Fisheries and Oceans Science | Pêches et Océans Sciences

C S A S

Canadian Science Advisory Secretariat

S C C S

Secrétariat canadien de consultation scientifique

Research Document 2001/071

Document de recherche 2001/071

Smoothing Length Frequency Data via Kernel Density Estimation

Lissage de données de fréquence de longueur à l'aide de l'estimation de la densité par la méthode du noyau

B.P. Healey and N.G. Cadigan

Science, Oceans, and Environment Branch
Fisheries and Oceans Canada
P.O. Box 5667
St. John's, NF
A1C-5X1

ISSN 1480-4883
Ottawa, 2001

Canada

Abstract

We use kernel smoothing to estimate the weekly length frequency of the commercial landings in NAFO Divisions 3Ps and 3KL. This is done for each gear type used in the fishery. The estimates are based on irregular length frequency sampling conducted by commercial fishers and sentinel participants. The goal of the smoothing is to predict weekly proportions-at-length in 1998-2000, for selected gear types and regions that have reported landings. This information is required in the analysis of tag recapture data (see Cadigan and Brattey, 2000).

Résumé

Nous utilisons le lissage par noyau pour estimer les fréquences de longueur des débarquements commerciaux hebdomadaires dans les divisions 3KL et la sous-division 3Ps de l'OPANO. Cette procédure est effectuée pour chaque type d'engin utilisé dans la pêche. Les estimations sont fondées sur l'échantillonnage irrégulier des fréquences de longeur réalisé par les pêcheurs commerciaux et les participants à la pêche sentinelle. Le lissage vise à prédire les proportions hebdomadaires des poissons selon la longueur sur la période 1998-2000, pour certains types d'engin et les régions où l'on signale des débarquements. Ces données sont nécessaires pour l'analyse des données de recapture dans des études de marquage. (voir Cadigan et Brattey, 2000).

Introduction

Weekly estimates of catch-at-length are required by Cadigan and Brattey (2000) to estimate weekly stock size for cod in NAFO Divisions 3K and 3L, and Subdivision 3Ps using exploitation rate estimates obtained from tagging data. Estimates of catch-at-length are obtained from estimates of weekly landings and the length composition of the catch. In recent assessments of the 2J3KL and 3Ps cod stocks, this has been accomplished by averaging length frequencies for geographic-gear combinations. However, small sample sizes and intermittent sampling are common, especially for the post-moratorium 2J3KL cod fishery. This has meant that time periods with no length frequency samples available (most often a week, but possibly a month or a quarter) had to be interpolated. The interpolation decisions were subjective and resulted in abrupt changes in the length frequency estimates over short time scales. It is desirable to have a procedure that can automatically interpolate length frequencies in a more reasonable manner.

Sample Length Frequency and Weight

Length frequency sampling data is available from three sources: commercial sampling (traditionally on-board monitors), port sampling, and the sentinel survey. Hereafter, the commercial and port samples shall be jointly referred to as commercial data. We use all the length frequencies from 1998 to 2000, in both 2J3KL and 3Ps. The available data are disaggregated by length (1cm groups ranging from 1 to 200cm), gear type, NAFO unit area (e.g. 3Kd), and sample source (commercial or sentinel). For gear types, we consider only gillnet (excluding the experimental 3.25” mesh used by sentinel participants), handline, linetrawl and ottertrawl (includes seine). The length classes considered in this analysis are 40-120cm. The NAFO sub-unit areas are aggregated into six regions (Table 1; Figure 1) defined by Cadigan and Brattey (2000).

The number of length frequency samples selected for analysis was 10961. Of this total, 3646, 3445, and 3870 samples were taken in 1998, 1999, and 2000, respectively. The regional totals (following the region order in Table 1) are: 3419, 1547, 2605, 1426, 1562, and 402, respectively. The majority of sampling came from gillnet catches (n=8330); the number of linetrawl (1757), handline (566), and ottertrawl (308) samples are much smaller. Approximately 81% of samples came from sentinel enterprises.

Density Estimation

Given length frequency samples, how should proportions-at-length be estimated? Consider the simple case in which there is just one sample. An estimate of proportion-at-length is:

$$\hat{p}_l = \frac{n_l}{\sum_{l=1}^{L} n_l} \qquad [1]$$

where n_l is the number sampled at length l, and L is the total number of length classes that were sampled. The problem with this approach is that the result may be non-smooth, particularly if the sample size is small.

Various methods can be used to produce smoother estimate of $\hat{p}_l$. A common method is kernel smoothing, which involves locally weighted averages of length frequencies for neighboring length classes. A kernel function is used to define the local neighborhood. For extended details and discussion on kernel smoothing, refer to Silverman (1986) or Härdle (1990).

For the one-sample case considered above, the kernel estimate of p_l is:

$$\hat{p}_l = \frac{\text{[illegible]}}{\sum_{i=1}^{L} \text{[illegible]}}, \text{ where } \text{[illegible]} = \sum_{i=1}^{L} K(l,l_i) n_i. \quad [2]$$

The function $K(l,l_i)$ is the kernel function; it provides the weight given to the numbers sampled in each length class l_i for predicting the proportion (p_l) in length class l. The kernel function is such that:

$$0 \le K(l,l_i) < 1, \text{ and } \sum_{i=1}^{L} K(l,l_i) = 1 .$$

The value of $K(l,l_i)$ is large if l is close to l_i; otherwise, it tends to 0 as $|l - l_i|$ increases. That is, length classes (l_i) close to the point under consideration (l) get more weight in predicting p_l.

When there are multiple samples within a week, length frequencies can be constructed by adding the samples within a region/gear type/week cell if sampling is random and if the entire catch of a vessel is sampled. That is, if we denote the total length frequency as

$$n_l = \sum_{j=1}^{S} n_{jl}, \text{ where } S \text{ is the total number of samples.}$$

then proportions-at-length can be estimated using [1] or [2] and the total length frequency.

The more important problem for us is predicting length frequencies in weeks where no sampling is available. To fill these gaps, we use a bivariate smoother, smoothing over length and weeks simultaneously.

We smooth the length frequencies in two dimensions – time (week) and length. Let $p_{l(t)}$ denote the proportion at length in week t. The bivariate kernel estimate of $p_{l(t)}$ is:

$$\hat{p}_{l(t)} = \frac{\text{[illegible]}}{\sum_{l} \text{[illegible]}} \quad [3]$$

where $\text{[illegible]} = \sum_{i=1}^{L}\sum_{j=1}^{W} K(l-l_i, t-t_j)n_{ij}$, In [3], L is the number of length classes, and W is the number of weeks where sampling has occurred. Notice that the kernel function used now involves two variables, length and week. The kernel function used is the bivariate normal (or Gaussian) kernel:

$$K(x,y) = \frac{1}{2\pi k h_x h_y}\exp\left[-\frac{1}{2}\left\{\left(\frac{x}{h_x}\right)^2 + \left(\frac{y}{h_y}\right)^2\right\}\right], \qquad [4]$$

where $h_x > 0$ and $h_y > 0$ are *bandwidths* in the x and y directions, respectively. The constant k standardizes $K(x,y)$, ensuring it sums to one. The bandwidths scale the independent variables, and determine the size of the local neighborhood for smoothing; the pair (h_x, h_y) defines an elliptical neighbourhood in the xy-plane for local averaging.

It is also informative to examine the bivariate estimate of the density of length frequency samples, or the proportion of sampling at length l in week t from all fish sampled (i.e. for all lengths and weeks). The bivariate kernel density estimate of length frequency sampling is:

$$\hat{p}_{lt} = \frac{\text{[illegible]}}{\sum_{i=1}^{L}\sum_{j=1}^{T} \text{[illegible]}}, \text{ where } \text{[illegible]} = \sum_{i=1}^{L}\sum_{j=1}^{W} K(l-l_i, t-t_j)n_{ij}. \quad [5]$$

The only quantities left to specify in [5] are the bandwidths. In practice these are often chosen subjectively to capture the systematic variation in the data. Many automatic methods are available to choose bandwidths as well. We used PROC KDE in SAS for kernel smoothing. The simple

normal reference method was used to compute the bandwidths. Optimal bandwidths are selected by minimizing the approximate mean integrated square error.

Initial estimation results suggested that doubling the weekly bandwidths provided predictions that seemed more appropriate at the surface boundaries.

Length frequency samples coming from commercial sources are often a sub-sample of the total catch. The standard approach to deal with sub-sampling is to increase the sub-sampled length frequencies by the sampling fraction (s_f), the ratio of catch weight to sampled weight. (If the complete catch is sampled, the weight is simply 1.) That is, $n'_{lt} = wn_{lt}$ and then replace n_{lt} by n'_{lt} in equation [5].

For sentinel length frequencies, the entire catch is sampled. Sample weights were computed using the following length-weight relationship from Gavaris and Gavaris (1983):

$$\log(weight) = 3.0879 \log(length) - 5.2106.$$

where length has units of cm and weight is measured in kg. Commercial sample weights were compared with the predicted weights using the above weight-length relationship, and if

$$\left| \frac{\text{sample weight - predicted weight}}{\text{predicted weight}} \right| > 20\% ,$$

then the measured sample weight was replaced by the predicted sample weight. Samples having unknown or unavailable turnout weights were weighted using the sample weight (measured or predicted as appropriate).

In this application, the weighting of each length frequency is proportional to the *turnout* weight, the total catch weight, and not the sub-sampling fraction. This is incorrect, and future applications of this approach will use the sampling fraction weighting.

Results

Bivariate length frequency sampling densities were predicted for each of the 19 region-gear combinations in which length-samples were available. For example, Figure 2 contains all length frequency samples taken in the 3K inshore region from gillnet catches. The bandwidths we use for

each region-gear combination are presented in Table 2. Recall that the week bandwidths are twice the optimal bandwidths selected by PROC KDE. Larger bandwidths are indicative of sparse sampling information (for example, the estimated 3PS_WB ottertrawl bandwidths). These bandwidths are used with equation [5] to produce estimates of bivariate length frequency density as a function of length and week sampled. To save space, only selected kernel density estimates are shown in Figure 3. In these panels, the x-axis indicates length (cm), and the y-axis represents weeks (referenced to Jan.01, 1997). These plots suggest that the lengths of fish captured in the fishery are fairly consistent, but that there are temporal differences in the numbers of fish measured.

