measurement variables to be analyzed. Although there are many different methods to select an optimal bandwidth for a kernel density estimation, including least-squares cross-validation and biased cross-validation, the bandwidths chosen using these methods resulted in over-smoothing and under-smoothing, respectively, of the density estimation in this study. We found that the optimal bandwidth was given by equation 8 and that a variable bandwidth was a better estimator of the multimodal distribution displayed in Fig. 1 than previous methods.

This study demonstrated a new methodology for the classification of instars and a novel solution for the selection of an optimal bandwidth or bin width to use when estimating density measurements of sclerotized body parts using nonparametric methods such as histograms. We also describe a kernel estimator, which represents a new standard to assess the results of instar classification. This method can improve the separation of sampled larvae into instars and the estimation of instar distributions from field-collected data. The adaptive kernel smoothing method is an efficient and robust methodology that can be used by researchers to correctly classify instars in their own datasets. Compared with previous methods, the adaptive kernel smoothing method exhibits two improvements: 1. The selection of bandwidth is based on the growth process of the sclerotized parts of the insect. 2. The application of an adaptive kernel density estimation to the density distribution of the measurements more accurately captures the multimodal features of the measurement density distributions. Histograms for each variable are depicted in Figs. 2 and 3 and illustrate the difficulty of separating instars for *A. Tillyardianum* using histograms. The adaptive kernel smoothing method produced nearly the same results as those of Crosby. Because of the variation in growth rates and other processes during the different stages of growth of sclerotized body parts, the use of the adaptive kernel smoothing method to separate instars might produce different results when measurements of different sclerotized parts are employed. In previous studies, HCW and HCL have often been adopted. The use of these two variables is also recommended in this study. Another solution is to establish a multidimensional kernel density to estimate instars based on multiple instar variables. The method for the classification of instars proposed in this study offers researchers a new tool to differentiate instars. When applying this method, related parameters such as the Crosby ratio and bm value (Table 2) should be calculated for the potential instar classifications (different m values). Researchers can then choose the optimum classification based on these parametric values and their own research experience.

The increases in the sizes of sclerotized body parts that occur in insects between instars are multidimensional; thus, a univariate model is not sufficient to describe differences between instars, and instar classifications obtained using different variables may not agree. Therefore, future studies should focus on the use of multivariate kernel smoothing methods to estimate instar distributions and classify instars.

Acknowledgments

We thank Degui Yang (SCAU, College of Mathematics and Informatics) for his comments on our mathematical reasoning and on an earlier version of the manuscript. This work was supported by the Natural Science Fundation Project of Guangxi (2012GXNSFBA053050), Special Basic Scientific Research Project of Guangxi Academy of Agricultural Sciences (2013YZ06, 2014YQ32, 2015YM15, 2015YT38), Guangxi Key Laboratory of Biology for Crop Diseases and Insect Pests (14-045-50-ST-12).

References Cited

Abramson, I. S. 1982. On bandwidth variation in kernel estimates a square root law. Ann. Stat. 10: 1217–1223.

Beaver, R. J., and J. P. Sanderson. 1989. Classifying instars of the navel orangeworm (Lepidoptera: Pyralidae) based on observed head capsule widths. J. Econ. Entomol. 82: 716–720.

Boyd, J. P., and D. H. Gally. 2007. Numerical experiments on the accuracy of the Chebyshev–Frobenius companion matrix method for finding the zeros of a truncated series of Chebyshev polynomials. J. Comput. Appl. Math. 205: 281–295.

Chapman, R. F. 1998. The insects: structure and function. Cambridge University Press, Cambridge, UK.

Chen, Y., and S. J. Seybold. 2013. Application of a frequency distribution method for determining instars of the beet armyworm (Lepidoptera: Noctuidae) from widths of cast head capsules. J. Econ. Entomol. 106: 801–805.

Crosby, T. K. 1973. Dyar's rule predated by Brooks' rule. N. Z. Entomol. 5: 175–176.

