Statistical Review of Osprey Monitoring in the Labrador and Eastern Quebec Low-level Training Area

Prepared for:

Institute for Environmental Monitoring and Research
P.O. Box 1859, Stn. "B"
114 Hamilton River Road
Happy Valley-Goose Bay, Labrador
A0P 1E0

By:

Jacques Allard, Ph.D.
Atlantic Statistical Analysis Inc.
21 Colburne Street
Moncton N.B. E1C 8Z9

Tel.: 506-866-2655
E-Mail: mathstat@nbnet.nb.ca

2008 09 19

1 INTRODUCTION 4

2 HISTORY OF THE OSPREY MONITORING PROGRAM 5

2.1 Information reviewed 5

2.2 Chronology of osprey monitoring in the LLTA 5

3 DESCRIPTION OF THE 1999-2007 MONITORING PROGRAM 7

3.1 Sampling method 7

3.2 Outcomes observed 7
3.2.1 Nest Occupancy (Nest Activity) 7
3.2.2 Nesting Success and Nesting Productivity 7

3.3 Statistical analysis 8
3.3.1 Use of several statistical methods throughout the years 8
3.3.2 Significance threshold 8
3.3.3 Confidence intervals 8

4 REVIEW OF THE 1999-2007 MONITORING PROGRAM 9

4.1 Observational study 9

4.2 The interpretation of the Nest Occupancy Rate 9

4.3 Sampling method 12

4.4 Alternative hypothesis 13

4.5 Statistical significance threshold and power 13
4.5.1 Type I error (false positives) 13
4.5.2 Type II error (false negatives) and power 14

4.6 Multiple testing 15

4.7 Statistical tests for single year comparisons 15
4.7.1 Impact of the quota sampling approach on Nest Occupancy Rate estimate 15
4.7.2 Test on Nest Occupancy Rate and Nesting Success 16
4.7.3 Discrete variable: Nesting Productivity 18

5 MULTI-YEAR MODELS 23

5.1 Nesting Success model based on presence/absence of LLT activity 23

5.2 Nesting Success model based on the intensity of LLT activity or exposure 24

5.3 Nesting Productivity model based on presence/absence of LLT activity 25

5.4 Nesting Productivity model based on intensity of LLT activity 25

6 ANALYSIS OF THE 1999-2007 RESULTS 27

6.1 Nesting Success 27

6.2 Nesting Productivity 28

6.3 Nest Occupancy Rate 30

7 CONCLUSION 32

7.1 Interpretation of the current results 32

7.2 Special circumstances 32

7.3 Periodical monitoring 32

7.4 Database organization 33

7.5 Statistical tests 33
7.5.1 Single-year analyses 33
7.5.2 Multi-year analyses and intensity impact 33
7.5.3 Probability of type I and type II error 33
7.5.4 Confidence intervals 34

8 ACKNOWLEDGMENTS 35

1 Introduction

The mandate of the Institute for Environmental Monitoring and Research (IEMR) for the low-level training area (LLTA) in Labrador and Eastern Quebec includes the responsibility “to provide independent verification of environmental effects” of the low-level training (LLT). Within that mandate, the Institute has been carrying out monitoring of osprey reproduction activities for several years.

The present document includes a review of the statistical methodology used up to now in the Osprey Monitoring Program and a statistical review of the accumulated results. It also contains recommendations on the statistical methodology and on the usefulness of continuing the program.

2 History of the Osprey Monitoring Program

2.1 Information reviewed

The following information was used:

- An overview of the history of osprey nest monitoring in the LLTA.
- A database of osprey nest observations from 1994 to 2002 (Nest_outcomes_1994-2002.xls).
- A larger database of osprey nest observations from 1999 to 2001 (including the observations in the previous item) (Nest_outcomes_1999-2001.xls).
- A database of nest observations for various species including the osprey, in 1998 (Osprey_nests_98_mod.xls).
- Annual reports from the program “Osprey Monitoring in the Low-Level Training Area of Labrador” from 1999 to 2007.
- A draft report from a Osprey Workshop held in 2003 in St.John`s, N.L., titled “Post-Mitigation Monitoring of Ospreys in the Low-Level Training Area of Labrador”.
- A description of the various monitoring programs, supplied by the IEMR for the reviewer.

2.2 Chronology of osprey monitoring in the LLTA

The IEMR reports several studies of osprey nesting. Table 1 presents a summary of the studies and the results as described in various documents submitted to the reviewer.

Table 1. History of osprey monitoring related to LLT activity

Period	Description	Results
1991	Department of National Defence (DND) initiates a program of annual osprey nest monitoring to identify active nests in a mitigation program which establishes a 2.5 nm exclusion zone around nests.	
1994	Nesting success and survival is compared between 5 nests inside and 5 nests outside the 2.5 nm exclusion zone.	No statistically significant difference is found; the probability of type II error was not reported.
1995, 1996	Nesting success and survival is estimated again.	Results are not presented.
1996-1998	Several studies of nesting success as a function of flight frequency or exclusion zone size.	Results are not presented.
1999-2007	Systematic survey of nesting activity, nesting success and nest productivity in a subarea of the LLTA and in a corresponding area outside the LLTA.	**The results of this survey are the main focus of the present review and are described and discussed extensively below.**

The IEMR has also conducted a workshop on the Osprey Monitoring Program in October 2003.

3 Description of the 1999-2007 Monitoring Program

The 1999-2007 Monitoring Program appears to be the most extensive program among those reported. We now describe this program in detail.

3.1 Sampling method

The sampling is carried out by helicopter. The sampling method is described as follows in the LLTA Osprey Monitoring Program 2007 Report:

“Prior to departure, a route was chosen that maximized nests encountered (by chance) while minimizing travel distance/time (and associated cost). The goal was to locate a sample of 30 active nests in each Study Area. Note that this sample may have included new nests located during the surveys. As such, nests used in 2007 may or may not have been included in previous monitoring years.”

Further details were provided to the reviewer by Mr. Perry Trimper, the project manager.

The sampling route targets sites known to have held a nest in previous years and starts with such sites closest to Goose Bay. Any new previously unknown nest site encountered is included in the current year sample and added to the database for later years. Sampling proceeds until 30 active nests have been found, expanding beyond the sites in the database if necessary.

3.2 Outcomes observed

Three outcomes have been analysed in the LLTA Osprey Monitoring Program:

- Nest activity, defined as presence of signs of breeding activity at a nest. We prefer to use the expression “Nest occupancy”.
- Nesting success, defined for active nests only, as the presence of at least one young fledged.
- Nesting productivity, defined for active nests only, as the number of young fledged.