From the bivariate density, we can compute the weekly proportions at length. For a given week, we compute this density by scaling the bivariate estimate of density by the sum of the bivariate density in this week. This gives the estimate in equation [3]. The conditional densities (proportions-at-length each week) corresponding to Figure 3 are presented in Figure 4. The proportions-at-length are the results required in the tagging analyses we mentioned in the Introduction.

To assess the goodness-of-fit, we compared the observed and predicted proportions-at-length assesses the goodness-of-fit of the length frequency estimates. There are 156 weeks and 81 length groups for each region/gear type cell, so individual comparisons are impractical. Predicted density and observed sampling frequency were averaged over all weeks and sixteen 5cm length classes (the final class is a 6cm class; 115-120cm, inclusive) for each region-gear type (Figure 5). The averaged observations were weighted by turnout weight and by sample size, the weights used in our density estimation. These comparison plots are presented for all nineteen region-gear type cells (in contrast to Figures 3 and 4) to demonstrate that the estimation is adequate in all cases.

Discrepancies between observed and predicted numbers are small; however, they have a systematic pattern that is typical of smoothing. By attempting to fit the peaks of the data, the estimates of adjacent length groups are over-estimated. This pattern is a predictable consequence of smoothing, and represents the bias-variance trade-off that is involved in smoothing. We can remove the residual pattern by smoothing with smaller length bandwidths, but the cost is increased variability in our length frequency estimates. Shrinking the length bandwidths to zero would be analogous to smoothing over time only. As the magnitude of the potential bias is small, we feel the estimates in Figure 5 are reasonable.

Acknowledgements

The authors would like to thank Joe Firth and Dawn Maddock Parsons for providing data, Dan Porter and Gus Cossitt for typesetting the figures, and John Brattey for providing Figure 1.

References

Cadigan, N., and J. Brattey. 2000. Lower bounds on the exploitation of cod (*Gadus morhua*) in NAFO Divs. 3KL and Subdiv. 3Ps during 1997-1999 from tagging experiments. Can. Stock Assess. Res. Doc. 2000/073.

Gavaris, S., and Gavaris, C.A. 1983. Estimation of catch at age and its variance for groundfish stocks in the Newfoundland region. *In* Sampling commercial catches of marine fish and invertebrates. *Edited by* W.G. Doubleday and D. Rivard. Can. Spec. Publ. Fish. Aquat. Sci. 66. pp.178-182.

Hardle, W. 1990. Smoothing Techniques: With Implementation in S, New York: Springer-Verlag.

SAS Institute Inc., 1999. SAS OnlineDoc, Version 8. Cary, NC, USA.

Silverman, B.W. 1986. Density Estimation for Statistics and Data Analysis, London: Chapman and Hall.

Region ID	Unit Area(s)	Location
3K_IN	3K d,h,i	3K Inshore
3L_INN	3L a,b	3L Inshore (North)
3L_INS	3L f,j,q	3L Inshore (South)
3PS_PB	3Ps c	Placentia Bay
3PS_WB	3Ps a,b,c,e	3Ps West of Burin Peninsula
3PS_OF	3Ps f,h	3Ps Offshore

Table 1: Region Definitions.

Region	Gear	h_wk	h_L
3K IN	Gillnet	12.19	1.11
3K IN	Handline	14.17	2.09
3K IN	Linetrawl	13.66	2.39
3L INN	Gillnet	15.09	1.28
3L INN	Handline	19.17	2.70
3L INN	Linetrawl	25.12	3.21
3L INS	Gillnet	12.25	1.05
3L INS	Handline	15.69	2.02
3L INS	Linetrawl	18.32	2.69
3PS OF	Gillnet	18.60	2.86
3PS OF	Linetrawl	22.00	4.72
3PS OF	Ottertrawl	14.62	2.64
3PS PB	Gillnet	15.86	1.54
3PS PB	Handline	26.35	2.77
3PS PB	Linetrawl	22.24	2.49
3PS WB	Gillnet	16.75	2.62
3PS WB	Handline	14.17	4.98
3PS WB	Linetrawl	15.26	1.90
3PS WB	Ottertrawl	35.67	5.50

Table 2: Estimated Bandwidths.

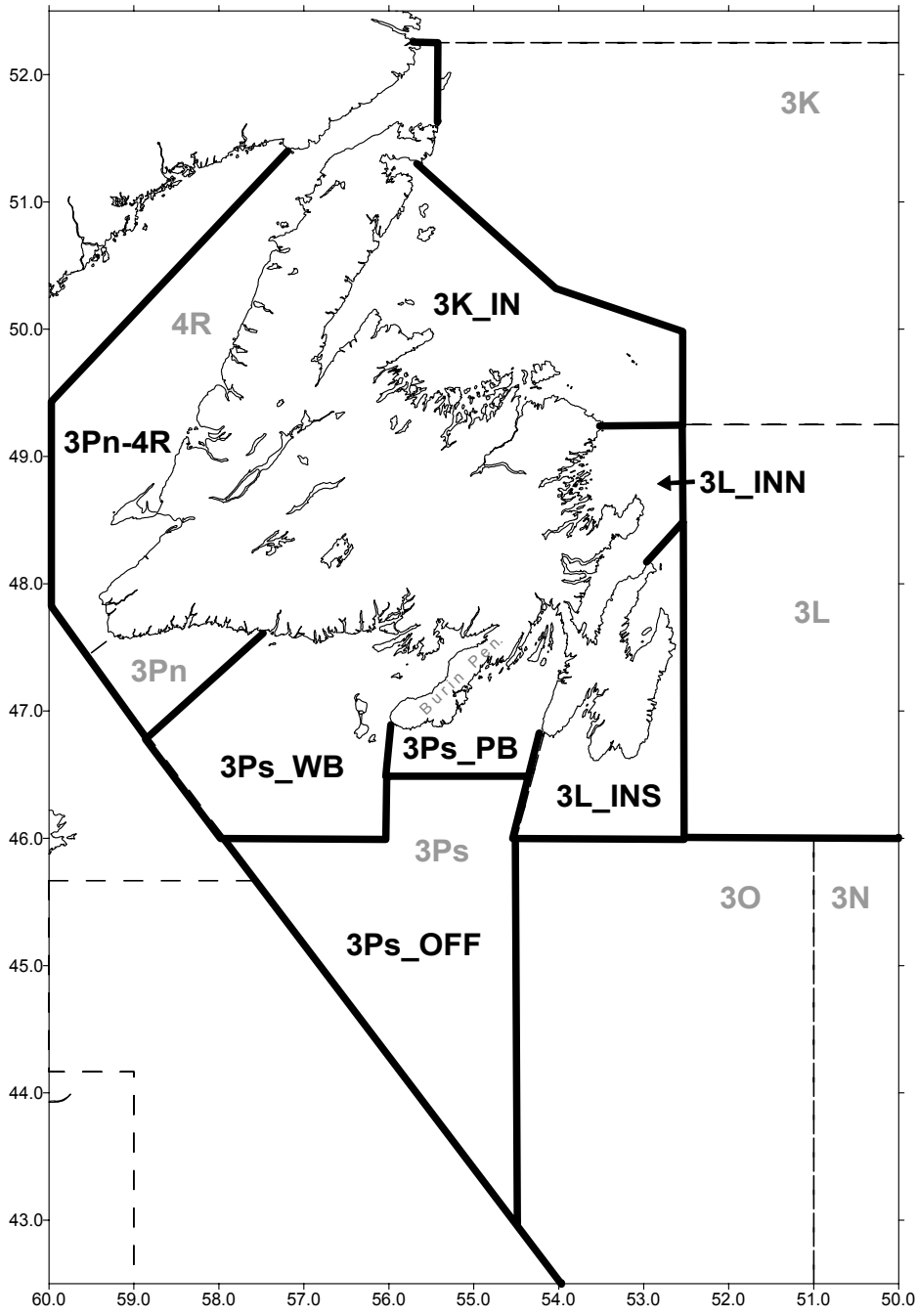


Figure 1. NAFO Divisions and boundaries of sub-areas used in the analysis of cod tagging data: 3K_IN=3K Inshore, 3L_INN=3L Inshore North, 3L_INS=3L Inshore South, 3PS_PB=3Ps (Placentia Bay), 3PS_OF=3Ps offshore, 3PS_WB=3Ps West of the Burin Peninsula.

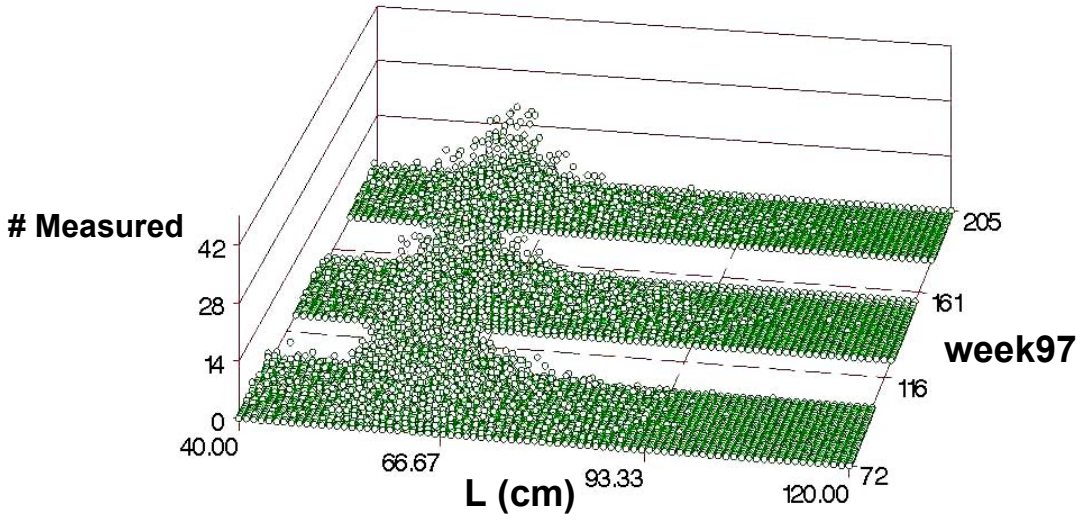


Figure 2: Length-frequency sampling from 3K (inshore) gillnet catches. Weeks (week97 axis) are referenced to Jan. 01, 1997.