Crosby, T. K. 1974. Studies on Simuliidae (Diptera), with particular reference to *Austrosimulium*, Ph.D. dissertation, University of Canterbury, Christchurch, New Zealand.

Dallara, P. L., M. L. Flint, and S. J. Seybold. 2012. An analysis of the larval instars of the walnut twig beetle, *Pityophthorus juglandis* Blackman (Coleoptera: Scolytidae), in northern California black walnut, *Juglans hindsii*, and anew host record for *Hylocurus hirtellus*. Pan-Pac. Entomol. 88: 348–366.

De Moor, F. C. 1982. Determination of the number of instars and size variation in the larvae and pupae of *Simulium chutteri* Lewis 1965 (Diptera: Simuliidae) and some possible bionomical implications. Can. J. Zool. 60:1374–1382.

Fink, T. J. 1984. Errors in instar determination of mayflies (Ephemeroptera) and stoneflies (Plecoptera) using the simple frequency, Janetschek, Cassie and Dyar's law methods. Freshwater Biol. 14: 347–365.

Forsghler, B. T., and G. L. Nordin. 1991. Instar determination in the cottonwood borer, *Plectrodera scalator* (Fab.) (Coleoptera: Cerambycidae). Coleopts. Bull. 45: 165–168.

Got, B. 1988. Determination of instar of the European corn borer (Lepidoptera: Pyralidae) based on a distributionmodel of head capsule widths. Ann. Entomol. Soc. Am. 81:91–98.

Hammack, L., M. M. Ellsbury, R. L. Roehrdanz, and J. L. Pikul. 2003. Larval sampling and instar determination in field populations of northern and western corn rootworm (Coleoptera: Chrysomelidae). J. Econ. Entomol. 96: 1153–1159.

Kishi, Y. 1971. Reconsideration of the method to measure the larval instars by use of the frequency distribution of head-capsule widths or lengths. Can. Entomol. 103: 1011–1015.

Leibee, G. L., B. C. Pass, and K. V. Yeargan. 1980. Instar determination of clover root curculio, *Sitona hispidulus* (Coleoptera: Curculionidae). J. Kans. Entomol. Soc. 53: 473–475.

Loerch, C. R., and E. A. Cameron. 1983. Determination of larval instars of the bronze birch borer, *Agrilus anxius* (Coleoptera:Buprestidae). Ann. Entomol. Soc. Am. 76: 948–952.

Logan, J. A., B. J. Bentz, J. C. Vandygriff, and D. L. Turner. 1998. General program for determining instar distributions from headcapsule widths: example analysis of mountain pine beetle (Coleoptera: Scolytidae) data. Environ. Entomol. 27: 555–563.

Math Works. 2007. MATLAB reference guide. Math Works, Natick, MA.

McClellan, Q. C., and J. A. Logan. 1994. Instar determination for gypsy moth (Lepidoptera:Lymantriidae) based on the frequency distribution of head capsule widths. Environ. Entomol. 23: 248–253.

McNeil, J. N. 1978. The number of larval stages of *Thymelicus lineola* (Lepidoptera:Hesperudae) in Eastern Canada. Can. J. Zool. 110: 1293–1295.

Panzavolta, T. 2007. Instar determination for *Pissodes castaneus* (Coleoptera: Curculionidae) using head capsule widths and lengths. Environ. Entomol. 36: 1054–1058.

Price, P. W., and T. P. Craic. 1984. Life history, phenology and survivorship of a stem-calling sawfly, *Euura lasiolepis* (Hymenoptera: Tenthredinidae), on the arroyo willow, *Salix lasiolepis*, in Northern Arizona. Ann. Entomol. Soc. Am. 77: 712–719.

Schmidt, F. H., R. K. Campbell, and S. J. Trotter. 1977. Errors in determining instar numbers through head capsule measurements of a Lepidopteran—a laboratory study and critique. Ann. Entomol. Soc. Am. 70: 750–756.

Received 7 April 2015; accepted 10 October 2015.