

Technical Report

Imputation for Missing Values in the May Breeding Waterfowl and Habitat Survey

Clinton T. Moore
Statistician

12 May 1995

U.S. Fish and Wildlife Service
Office of Migratory Bird Management

The May Breeding Waterfowl and Habitat Survey has provided a continuous series of population data since 1955 for only a minority of the current 52 survey strata (Table 1). Data exist back to 1955 for strata 21-40 and for stratum 50; for the remaining strata, data are available beginning 1956-1965. Strata 7 and 50 are characterized by long periods (≥ 12 years) during which no survey was conducted. In a few other strata, missing data occur in isolation. For certain time periods in some strata, data are available for only a single transect, limiting their usefulness in the calculation of variance estimates.

Survey data provide the basis for the annual estimation of breeding waterfowl population size and prairie pond numbers (Smith 1995). The estimates are important to waterfowl managers both within and outside of the Office of Migratory Bird Management. An estimate of population size for years in which no survey was conducted may be valuable in some management endeavors.

This report summarizes an approach I used for the imputation of missing values in the database of population estimates. The method predicts values for missed surveys by smoothing the time series of estimates for each species-stratum combination. A variance is calculated as the sum of a prediction variance component and a sampling variance component. In a very few cases where a suspected outlying population estimate far exceeded the next largest estimate in the data series, I replaced the observation with an imputed value. The imputed values are flagged in the database, so that selective analysis of the database may be performed.

Methods

Estimates of visibility-corrected population size for 23 species and for ponds were segregated by survey stratum. In most (48 of 52) strata, the longest period of missed surveys was ≤ 4 years; in strata 7, 41, 42, and 50, the longest was ≥ 10 years (Table 1). Because long data gaps such as those observed in the latter 4 strata reduce the performance and reliability of nonparametric smoothing procedures, I selected data from neighboring strata to augment the sparse data sets. I augmented data for stratum 7 with data from strata 1-2, for strata 41 and 42 from strata 43-44, and for stratum 50 from stratum 24 (Fig. 1).

The smoothing method I employed was locally-weighted linear regression (loess), with no attempt to remove the influence of outliers through additional robustness iterations (Cleveland 1979). Loess is a particularly advantageous smoothing procedure for this application because it is less prone than other smoothing methods to suffer bias near the boundaries of a data series (Hastie and Loader 1993). Cleveland (1979) also provided a means of calculating approximate variances for smoothed estimates (Fig. 2). For augmented strata, I constructed a

partial linear model by incorporating a stratum indicator variable in the loess smooth (Härdle 1990). In such a model, the loess smooth becomes a series of parallel curves, one curve for each stratum (Fig. 3). In this way, neighboring strata help to determine the shape of the curve, but each stratum maintains its relative difference in mean population size.

A tension parameter f determines the degree of curvature in the smooth. Values of f near 0.0 place emphasis on lowering prediction bias at the cost of increasing prediction variance, and as such, produce curves that follow the data somewhat closely but have generally wide confidence bands. Alternatively, values of f near 1.0 place more emphasis on smaller variance than on smaller bias: these curves have narrow confidence bands but lack flexibility to closely follow the data.

The optimal degree of trade-off between precision and bias likely varies among species and strata. For example, a large value of f may be unreasonable for an abundant species in an intensively-surveyed stratum. Likewise, too low a tension parameter is inappropriate for a species sampled in a low-precision setting. After eliminating suspected outliers, I estimated f for each species-stratum series (including augmented series) by minimizing cross-validated predicted mean squared error (Cleveland 1979). I grouped strata into survey crew areas and obtained crew area means of \hat{f} (Table 2).

Mean \hat{f} varied between 0.087-1.000 (Table 2). Predicted values from series smoothed with very low tension tend to be highly variable, especially at either end of a series of data, because a local trend is estimated on essentially few data points. Most missing data values in the May survey occurred during the first few years of the program, and imputation of these values relies on extrapolation from the remainder of the series. Smooths computed with very low tension may provide extrapolated values that follow abruptly changing local trends which are unsupported by the data. Though low-tension smooths may be warranted for purposes of modeling patterns in data, such smooths may be unreasonable for extrapolation purposes. For this reason, I arbitrarily selected 0.5 as a minimum value for smoothing any series.