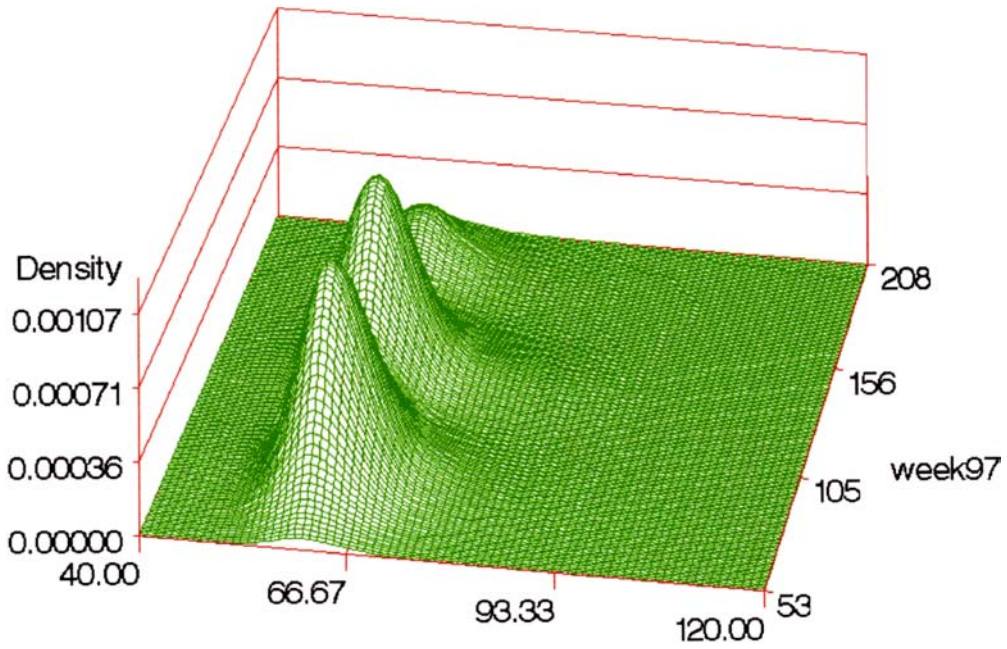


Figure 3a: Estimated Bivariate Kernel Density.

Gear: Gillnet, Region: 3K_IN.

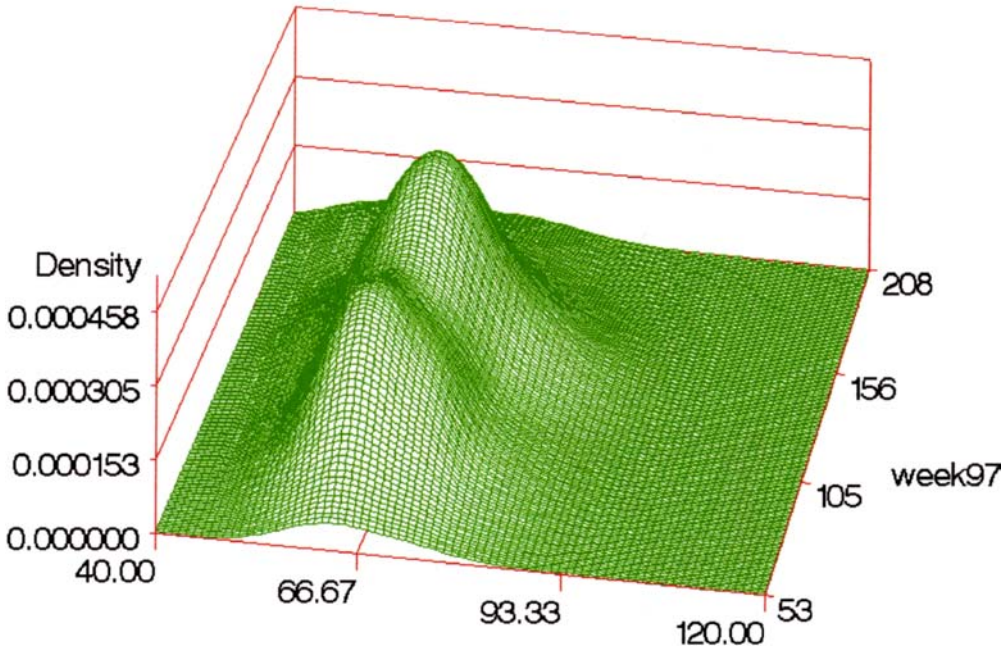


Figure 3b: Estimated Bivariate Kernel Density.

Gear: Handline, Region: 3L_INN.

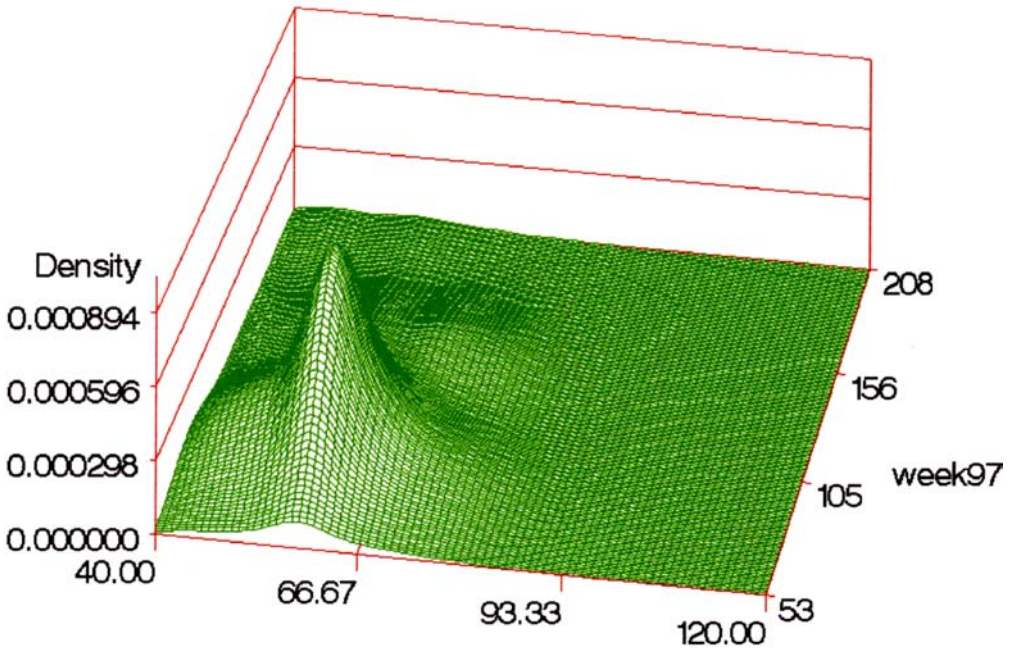


Figure 3c: Estimated Bivariate Kernel Density.

Gear: Linetrawl, Region: 3L_INS.

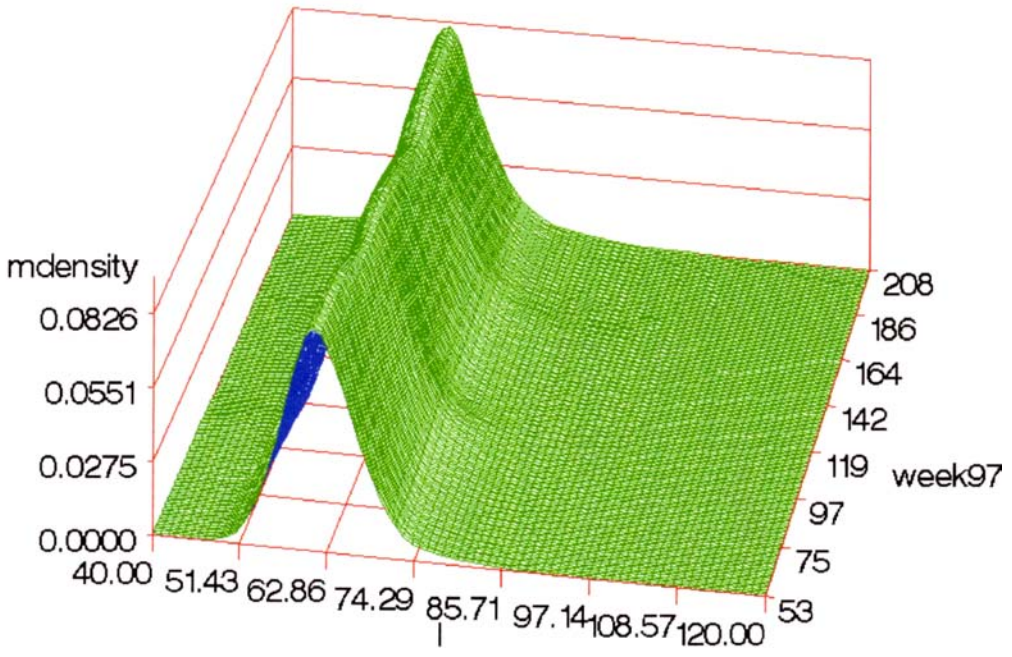


Figure 4a: Conditional Weekly Density Estimate.

Gear: Gillnet, Region: 3K_IN.