3.2.1 Nest Occupancy (Nest Activity)

For an available nest, nest occupancy is defined by the presence/absence of signs of breeding activity at the nest. For each nest, nest occupancy is a dichotomous (binary) variable. The nest occupancy rate (nest activity rate) is reported.

3.2.2 Nesting Success and Nesting Productivity

The two outcomes, nesting success and nesting productivity, are closely related. A nest’s success is defined as having 1 or more young fledged while a nest’s failure is defined has having 0 young fledged. An active nest’s productivity is defined as the number of young fledged observed. This number appears to be between 0 and 3, with 4 having been observed extremely rarely in the areas studied here.

3.3 Statistical analysis

3.3.1 Use of several statistical methods throughout the years

In the 1999-2007 Monitoring Program, difference in nest occupancy rates and in nesting success were said to have been tested using a one-tailed Fisher exact test up to 2005 and a logistic regression (on a dummy variable representing the experimental/control area) afterwards. No motivation was given for the change.

We notice numerous errors and inconsistencies in the statistical sections of the reports. The following are example of errors or inconsistencies:

- In the 2003 report, Table 2, the one-tailed Fisher exact test p-value for the 2003 nest occupancy rate (Nesting Activity) is reported as 0.109: it appears that the wrong tail was used in the test - the correct value is 0.957.
- In the 2004 report, Table 4, results reported for 1999-2003 as those of a one-tailed Fisher exact test are different from those reported in 2003. In fact, contrary to the statement in the methodology and the column headers, the 2004 report p-values are as follows: for nesting success 2004 only, the two-tailed Fisher exact test p-value (0.209) is reported; for all other nest occupancy rate (nesting activity) and nesting success cases, the one-tailed Logistic Regression p-value is reported.
- In the 2007 report, Table 4, the two-tailed logistic regression p-value for 2007 nest occupancy rate (nesting activity) is reported as 0.578. The correct value is 0.766 as shown in Appendix B, section 2. (The incorrect value appears to be that for an “all-year” analysis shown in Appendix B, section 1.)

Productivity was tested using the two-sample t-test. While not specified in the methodology, spot-checks make us believe that the equal variance (homoscedastic) version of the test was used.

The documentation indicates that other methods were used through the years including ANOVA. The 1999-2007 Monitoring Program reports also included results from other analyses in appendices.

3.3.2 Significance threshold

The statistical significance threshold was chosen to be α=0.20, leading to a probability of Type I errors (false positives) of 20%. This choice appears to have been made in order to limit the probability of Type II errors (false negatives).

3.3.3 Confidence intervals

While estimates of the parameters (nest occupancy rate, nesting success and nesting productivity) are given, confidence intervals are not supplied either for the parameters or for the parameter differences between the experimental and control values.

4 Review of the 1999-2007 Monitoring Program

In this section, we review the 1999-2007 Monitoring Program and its results.

4.1 Observational study

The gold standard in research on the effect of a treatment is the “randomized controlled experiment” where subjects are randomly assigned to a treatment group and to a control group. Random assignment insures that any observed difference in outcome between the groups is due to either the treatment or to chance. Probability theory is used to quantify the probability that the observed outcome is due to chance alone (the observed statistical significance level or “p-value”). If that probability is small (0.05 is a traditional probability threshold), the researcher will tend to conclude that the observed difference is due to the treatment and that there may be a cause-and-effect relationship.

In many situations, it is not possible to carry out a randomized controlled experiment. In these situations, an “observational study” is often carried out. A comparison between an existing control group (subjects not receiving the treatment) and an existing treatment group (subjects receiving the treatment) is an example of observational study. The LLTA Osprey Monitoring Program is an observational study.

Observational studies are subject to biases. A bias occurs when some factor other than the treatment can cause a difference in the outcome. A bias can be overt (the researcher is aware of the bias) or hidden (the researcher is not aware of the bias). Overt biases can sometimes be corrected for (e.g. by using a stratified analysis). Hidden biases cannot be corrected for.

In the LLTA Osprey Monitoring Program, the treatment group consists of nesting sites within a subarea of the LLTA and the control group consists of nesting sites within an adjacent area located to the east of the treatment area. The LLTA Osprey Monitoring Program reports do not mention any overt biases or correction for overt biases. However, one can suspect that hidden biases could be present. For example, the treatment and control areas are adjacent, the control area being situated immediately to the east of the treatment area. The east-west relationship between the groups suggests that some systematic differences in weather patterns (including winds) could be present. The geography, including the configuration of rivers and lakes (food sources), the rock formation and the forested land (availability of nesting sites) could also be systematically different.

4.2 The interpretation of the Nest Occupancy Rate

We consider that nest occupancy may be difficult to interpret. An important decline in nest occupancy would very likely indicate an important and sudden change in the number of nesting pairs. Conversely, however, a progressive decline in the number of nesting pairs may not be reflected in a progressive decline in nest occupancy.

We believe that it would be difficult to model properly the relationship between the rate of change of the population and the nest occupancy rate. The nest occupancy rate depends on the lifespan of the nests (probability that an occupied or a non-occupied nest survives 1, 2... years) and on the propensity of the population to build new nests. The problem is further complicated by external factors: for example, nests built on rock outcrops or man-made structures are likely to have a longer lifespan than nests built at the top of tall spruces. As a consequence, a change in nest occupancy rate could be caused by a change in nest lifespan (e.g. a change in winds) or by a change in the population propensity to build nests.

To the best of our knowledge, the detailed information required to construct such a model is not available.

These difficulties would be compounded when comparing two groups since differences in the parameters of the model could become confounding factors.

Combining the sampling issues (next section) and the interpretability issues, we conclude that the nest occupancy rate is unlikely to be a very useful indicator of the impact of low level flights on osprey reproduction.

The main reviewer has pointed out that raptors tend to occupy “territories” (e.g. a rock outcrop above a good fishing spot) and that the interesting statistic may be “territory occupancy rate”, suggesting that the nest occupancy rate may have been implicitly taken to be a surrogate statistic for the territory occupancy rate.

Of course, even as a surrogate for the territory occupancy rate, the nest occupancy rate suffers from the issues described above. However, the reviewer observation suggests that the number of active nests in some preselected areas (i.e. territories) could be a better population index. Such areas would be small, historically occupied areas, like a section of a river bank (say 1 km deep) or a particular rock formation that can be swept systematically on an annual basis. This approach would be similar to the marine fish sampling using fixed stations (e.g. fish count from standardized trawls done at the same locations each year). However, “fixed station sampling” can lead to false conclusions when the location of the population shifts even slightly within the area of observation.