I smoothed values for each series (excluding suspected outliers) using $f = \max(0.5, \text{mean } \hat{f})$ appropriate for the species and crew area containing the stratum. Missing values of population size corresponding to non-surveyed years were replaced by predicted values from the loess smooth. For each series, I averaged sampling variance estimates for survey years based on ≥ 2 transects. I added the average sampling variance to prediction variance for each predicted value, and I imputed the sum for the variance estimate in years for which the survey was not conducted or was conducted on only a single transect. Imputation of population size and variance was also performed for suspected outliers (Table 3). Finally, I added a code field to the database to indicate imputation status: 0

(no imputation), 1 (stratum not surveyed, population size and variance imputed), 2 (suspected outlier, population size and variance imputed), and 3 (single-transect survey, variance imputed). The code field provides a means of removing imputed values from the data prior to its analysis.

I thank J.R. Sauer, W.L. Kendall, G.W. Smith, and J.R. Kelley, Jr. for their insightful comments in the preparation of this report.

References

- Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* 74:829-836.
- Härdle, W. 1990. Applied nonparametric regression. Cambridge Univ. Press, Cambridge, England. 333pp.
- Hastie, T., and C. Loader. 1993. Local regression: automatic kernel carpentry. *Stat. Sci.* 8:120-143.
- Smith, G. W. 1995. Review of the aerial waterfowl breeding ground and habitat surveys in North America. *Natl. Biol. Serv. Biol. Rep. In Press.*

..
..
..

Table 1. (cont.)

¹Strata 26-49 and 75-76 are grouped into 6 operational crew areas as designated in the standard operating procedures for the survey (Standard operation procedures for aerial waterfowl breeding ground population and habitat surveys, Canadian Wildlife Service and U.S. Fish and Wildlife Service, 1977). Other crew areas designed for purposes of this study are: A (Alaska-Yukon, strata 1-12), N (Northwest Territories-northern Alberta, strata 13-20, 77), and S (northern Saskatchewan-northern Manitoba-western Ontario, strata 21-25, 50).

Table 2. Estimated smoothing parameter (\hat{h}) for species-stratum loess smooths averaged over strata within crew area.

Species	Crew Area (<i>N</i> strata)								
	1 (4)	2 (4)	3 (7)	4 (5)	5 (4)	6 (2)	S (6)	N (8)	A (12)
1290 - COME	0.520	0.458	0.442	0.524	0.590	0.753	0.662	0.324	0.676
1320 - MALL	0.087	0.120	0.445	0.600	0.638	0.917	0.421	0.595	0.594
1330 - ABDU	0.705	0.752	0.802	1.000	.	.	0.461	0.997	.
1350 - GADW	0.130	0.387	0.555	0.422	0.303	0.260	0.332	0.498	0.687
1370 - AMWI	0.109	0.187	0.747	0.424	0.335	0.571	0.320	0.626	0.526
1390 - AGWT	0.239	0.221	0.353	0.536	0.569	0.682	0.505	0.293	0.606
1400 - BWTE	0.222	0.237	0.625	0.483	0.698	1.000	0.658	0.478	0.537
1420 - NSHO	0.239	0.189	0.543	0.544	0.655	0.856	0.694	0.801	0.693
1430 - NOPI	0.154	0.191	0.455	0.677	0.307	1.000	0.785	0.677	0.581
1460 - REDH	0.370	0.319	0.593	0.501	0.348	0.605	0.460	0.573	0.706
1470 - CANV	0.834	0.461	0.677	0.851	0.368	0.699	0.661	0.681	0.751
1490 - LESC	0.167	0.257	0.265	0.717	0.583	0.356	0.721	0.451	0.377
1500 - RNDU	0.514	0.324	0.128	0.598	0.599	0.814	0.766	0.726	0.435
1510 - COGO	0.734	0.620	0.578	0.495	0.796	0.648	0.846	0.262	0.597
1530 - BUFF	0.906	0.217	0.534	0.819	0.792	0.790	0.511	0.467	0.507
1540 - OLDS	0.831	.	0.769	.	.	0.731	0.981	0.705	0.361
1600 - COEI	0.419	0.634
1630 - BLSC	0.886	0.655	0.842	.	0.939	0.840	0.526	0.440	0.468
1670 - RUDU	0.755	0.747	0.585	0.325	0.748	0.953	0.820	0.559	0.624
1710 - GWFG	0.436	.	0.693	0.660	0.515	.	0.615	0.439	0.619
1720 - CAGO	0.429	0.442	0.437	0.426	0.356	0.333	0.583	0.441	0.560
1800 - TUSW	0.625	0.714	0.576	1.000	0.767	0.630	0.916	0.411	0.369
2210 - AMCO	0.494	0.425	0.506	0.502	0.357	0.668	0.656	0.601	0.349
Unadj Ponds	0.700	0.278	0.393	0.795	0.577	1.000	.	.	.
Visib-adj Ponds	0.494	0.597	0.517	0.851	0.817	0.881	.	.	.