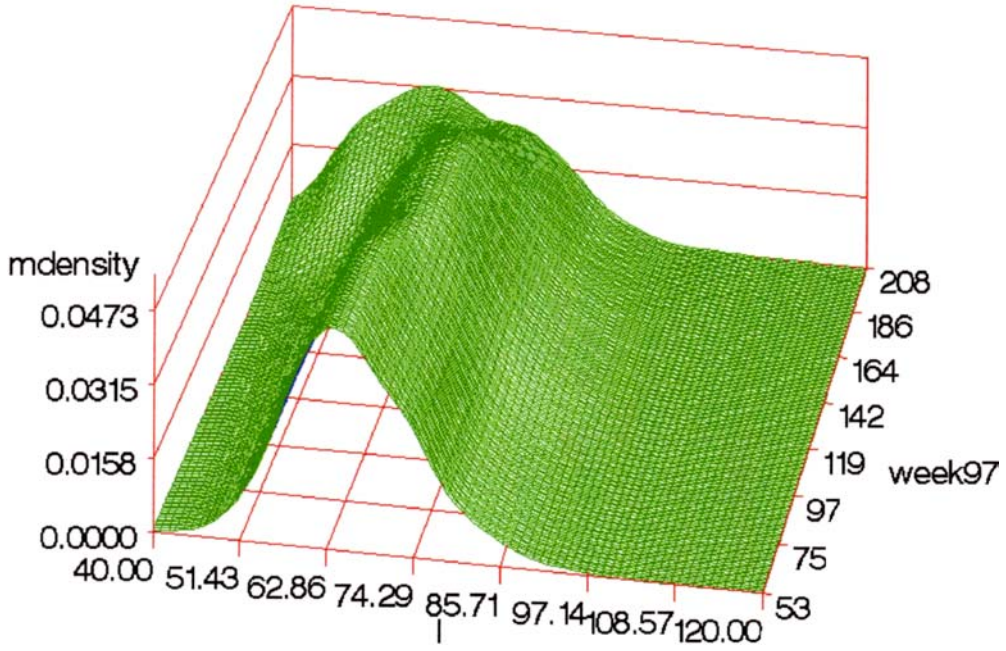


Figure 4b: Conditional Weekly Density Estimate.

Gear: Handline, Region: 3L_INN.

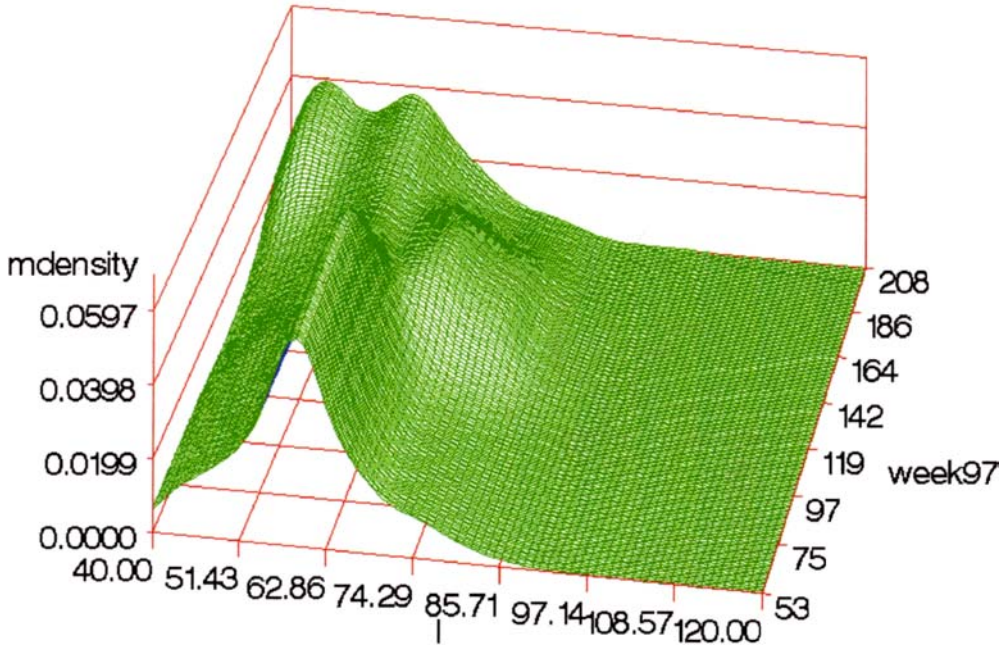


Figure 4c: Conditional Weekly Density Estimate.

Gear: Linetrawl, Region: 3L_INS.

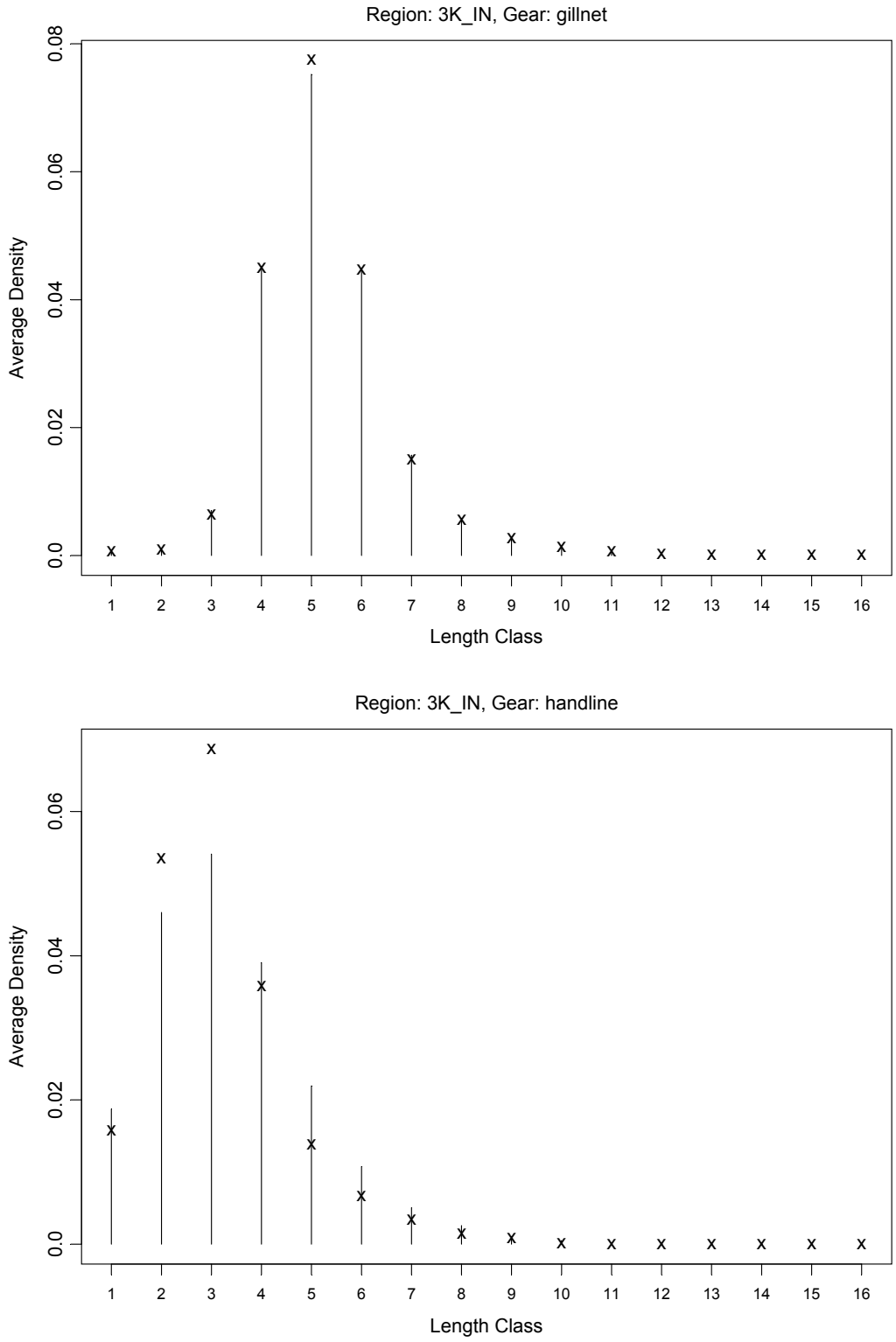
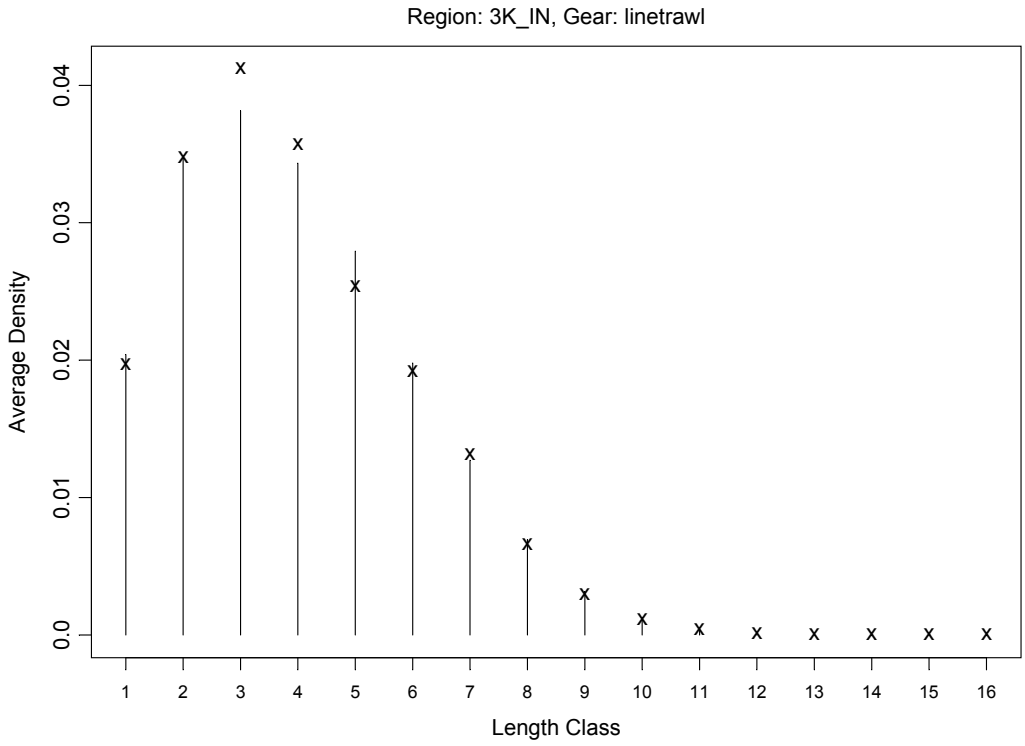


Figure 5: Comparison of predicted (vertical lines) and observed ("x") length frequency densities, averaged by length class (see text). The average observed proportions are weighted by turnout weight and sample size, the weights used in estimating the densities.