Technical Details

To illustrate the difficulty in interpreting the nest occupancy rate, we have run a simulation of two populations, one stable and one declining. Both populations started with approximately 400 nesting pairs and 800 nests (i.e. 50% occupancy). The number of nesting pairs remained around 400 in the stable population, while it declined by approximately 10% per year in the declining population. We made some reasonable assumptions on the lifespan of unused nests and on the propensity of the nesting pairs to build new nests. Figure 1 shows the result of a single 30 year simulation for each population.

For the stable population, the number of nesting pairs, the number of nests and the nest occupancy remains relatively constant. For the declining population, the number of nesting pairs and the number of nests become negligible after 20 years. However, the nest occupancy rate remains approximately constant (but it is on average 10% below that of the stable population).

This result is due to the eventual collapsed of unused nests: in a declining population, the number of available nests will also decline and the nest occupancy rate will remain constant.

All other things being equal [ceteris paribus]***, the nest occupancy rate is related to the rate of change of the population numbers (and not to the population numbers themselves).***

Ceteris paribus*, if a population starts a continuous decline at a constant rate, nest occupancy will show a single, one-time, shift to a lower value and then remain at that lower value.*

Ceteris paribus*, if the population number suddenly changes to a lower level and then remains at that lower level, nest occupancy will show a brief (and probably undetectable) dip and then return to its previous level.*

Similarly, even if population A and population B have different numbers or density, the nest occupancy rates can be the same.

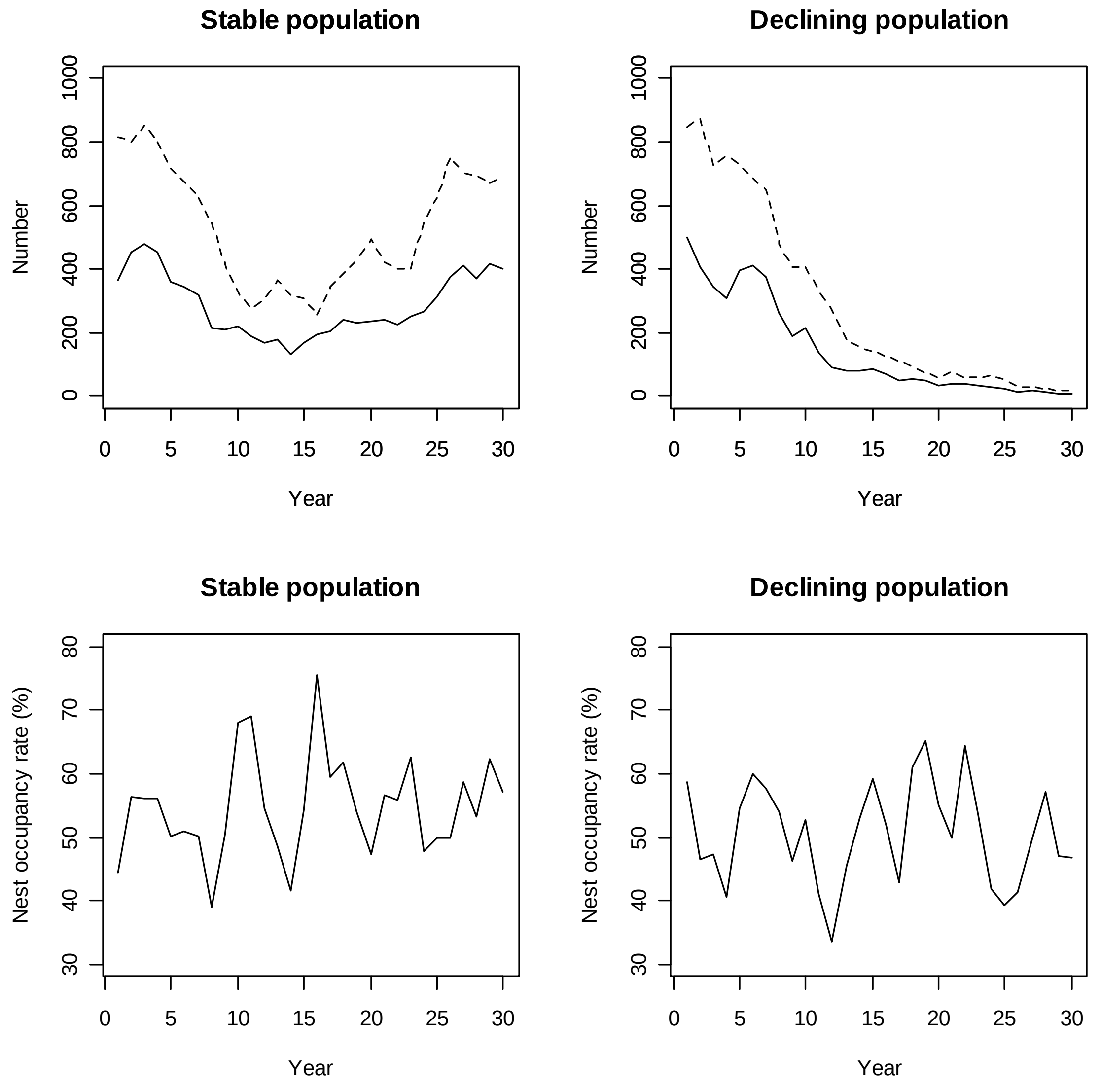

Figure 1. Results of a 30 year simulation of population size (number of nesting pairs – solid line in the top panels), number of nests available (dashed line in the top panels) and nest occupancy rate (lower panels) for a stable population and for a population declining by 10% per year on average. The nest occupancy rate is very similar for both populations.

4.3 Sampling method

The sampling method is clearly not equivalent to random sampling. The sampling is biased in favour of sites previously used and in favour of sites closer to Goose Bay (in fact, the 2007 sample sites are in the northern half of the treatment and of the control areas, just south of Goose Bay).

We note that treatment area nests that are closer to Goose Bay are likely to be more exposed to low level flights since planes must depart and return to Goose Bay. [This may actually be beneficial to the study since this can amplify any effect from the low level flights: nests further from Goose Bay are likely to be exposed to fewer flights.]

While these biases appear to apply equally to the control and the treatment group, subtle differences can be present: for example, it was mentioned that many more sites are available in the control area than in the treatment area; this could imply that selected nests are closer to Goose Bay in the control area than in the treatment area (this does not appear to be true for the 2007 samples) or that the control sample may contain more "old" sites than the treatment sample, since the "quota" may be reached faster in the control area.