Table 3. Value of suspected outliers, maximum of 10 nearest (temporally) values, and imputed value and standard error.

Species	Year	Stratum	Suspected Outlier	Nearest-10 maximum	Imputed value	
					Estimate	SE
1370 - AMWI	1958	34	570,706	223,272	125,684	45,086
1370 - AMWI	1958	35	195,323	82,974	43,815	24,510
1370 - AMWI	1958	40	190,147	73,662	42,097	12,197
1490 - LESC	1955	25	248,920	94,063	54,022	48,969

Figure 1. Strata for the May Breeding Waterfowl and Habitat Survey.

Figure 2. Loess smooth (tension $f = 0.594$) applied to population estimates of mallards in stratum 3. Curves above and below the smooth represent ± 1 square root of prediction variance (large dashes) and ± 1 square root of prediction plus average sampling variances (small dashes). Filled symbols represent imputed values for years with missed surveys.

Figure 3. Loess smooth (tension $f = 0.500$) used in a partial linear model for population estimates of mallards in strata 24 (dashed line, Δ) and 50 (solid line, \odot). The model forces the 2 data series to share the same curvature but not necessarily the same mean population size. Filled symbols represent imputed values for years with missed surveys in stratum 50.

Figure 1



Figure 2

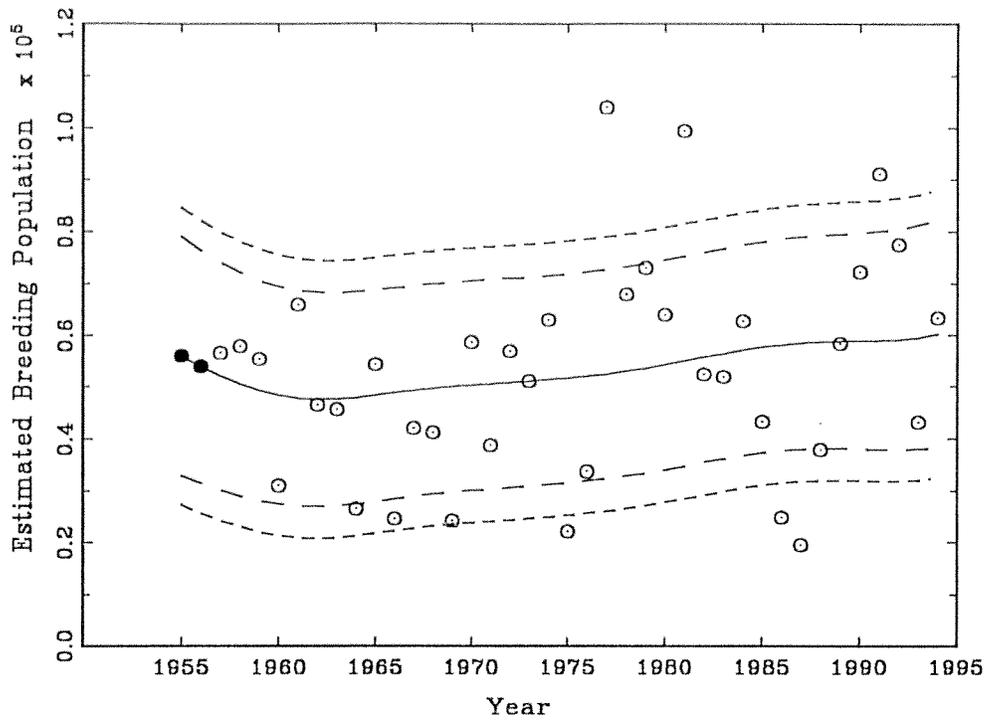


Figure 3