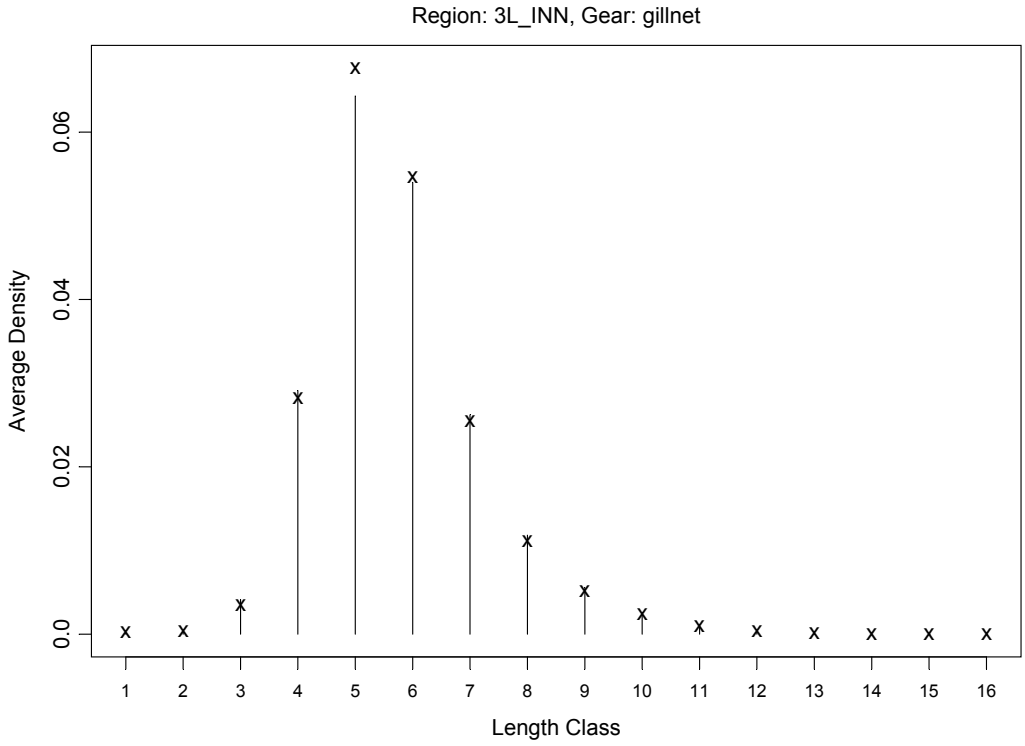
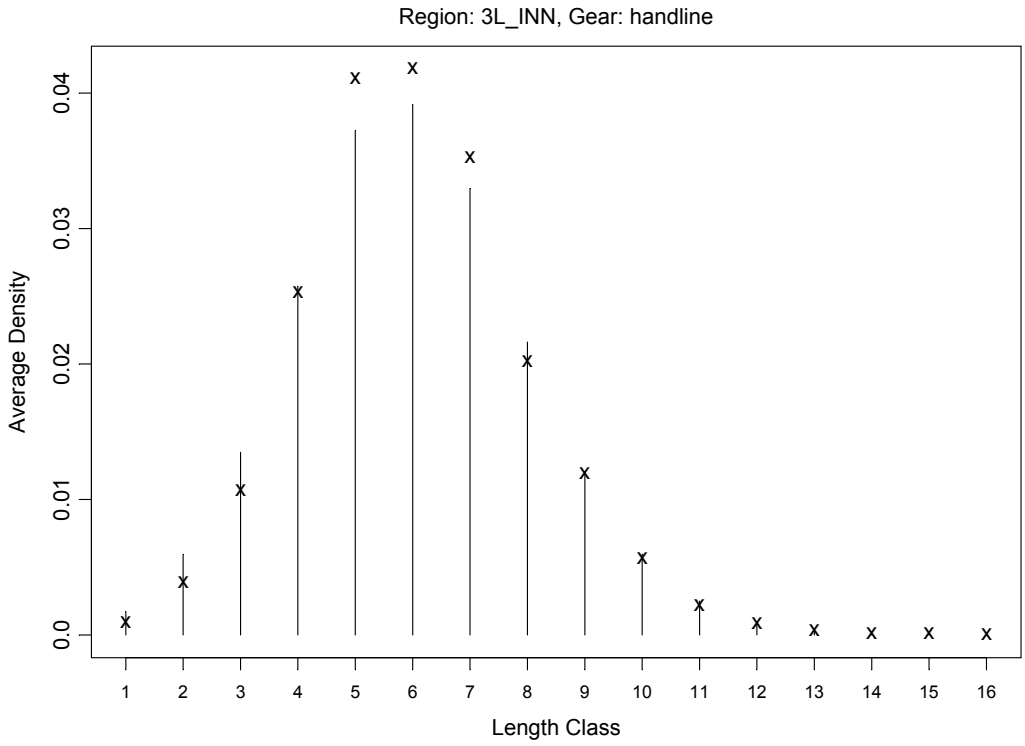


Figure 5 (cont.)



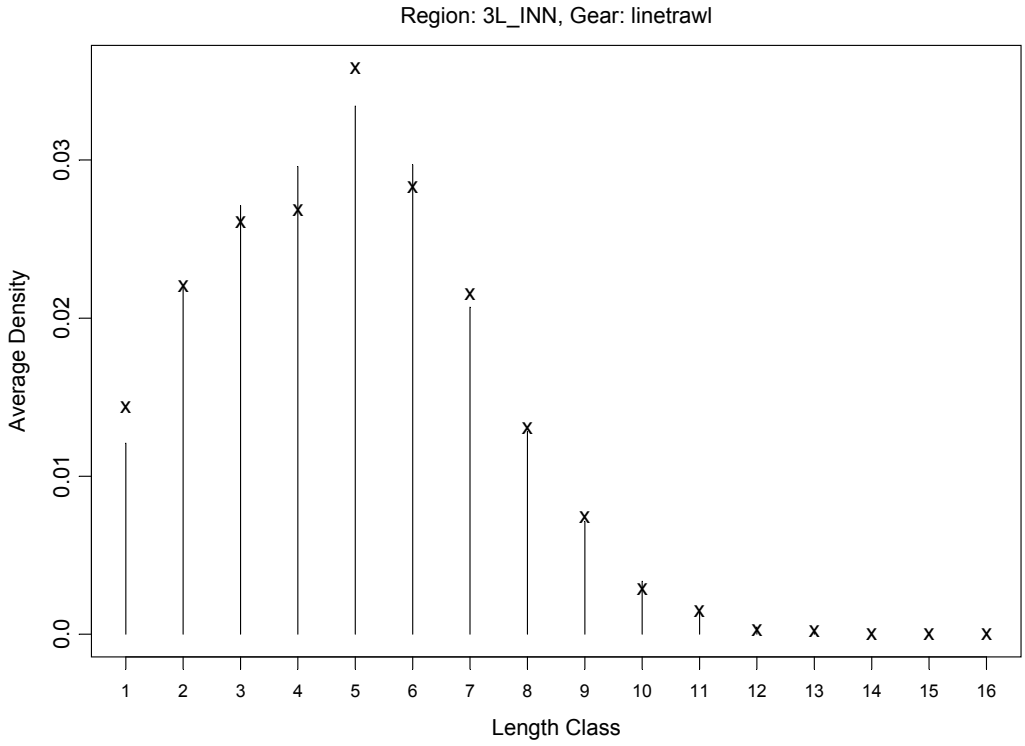
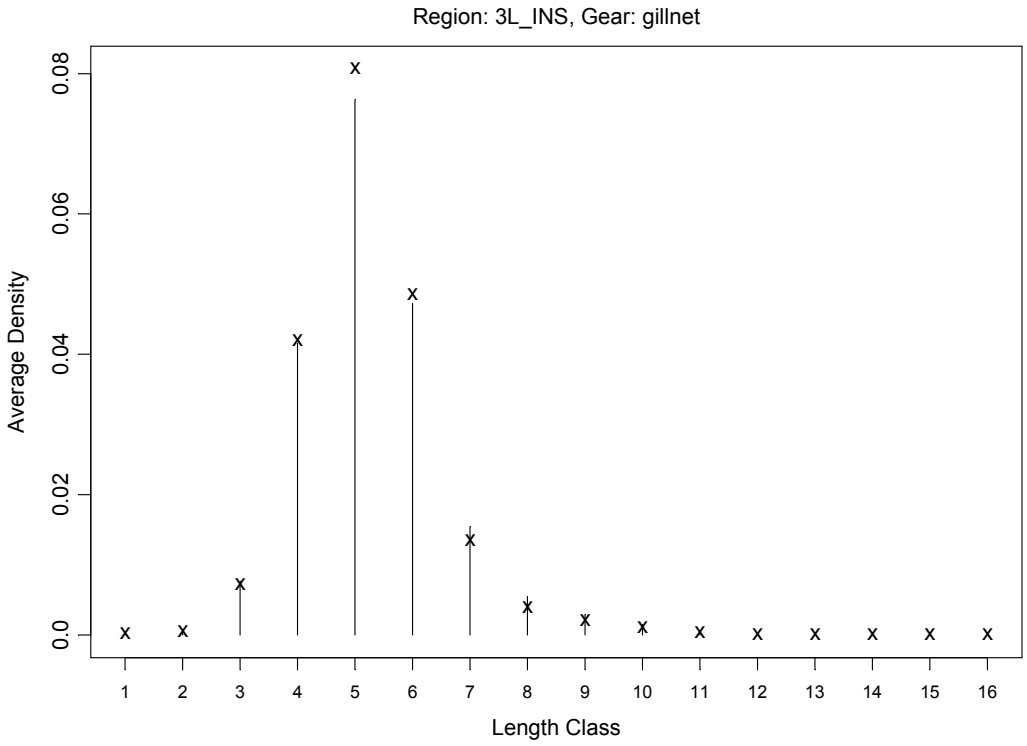


Figure 5 (cont.)