A goal of the sampling procedure is to minimize fuel usage. This suggests that nests most likely to be occupied may be targeted and found earlier than non-occupied nests. As consequence, the nest occupancy rate may be over-estimated.

Overall, we consider that these characteristics of the sampling method are likely to have a major impact on the estimates of the nest occupancy rate and a lesser impact on the estimates of the nesting success rate and the nest productivity.

We cannot determine whether or not the sampling method has an impact on the comparisons between the experimental area and the control area.

We conclude that the sampling method is not similar to simple random sampling.

4.4 Alternative hypothesis

The choice of an alternative hypothesis is not consistent between reports. We consider that the alternative hypothesis should correspond to a negative impact of the LLT activity: decreased nest occupancy rate, decreased nesting success rate and decreased nesting productivity.

We consider that the statistical tests should be one-tailed tests, with the direction (tail) corresponding to an alternative hypothesis describing a negative impact of LLT on reproduction activity.

4.5 Statistical significance threshold and power

4.5.1 Type I error (false positives)

The empirical significance level (p-value) and the significance threshold play a central role in interpreting statistical tests. The significance threshold of 0.05 is very widely used. However, I consider that the following advice gives a more nuanced view of the significance threshold:

> "As a guide, it could be said that, when one's attitude is a priori "neutral" to a particular type of discrepancy, one begins to be slightly suspicious of a discrepancy at the α=0.20 level, somewhat convinced of its reality at the 0.05 level, and fairly confident of it at the 0.01 level. In practice, an experimenter's prior belief in the possibility of a particular type of discrepancy must affect his attitude. If the alternative hypothesis was plausible a priori, the experimenter would feel much more confident of a result significant at the 0.05 level than if it seemed to contradict all previous experience."
>
> *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, Box, Hunter, Hunter, Wiley-Interscience, 1978.

According to this scale, the choice of a significance threshold of α=0.20 for the osprey study corresponds to rejecting the null hypothesis when a discrepancy makes one "slightly suspicious". Assuming that there is actually no discrepancy, a significance threshold of α=0.20 also means that one will conclude incorrectly that there is a discrepancy (type I) 20% of the time.

The choice of the significance threshold was attributed to a desire to reduce the probability of a type II error, i.e. of concluding that there is no discrepancy when there is one (false negative).

The choice of type I error should depend on the consequence of incorrectly rejecting the null hypothesis. If rejecting the null hypothesis leads to a very large cost or to a high risk intervention, one will want to keep the probability of type I error low.

Given the information available to us, the consequences of a type I error (incorrectly concluding that the LLT activity affects osprey reproduction) would be important: they could range from wasteful spending on further studies and to the closing of the LLT program. We consider that α=0.20 is inappropriate.

We recommend that, when interpreting statistical tests, the above guidelines from Box, Hunter, Hunter be used. In particular, we recommend that the significance threshold for tests on single year data be around α = 0.05.

4.5.2 Type II error (false negatives) and power

It is important to know the power of the statistical tests used for a well chosen alternative hypothesis. The probability of type II error, β, for a given alternative is *1–Power* for that alternative. Therefore, knowing the "power curve" of a test is equivalent to knowing the probability of type II error for all alternative hypotheses.

While the power curve gives more information than the probability of type II error for a single alternative, it is often convenient to use the probability of type II error for a single alternative to describe the properties of a test.

A "well chosen alternative hypothesis" should reflect an "important" biological change. The corresponding probability of type II error should reflect the importance of the biological change. While many approaches are possible, we suggest two:

- The alternative hypothesis could be a LLT effect greater than a fraction of the natural temporal variation of the parameter. For example, if nesting success rate varies from 20% to 80% between years in the control area, the alternative hypothesis could be that the LLT effect is one half of the natural variation, i.e. (80-20)/5 = 30%. Given that this change is less than the natural variation, we would propose a probability of type II error between β=0.05 and β=0.10 (i.e. a power between 0.95 and 0.90).
- The alternative hypothesis could be a LLT effect greater than the smallest effect believed to be irreversible. For example, if one considers that the LLTA osprey population could bounce back from a 50% decline but not from a deeper decline, a 50% LLT effect would be a natural alternative hypothesis. Given that such a change (resulting in extinction of the local population) would be biologically very important, we propose that the probability of type II error β should 0.05 or less (i.e. a power of 0.95 or more) and preferably around 0.01.

We recommend that the target probability of type II error, β, be selected according to the importance of the biological change reflected by the alternative hypothesis chosen.

4.6 Multiple testing

If the null hypothesis is true, each time that a statistical test with a significance threshold α is used, the probability of committing a type I error is α. If a statistical test is used independently k times and the null hypothesis is true in all cases, the "family-wise" probability of committing at least one type I error is $1 - (1 - \alpha)^k$. For example, if α=0.05 and k=9, $1 - (1 - \alpha)^k = 0.37$ and if α=0.20 and k=9, $1 - (1 - \alpha)^k = 0.87$. Therefore, when k increases, the probability of committing at least one type I error increases.

The 1999-2007 Osprey Monitoring analyses present such a situation: for each outcome, a single test is applied 9 times (once for every year). Using α=0.20, one obtains a family-wise Type I error rate of 0.87, a very high value for the 9-year sequence of tests. This means that, if the LLT activity had no impact at all, one would very probably incorrectly conclude otherwise at the end of the 9-year sequence. This is clearly an undesirable situation.

Statisticians recognize that such "multiple comparisons" can falsely give the appearance of significance. One of the popular remedy is the Bonferroni method (often known as the Bonferroni correction) which consists of using a single test significance threshold of α/k. This method insures that the multiple comparisons will have a global probability of at least one type I error equal to α, which becomes the family-wise probability of error. Several other similar methods are available.

Applied to 9 years of data from the 1999-2007 Osprey Monitoring Program, a family-wise probability of Type I error would require setting the single test threshold to 0.20/9 = 0.022. Under this condition, only 2007 shows significant differences (notice that there was no LLT activity in 2006 and 2007).

The Bonferroni approach and other similar approaches have often been found to be too conservative, i.e. to have low power. More recent methods have been shown to perform better (the recent, say post-1990, literature on the subject is vast).

Using a multiyear statistical model which reduces the number of hypotheses avoids multiple tests. Such models will be discussed below.

Combining our remark on significance in the previous subsection with the increasing impact of multiple comparisons, the number of sampling years increases, we conclude that the choice of an annual test with α=0.20 significance threshold is not statistically sound.

If sequences of single-year statistical tests are presented, we recommend that a "multiple comparison correction" be used in the interpretation of the test results.

4.7 Statistical tests for single year comparisons

4.7.1 Impact of the quota sampling approach on Nest Occupancy Rate estimate

The "quota" approach (i.e. to sample until 30 active nests have been found) is unusual, but it is similar to the "inverse sampling" method in tag-recapture sampling. Assuming simple random sampling, the total number n of nests to be observed in order to obtain 30 occupied nests follows a negative binomial distribution with parameters p = proportion of occupied nests and r = 30. We have verified that the estimator is at most only slightly biased and that the usual estimator central limit theorem estimator of the standard error is satisfactory.

4.7.2 Test on Nest Occupancy Rate and Nesting Success

The 1999-2007 Osprey Monitoring Program reports used the one-tailed Fisher Exact test until 2005 and logistic regression from 2006. The earlier data were then retested using logistic regression.

The Fisher exact test is a test on a contingency table of the following type:

Table 2. Two x two contingency table

		X		
		A	B	Total
Y	U	n_{11}	n_{12}	$n_{1.}=n_{11}+n_{12}$
	V	n_{21}	n_{22}	$n_{2.}=n_{21}+n_{22}$
	Total	$n_{.1}=n_{11}+n_{21}$	$n_{.2}=n_{12}+n_{22}$	

The two-tailed Fisher exact test has been proven to be the uniformly most powerful test among unbiased tests for this type of problem (Lehmann, 1986, cited by Good, 2005).

If the null hypothesis is true ($p_0 = p_1 = 0.60$), the probability of Type I error is not exactly 0.05 because the number of possible values of the empirical significant level is finite and relatively small ($\min(n_{.1}, n_{.2}, n_{1.}, n_{2.}) +1$). A post-computation randomization process can yield a test with probability of Type I error = α, but it is rarely used.

The one-tailed version of the Fisher exact test is the exact permutation test. We have computed the power of the one-tailed Fisher exact test for various values of p_0, p_1 and the sample size $n_0 = n_1 = n$.

Since we have chosen to describe the impact of the LLT as a proportion of the original nesting success rate, the impact is more difficult to detect for smaller original nesting success rates (for example, a –10% impact on a success rate of 25% is only 2.5%).

The results are shown in the following tables. We note that probability of Type I error appears, in all cases, lower than the chosen significance threshold: again, this is due to the fact that counts are integer values and that 0.05 or 0.10 may not be close to any of the finite numbers of empirical significance levels possible for a given set of marginal distributions.

If the non-impacted (i.e. control area) success rate is p_0 = 25% (a relatively rare occurrence), an LLT impact of –50% gives an impacted success rate of 12.5%. For any reasonable sample size, the power of the test to detect such a small change is low.

Non-impacted success rates p_0 = 50% or p_0 = 75% are more interesting. In these cases, with sample sizes = 30, we consider that the power of the test is adequate for a LLT impact of 50% or more, i.e. cutting the nesting success rate in half.

We also note that, as expected, the choice of α affects the power of the test. For example, using a sample size of 30, a non-impacted success rate of p_0 = 75%, and an impact rate of –30%, the power of the test increases from 0.45 (β = 0.55) to 0.63 (β = 0.37) as α changes from 0.05 to 0.10.

We consider that, with 9 years of data collected, single-year nesting success rate analyses are not desirable. However, if such analyses are carried on, we recommend that a one-tailed Fisher exact test be used, that the power of the test be reported and taken into consideration in the decision making.

The same observations and recommendations would apply to analyses of the nest occupancy rate, if carried out.

Table 3. Power of the one-tailed Fisher exact test on the presence of a negative effect of the LLT activity on nesting success according to the impact as a percentage of the success rate (e.g. a 10% relative reduction changes a 50% success rate to a 45%) for $\alpha = 0.05$ and $\alpha = 0.10$ and sample sizes = 30.

Sample sizes = 30 / group ***α = 0.05***		Non-impacted success rate (p_0)			
		25%	50%	75%	90%
No impact	(α)	0.02	0.03	0.02	0.01
–10%	$p_1 = 0.9 \times p_0$	0.04	0.06	0.09	0.14
–20%	$p_1 = 0.8 \times p_0$	0.06	0.13	0.25	0.43
–30%	$p_1 = 0.7 \times p_0$	0.10	0.22	0.45	0.72
–40%	$p_1 = 0.6 \times p_0$	0.15	0.36	0.70	0.91
–50%	$p_1 = 0.5 \times p_0$	0.21	0.52	0.85	0.98
Sample sizes = 30 / group α = ***0.10***		Non-impacted success rate (p_0)			
		25%	50%	75%	90%
No impact	(α)	0.06	0.07	0.06	0.05
–10%	$p_1 = 0.9 \times p_0$	0.08	0.15	0.17	0.27
–20%	$p_1 = 0.8 \times p_0$	0.12	0.26	0.39	0.60
–30%	$p_1 = 0.7 \times p_0$	0.19	0.40	0.63	0.84
–40%	$p_1 = 0.6 \times p_0$	0.26	0.56	0.84	0.95
–50%	$p_1 = 0.5 \times p_0$	0.35	0.71	0.94	0.99

Table 4. Power of the one-tailed Fisher exact test on the presence of a negative effect of the LLT activity on nesting success according to the impact as a percentage of the success rate (e.g. a 10% relative reduction changes a 50% success rate to a 45%) for $\alpha = 0.05$ and $\alpha = 0.10$ and sample sizes = 50.

Sample sizes = 50 / group		Non-impacted success rate (p_0)			
$\alpha = 0.05$		25%	50%	75%	90%
No impact	(α)	0.03	0.02	0.03	0.02
–10%	$p_1 = 0.9 \times p_0$	0.06	0.09	0.15	0.26
–20%	$p_1 = 0.8 \times p_0$	0.09	0.19	0.41	0.66
–30%	$p_1 = 0.7 \times p_0$	0.17	0.35	0.71	0.92
–40%	$p_1 = 0.6 \times p_0$	0.25	0.57	0.91	0.99
–50%	$p_1 = 0.5 \times p_0$	0.39	0.79	0.98	1.00
Sample sizes = 50 / group		Non-impacted success rate (p_0)			
$\alpha = 0.10$		25%	50%	75%	90%
No impact	(α)	0.07	0.07	0.07	0.05
–10%	$p_1 = 0.9 \times p_0$	0.11	0.16	0.26	0.40
–20%	$p_1 = 0.8 \times p_0$	0.18	0.31	0.53	0.79
–30%	$p_1 = 0.7 \times p_0$	0.27	0.50	0.81	0.96
–40%	$p_1 = 0.6 \times p_0$	0.39	0.71	0.95	1.00
–50%	$p_1 = 0.5 \times p_0$	0.55	0.87	0.99	1.00

Table 5. Power of the one-tailed Fisher exact test on the presence of a negative effect of the LLT activity on nesting success according to the impact as a percentage of the success rate (e.g. a 10% relative reduction changes a 50% success rate to a 45%) for $\alpha = 0.05$ and $\alpha = 0.10$ and sample sizes = 100.