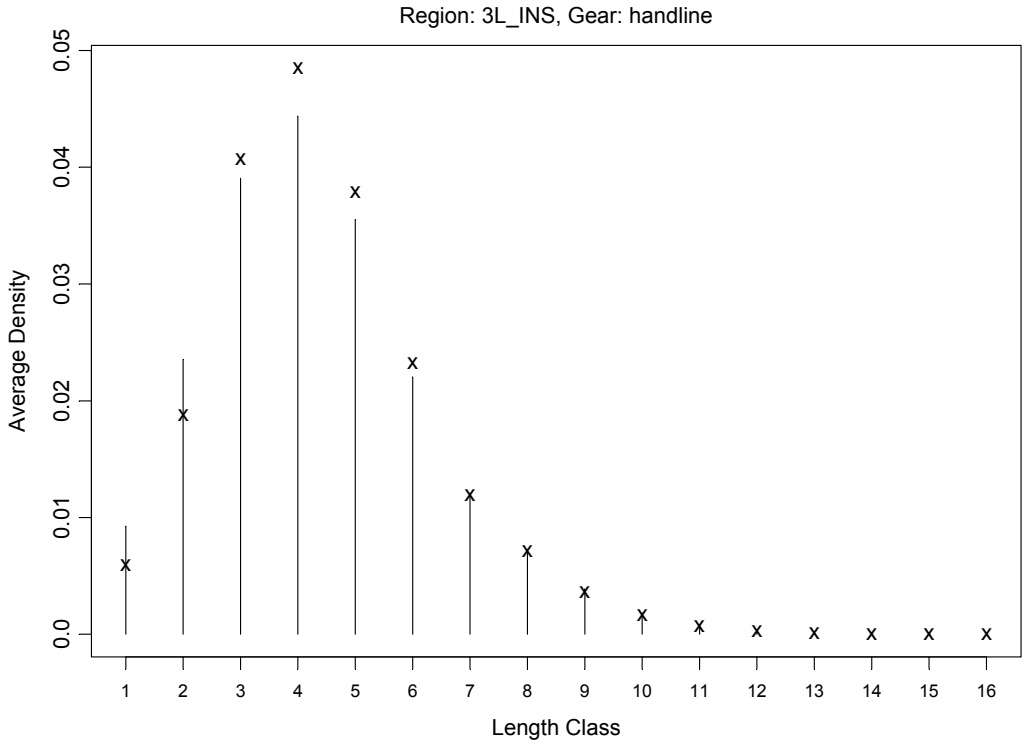
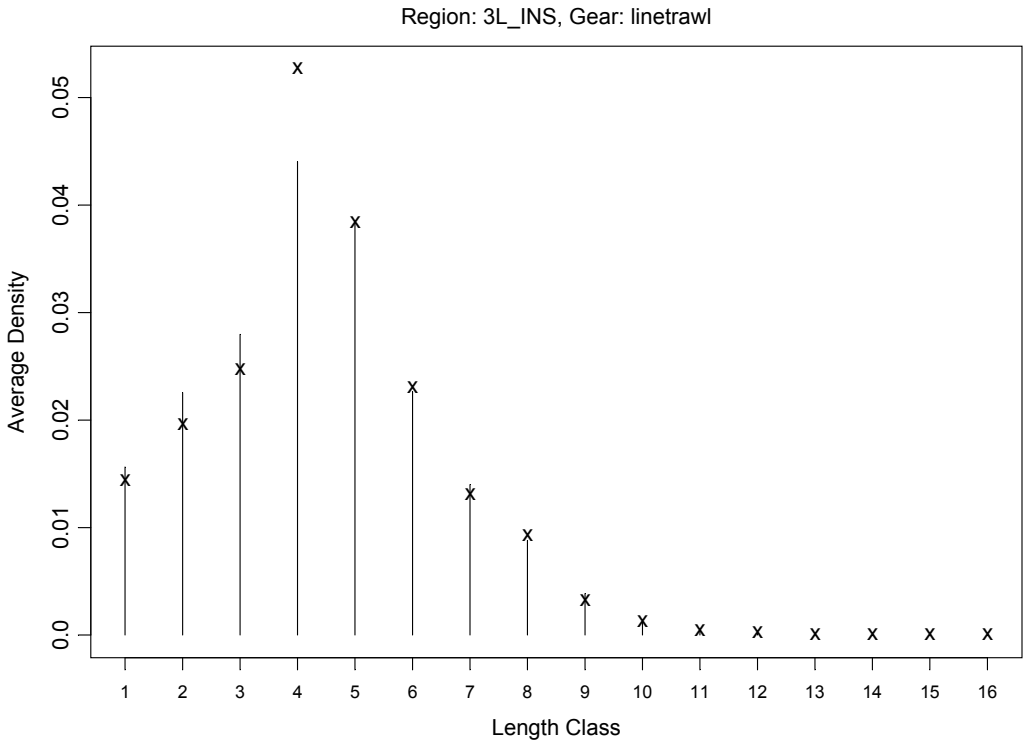


Figure 5 (cont.)



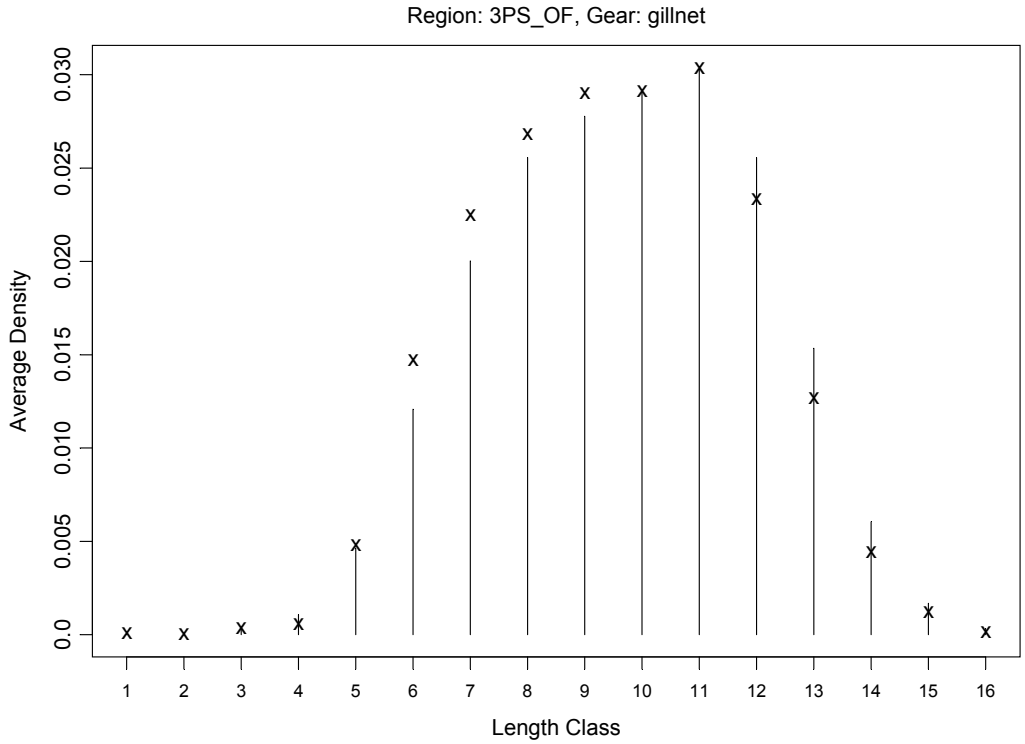
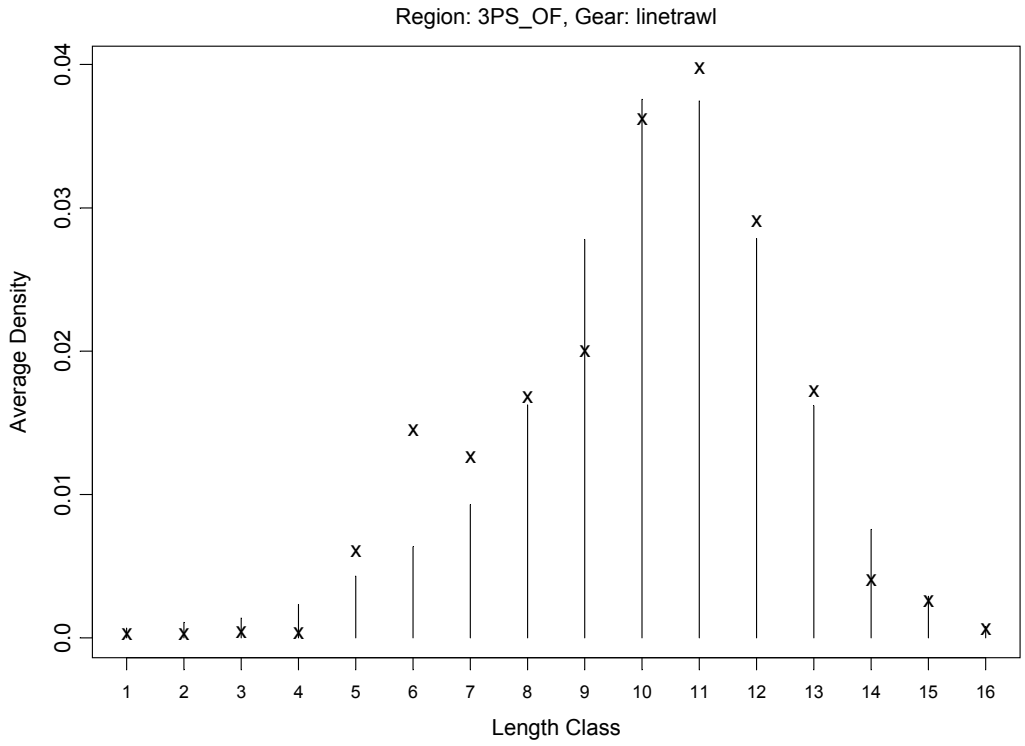


Figure 5 (cont.)