Sample sizes = 100 / group		Non-impacted success rate (p_0)			
$\alpha = 0.05$		25%	50%	75%	90%
No impact	(α)	0.04	0.04	0.03	0.03
–10%	$p_1 = 0.9 \times p_0$	0.09	0.13	0.26	0.49
–20%	$p_1 = 0.8 \times p_0$	0.18	0.35	0.69	0.94
–30%	$p_1 = 0.7 \times p_0$	0.31	0.65	0.94	1.00
–40%	$p_1 = 0.6 \times p_0$	0.47	0.85	1.00	1.00
–50%	$p_1 = 0.5 \times p_0$	0.69	0.98	1.00	1.00
Sample sizes = 100 / group		Non-impacted success rate (p_0)			
$\alpha = 0.10$		25%	50%	75%	90%
No impact	(α)	0.08	0.08	0.08	0.07
–10%	$p_1 = 0.9 \times p_0$	0.16	0.22	0.38	0.61
–20%	$p_1 = 0.8 \times p_0$	0.29	0.49	0.80	0.97
–30%	$p_1 = 0.7 \times p_0$	0.43	0.78	0.98	1.00
–40%	$p_1 = 0.6 \times p_0$	0.62	0.93	1.00	1.00
–50%	$p_1 = 0.5 \times p_0$	0.81	0.99	1.00	1.00

4.7.3 Discrete variable: Nesting Productivity

Three distributions of the number of fledged young observed in the control area or the LLTA are shown in the following table together with a blended, high variance distribution. The observed distributions are typical for low, middle and high productivity. The blended distribution is a 50-50 mixture of the low and high distributions constructed to obtain a high variance case. Our standard deviations differ slightly from those found in the annual reports, due to some ambiguities in the raw databases.

Table 6. Theoretical density distributions of the number of fledged young used in simulations.

	Probability			
Number	Low productivity (LLTA 2005)	Middle productivity (Control 2006)	Middle productivity (Composite)	High productivity (Control 2007)
	Probability			
0	0.70	0.25	0.37	0.04
1	0.20	0.31	0.16	0.11
2	0.07	0.34	0.22	0.39
3	0.03	0.09	0.25	0.46
4	0.00	0.00	0.00	0.00
Mean	$\mu = 0.43$	$\mu = 1.28$	$\mu = 1.35$	$\mu = 2.27$
Standard Deviation	$\sigma = 0.82$	$\sigma = 0.94$	$\sigma = 1.21$	$\sigma = 0.76$

The low and high productivity distributions are strongly asymmetrical. For annual comparisons, the ratio of the observed variances varies between 1.1 and 2.1.

We have examined one-tailed tests on the average productivity using the low, high and composite distributions. For equal sample sizes of 30 or more, the one-tailed t-test with unequal variances and the Welch-Satterthwaite estimate of the degrees of freedom and a randomization test give similar results. For the data in this study, the results of the equal variance and unequal variance version are very similar in essentially all cases. Given the observed distributions, the variance ratios and the sample sizes, this conclusion is consistent with the well established characteristics of the t-test.

The following tables show the power of the test for various non-impacted distributions and impact levels.

As for nesting success, any LLT impact in a low productivity situation is difficult to establish since that absolute impact will be small.

The middle and high productivity cases, $\mu = 1.35$ and $\mu = 2.27$, are the most interesting. We consider that, for single year comparisons, the power of the test is acceptable for LLT impact of –30% or more even with the current sample size of 30 in each area.

We consider that, with 9 years of data collected, single-year mean nesting productivity analyses are not desirable. However, if such analyses are carried on, we recommend that a one-tailed unequal variance t-test be used, that the power of the test be reported and taken into consideration in the decision making.

Table 7. Power of the one-tailed Welch variation of the unequal variances t-test on the presence of a negative effect of the LLT activity on Nest Productivity according to the impact as a percentage of the productivity (e.g. a 10% relative reduction changes a mean number of fledglings of 2 to a mean number of 1.80) for $\alpha = 0.05$ and $\alpha = 0.10$ and sample sizes = 30.

Sample sizes = 30 / group ***$\alpha = 0.05$***			Non-impacted mean nest productivity (H_0)		
			$\mu_0 = 0.43$	$\mu_0 = 1.35$	$\mu_0 = 2.27$
Relative reduction of the Mean productivity due to LLT activity	No impact	(α)	0.05	0.05	0.05
	–10%	$\mu_1 = 0.9 \times \mu_0$	0.08	0.11	0.22
	–20%	$\mu_1 = 0.8 \times \mu_0$	0.12	0.22	0.52
	–30%	$\mu_1 = 0.7 \times \mu_0$	0.17	0.38	0.79
	–40%	$\mu_1 = 0.6 \times \mu_0$	0.25	0.52	0.94
	–50%	$\mu_1 = 0.5 \times \mu_0$	0.36	0.72	0.99
Sample sizes = 30 / group ***$\alpha = 0.10$***			Non-impacted mean nest productivity (H_0)		
			$\mu_0 = 0.43$	$\mu_0 = 1.35$	$\mu_0 = 2.27$
Relative reduction of the Mean productivity due to LLT activity	No impact	(α)	0.11	0.10	0.10
	–10%	$\mu_1 = 0.9 \times \mu_0$	0.14	0.20	0.36
	–20%	$\mu_1 = 0.8 \times \mu_0$	0.20	0.34	0.67
	–30%	$\mu_1 = 0.7 \times \mu_0$	0.27	0.52	0.88
	–40%	$\mu_1 = 0.6 \times \mu_0$	0.39	0.66	0.98
	–50%	$\mu_1 = 0.5 \times \mu_0$	0.50	0.83	1.00

Table 8. Power of the one-tailed Welch variation of the unequal variances t-test on the presence of a negative effect of the LLT activity on Nest Productivity according to the impact as a percentage of the productivity (e.g. a 10% relative reduction changes a mean number of fledglings of 2 to a mean number of 1.80) for $\alpha = 0.05$ and $\alpha = 0.10$ and sample sizes = 50.

Sample sizes = 50 / group ***α = 0.05***			Non-impacted mean nest productivity (H_0)		
			$\mu_0 = 0.43$	$\mu_0 = 1.35$	$\mu_0 = 2.27$
Relative reduction of the Mean productivity due to LLT activity	No impact	(α)	0.05	0.05	0.06
	–10%	$\mu_1 = 0.9 \times \mu_0$	0.08	0.14	0.32
	–20%	$\mu_1 = 0.8 \times \mu_0$	0.15	0.29	0.72
	–30%	$\mu_1 = 0.7 \times \mu_0$	0.24	0.52	0.94
	–40%	$\mu_1 = 0.6 \times \mu_0$	0.35	0.74	0.99
	–50%	$\mu_1 = 0.5 \times \mu_0$	0.49	0.90	1.00
Sample sizes = 50 / group ***α = 0.10***			Non-impacted mean nest productivity (H_0)		
			$\mu_0 = 0.43$	$\mu_0 = 1.35$	$\mu_0 = 2.27$
Relative reduction of the Mean productivity due to LLT activity	No impact	(α)	0.10	0.11	0.11
	–10%	$\mu_1 = 0.9 \times \mu_0$	0.16	0.23	0.46
	–20%	$\mu_1 = 0.8 \times \mu_0$	0.25	0.42	0.83
	–30%	$\mu_1 = 0.7 \times \mu_0$	0.36	0.67	0.97
	–40%	$\mu_1 = 0.6 \times \mu_0$	0.48	0.83	1.00
	–50%	$\mu_1 = 0.5 \times \mu_0$	0.64	0.95	1.00

Table 9. Power of the one-tailed Welch variation of the unequal variances t-test on the presence of a negative effect of the LLT activity on Nest Productivity according to the impact as a percentage of the productivity (e.g. a 10% relative reduction changes a mean number of fledglings of 2 to a mean number of 1.80) for $\alpha = 0.05$ and $\alpha = 0.10$ and sample sizes = 100.

Sample size = 100 / group ***$\alpha = 0.05$***			Non-impacted mean nest productivity (H_0)		
			$\mu_0 = 0.43$	$\mu_0 = 1.35$	$\mu_0 = 2.27$
Relative reduction of the Mean productivity due to LLT activity	No impact	(α)	0.05	0.05	0.05
	–10%	$\mu_1 = 0.9 \times \mu_0$	0.11	0.20	0.52
	–20%	$\mu_1 = 0.8 \times \mu_0$	0.22	0.49	0.95
	–30%	$\mu_1 = 0.7 \times \mu_0$	0.36	0.77	1.00
	–40%	$\mu_1 = 0.6 \times \mu_0$	0.55	0.94	1.00
	–50%	$\mu_1 = 0.5 \times \mu_0$	0.74	0.99	1.00

Sample sizes = 100 / group ***$\alpha = 0.10$***			Non-impacted mean nest productivity (H_0)		
			$\mu_0 = 0.43$	$\mu_0 = 1.35$	$\mu_0 = 2.27$
Relative reduction of the Mean productivity due to LLT activity	No impact	(α)	0.10	0.10	0.10
	–10%	$\mu_1 = 0.9 \times \mu_0$	0.19	0.30	0.66
	–20%	$\mu_1 = 0.8 \times \mu_0$	0.34	0.63	0.97
	–30%	$\mu_1 = 0.7 \times \mu_0$	0.50	0.87	1.00
	–40%	$\mu_1 = 0.6 \times \mu_0$	0.69	0.97	1.00
	–50%	$\mu_1 = 0.5 \times \mu_0$	0.85	1.00	1.00

5 Multi-year models

5.1 Nesting Success model based on presence/absence of LLT activity

We are exploring the following type of model:

A. Code the flight intensity as follows (based on number of flights approximation).

Table 10. Flight Activity and Flight Intensity coding

	Flight data		Flight_Activity		Flight_Intensity (jets x1000)	
Year	Number of missions	Number of jets	LLTA	Control	LLTA	Control
1999	1728	2970	1	0	3	0
2000	2758	4780	1	0	5	0
2001	3045	5306	1	0	5	0
2002	2431	4333	1	0	4	0
2003	n/a	n/a	1	0	3	0
2004	n/a	n/a	1	0	2	0
2005	n/a	n/a	1	0	1	0
2006	0	0	0	0	0	0
2007	0	0	0	0	0	0

Data for 2003 to 2005 are unavailable. However, the LLTA usage declined progressively during those years. We have estimated the missing flight intensities by a linear interpolation between 2002 and 2006 values.

B. We consider the Year as a nominal variable for the following reasons: (1) at this time, there is no strong evidence of a temporal structure and (2) we are only interested in the LLT effect. (The data suggest that a cyclical model may hold. However, a longer sequence of observations would be necessary to reliably fit a cyclical model.)

C. Create a logistic regression model with nesting success as the response variable, Year and Flight_Activity as two nominal explanatory variables:

$$\mathrm{Log}(p_{\mathrm{Success}}/(p_{\mathrm{Failure}}) \sim \mathrm{Constant} + \mathrm{Year} + \mathrm{Flight\_Activity}$$

D. Consider the coefficient of Flight_Activity as the parameter of interest.

The power of this analysis for the effect of flight activity is as follows.

Table 11. Power of the logistic regression model on the presence of a negative effect of the LLT activity on nesting success according to the number of years of observation and the impact as a percentage of the success rate (e.g. a 10% relative reduction changes a 70% success rate to a 63%) for α = 0.05 and α = 0.10.

Sample size = 30 per group × year α = 0.05		Number of years of observation						
		9	10	12	15	20	25	50
Relative reduction of the Success rate due to LLT activity	No impact (α)	0.051	0.054	0.055	0.059	0.050	0.050	0.050
	–10%	0.34	0.35	0.39	0.49	0.54	0.65	0.88
	–20%	0.76	0.82	0.87	0.93	0.98	0.99	1.00
	–30%	0.97	0.98	0.99	1.00	1.00	1.00	1.00
	–40%	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	–50%	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Sample size = 30 per group × year α = 0.10		Number of years of observation						
		9	10	12	15	20	25	50
Relative reduction of the Success rate due to LLT activity	No impact (α)	0.100	0.107	0.104	0.110	0.097	0.106	0.096
	–10%	0.48	0.50	0.53	0.62	0.68	0.77	0.93
	–20%	0.84	0.90	0.93	0.97	0.99	1.00	1.00
	–30%	0.99	0.99	1.00	1.00	1.00	1.00	1.00
	–40%	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	–50%	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Technical notes

The power was computed in the following way for each number of years and LLT impact level desired. The non-impacted nesting success rate is distributed uniformly between 25% and 85%. The desired LLT impacted nesting success rate was computed (e.g. 28% - 2.8%). Logistic regression was applied to 2,500 datasets created by a pseudo-random process and the test statistic for the coefficient of the LLT impact dummy variable was recorded. The estimated power was then computed for the desired values of α.