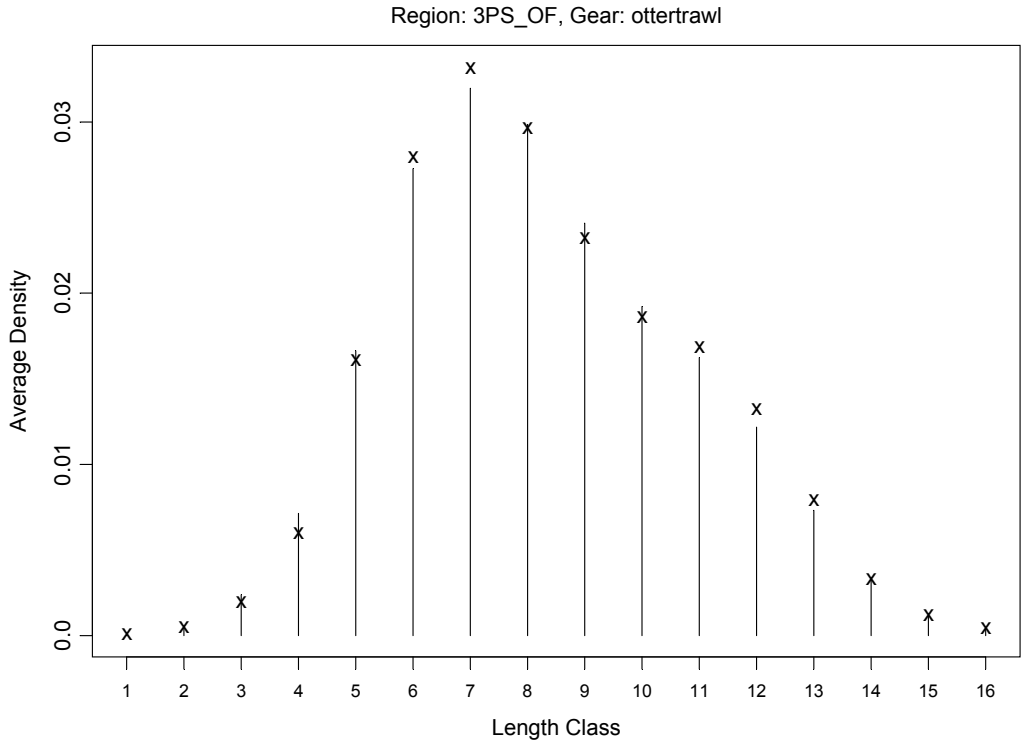
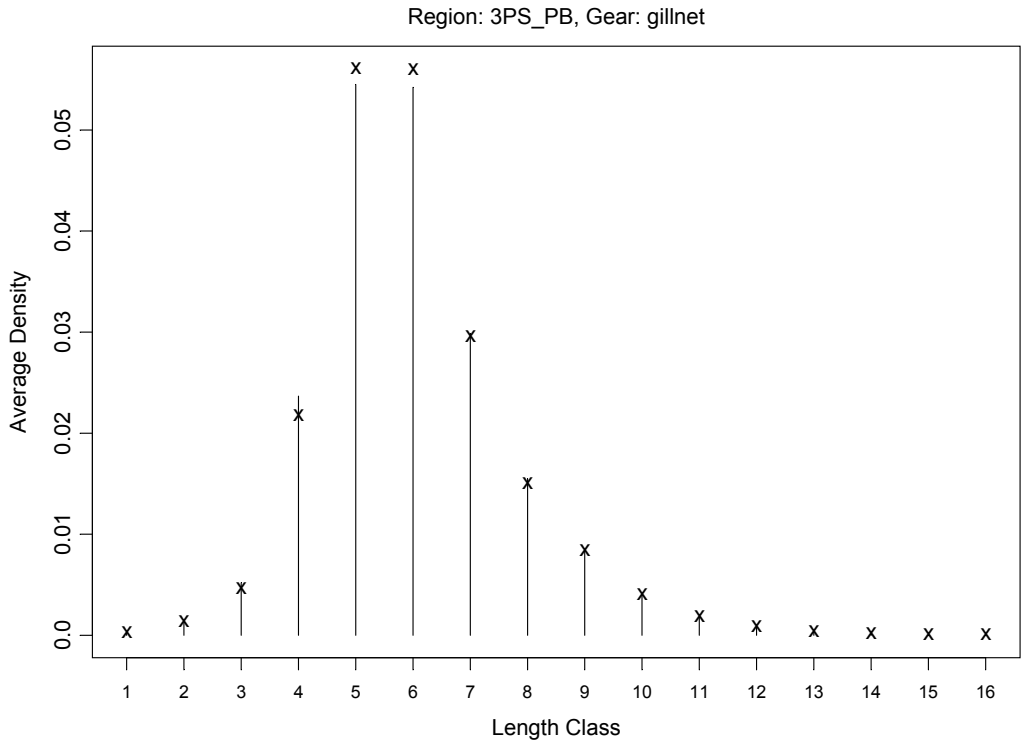


Figure 5 (cont.)



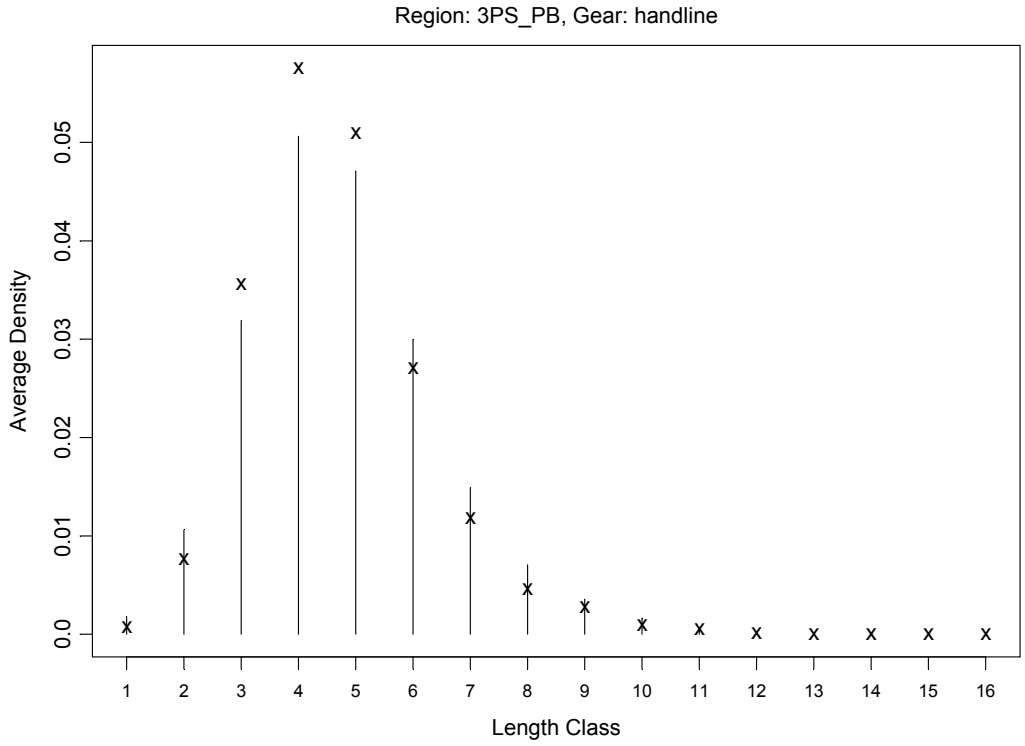
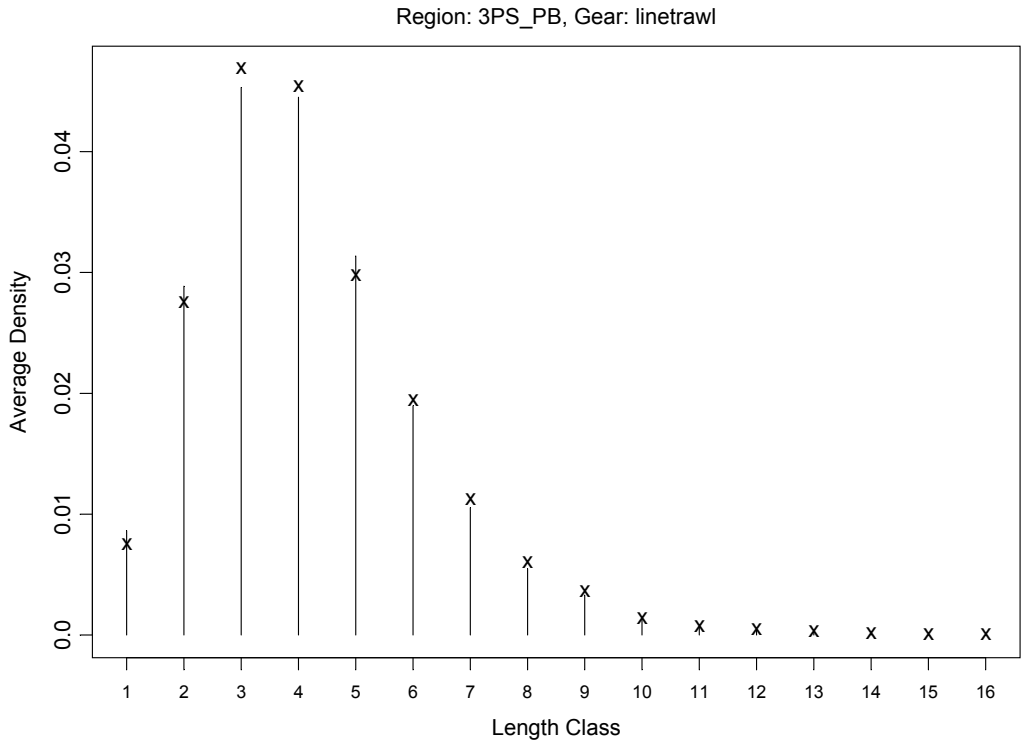


Figure 5 (cont.)



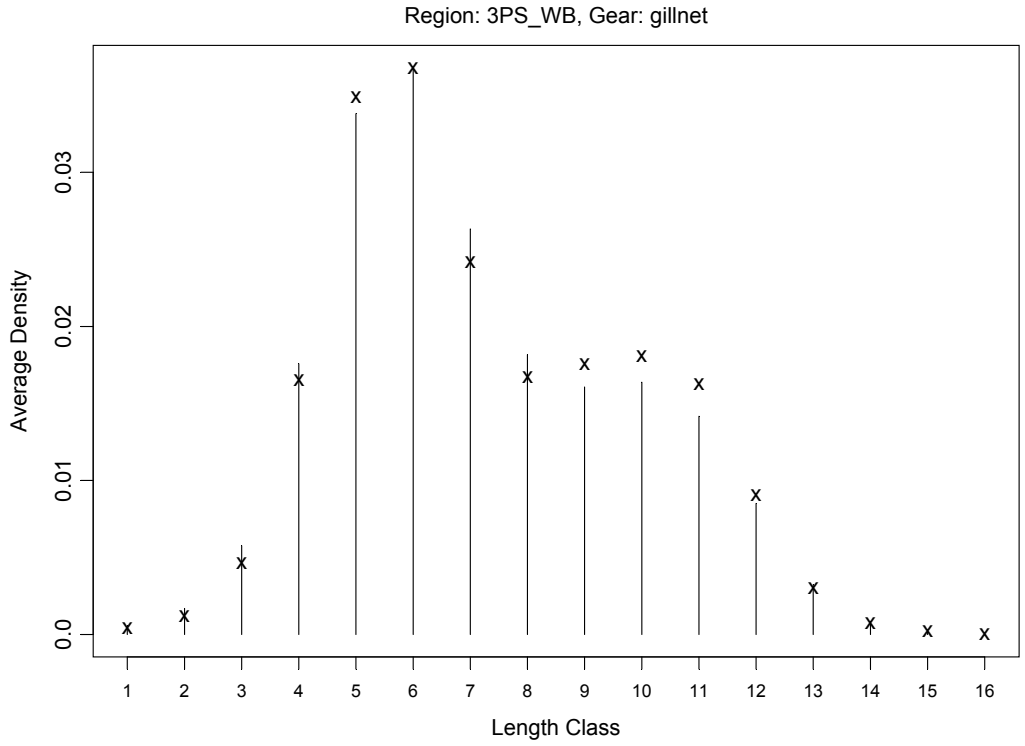
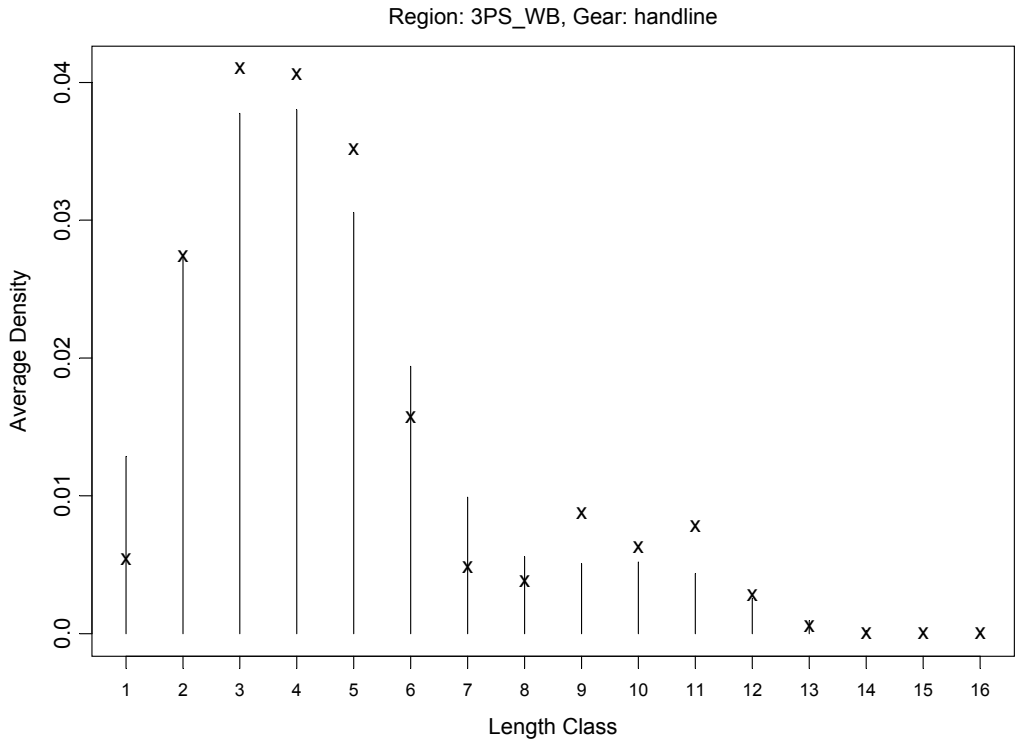


Figure 5 (cont.)



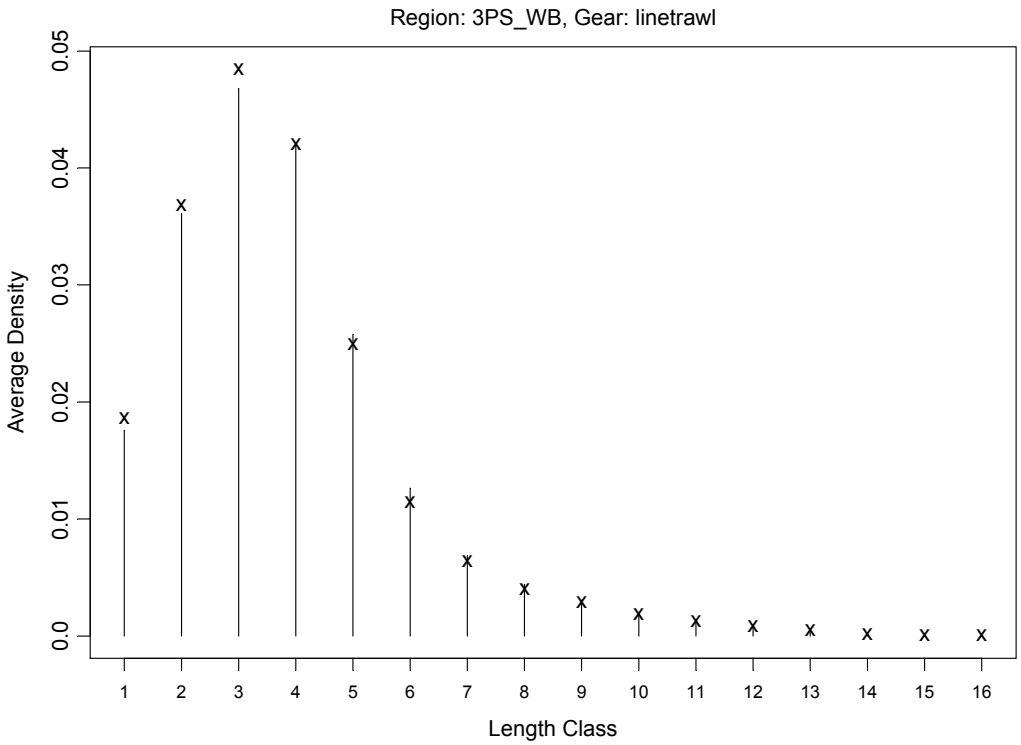


Figure 5 (cont.)

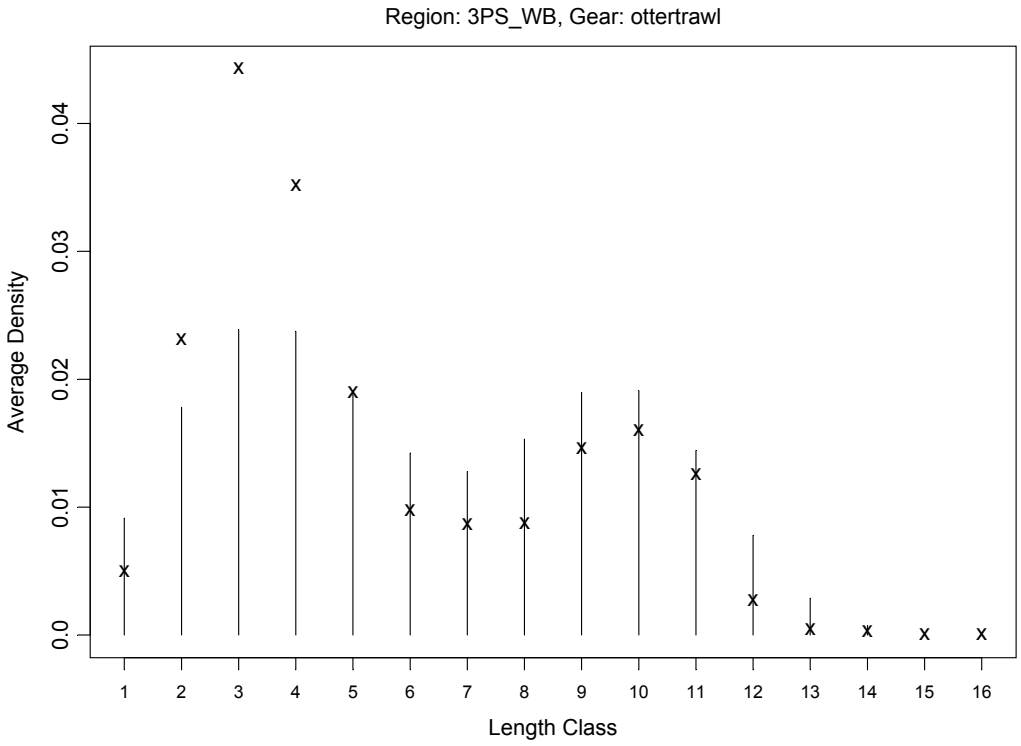


Figure 5 (cont.)