5.2 Nesting Success model based on the intensity of LLT activity or exposure

This model is similar to the previous model except for the usage of the number of jets (or number of missions – the two variables being highly correlated) as an explanatory variable.

$$\text{Log}(p_{\text{Success}}/(p_{\text{Failure}}) \sim \text{Constant} + \text{Year} + \text{Flight\_Intensity}$$

IEMR considers that some measure of exposure to LLT noise may become available. The model can use such a measure:

$$\mathrm{Log}(p_{\mathrm{Success}}/(p_{\mathrm{Failure}}) \sim \mathrm{Constant} + \mathrm{Year} + [\text{measure of exposure}]$$

5.3 Nesting Productivity model based on presence/absence of LLT activity

Considering again the low, middle and high productivity distributions of the number of fledglings shown in Table 6 (but not the artificial composite distribution), we observe that they appear to best fit a binomial distribution with parameters $n=3$ and $p=\mu/3$**.** For small values of μ, the binomial distribution $B(3, \mu/3)$ and the Poisson distribution $\mathrm{Poisson}(\mu)$ are similar and fit equally well; for $\mu>1.75$ approximately, however, the Poisson distribution does not fit essentially due to the apparent upper bound of 3 fledglings.

The standard deviations of the number of fledglings by year/area range from 0.77 to 1.20 and half are between 0.90 and 1.07. Therefore, the maximum variance ratio is 2.42. Given a sample size of 30 for each year/area, we consider that this is an acceptable level of heteroscedasticity for a simple analysis of variance.

Therefore, we propose to model the nesting productivity using a simple analysis of variance model.

$$\text{Number of fledglings} \sim \text{Constant} + \text{Year} + \text{Flight\_Activity}$$

5.4 Nesting Productivity model based on intensity of LLT activity

We propose to model the nesting productivity using a simple analysis of variance model.

$$\text{Number of fledglings} \sim \text{Constant} + \text{Year} + [\text{measure of exposure}]$$

Such a model is an approximation since one is likely to assume that the impact of the LLT is a proportional reduction of the number of fledglings (i.e. multiplicative) and the errors are not normal.

The number of observations, however, is very large (2 x 30 × [number of years]). Simulation results indicate that the proportion of false positives is very close to the chosen α at least for α between 0.05 and 0.10 and that the power is satisfactory for a number of years equal to 9 or more for LLT impact of –20% or more.

We also notice that the power of the test does not increase meaningfully after 9 years if we consider an impact of –30% or greater and that it increases only slowly for smaller impact.

Table 12. Power of the ANOVA model on the effect of the LLT activity on Nest Productivity according to the number of years of observation and the impact as a percentage of the success rate (e.g. a –10% relative reduction changes a 2.0 average number of fledglings to 1.8).

Sample size = 30 per group × year α = 0.05		Number of years of observation						
		9	10	12	15	20	25	50
Relative reduction of the Success rate due to LLT activity	No impact (α)	0.056	0.052	0.050	0.049	0.050	0.045	0.049
	–10%	0.32	0.32	0.38	0.43	0.57	0.65	0.92
	–20%	0.79	0.84	0.88	0.95	0.99	1.00	1.00
	–30%	0.97	0.99	1.00	1.00	1.00	1.00	1.00
	–40%	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	–50%	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Sample size = 30 per group × year α = 0.10		Number of years of observation						
		9	10	12	15	20	25	50
Relative reduction of the Success rate due to LLT activity	No impact (α)	0.102	0.109	0.109	0.099	0.102	0.095	0.096
	–10%	0.44	0.45	0.51	0.56	0.68	0.75	0.96
	–20%	0.87	0.90	0.93	0.97	0.99	1.00	1.00
	–30%	0.99	0.99	1.00	1.00	1.00	1.00	1.00
	–40%	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	–50%	1.00	1.00	1.00	1.00	1.00	1.00	1.00

6 Analysis of the 1999-2007 results

6.1 Nesting Success

The following table shows the results of applying the one-tailed Fisher exact test to the nesting success rate. Applying any multiple comparison correction, one would deduce that no significant difference has been found at the 0.05 threshold.

Without applying a multiple comparison correction, the only significant difference at the 0.05 threshold is for 2007, a year during which there was no LLT activity in the LLTA.

When the nesting success rate is low, the power of the test is naturally low because any LLT impact would be small and difficult to detect. We consider that, in all other cases, the power of the test for a -30% LLT impact is acceptable (~0.60-0.75).

Table 13. Year-wise permutation test for Nesting Success Rate. The alternative hypothesis states that the Nesting Success Rate is lower in the LLTA.

	Control Area		LLTA				
Year	Sample size	Nesting Success Rate	Sample size	Nesting Success Rate	LLTA Status	p Value	Power for $\alpha=0.05$ and LLT impact is –30%
1999	30	0.83	30	0.90	Activity	0.873	0.58
2000	30	0.33	30	0.53	Activity	0.966	0.13
2001	30	0.43	30	0.30	Activity	0.211	0.18
2002	30	0.43	30	0.50	Activity	0.781	0.18
2003	30	0.83	30	0.93	Activity	0.949	0.65
2004	30	0.87	30	0.70	Activity	0.105	0.65
2005	30	0.38	30	0.30	Activity	0.356	0.15
2006	29	0.77	30	0.53	**No activity**	0.052	0.48
2007	30	0.96	30	0.72	**No activity**	0.019	0.85

The results of applying a single logistic regression of nesting success on flight activity and on flight intensity are shown in Table 14 and Table 15, respectively.

In a logistic regression, $\log(p_{Success}/(p_{Failure})$ is modeled as a linear combination of the parameters. For nominal explanatory variables (year, flight_activity), a coding occurs. In the following table, the coefficient of *Flight_Activity=TRUE* is the most important. It is estimated to be 0.04, with a 95% confidence interval [-0.39,0.47]. Therefore, we estimate that $\log(p_{Success}/(p_{Failure})$ is 0.04 *greater* when there is LLT activity (*Flight_Activity=TRUE)* than when there is none (*Flight_Activity=False)*. This implies that the probability of Nesting Success is also *greater*. We cannot of course reject the null hypothesis that there is not LLT effect against the alternative that there is a negative LLT effect.

The other coefficients describe the year effect, relative to the base year, 1999.

Similarly for flight intensity, where, this time, the coefficient of ***Flight_Intensity***, 0.04, indicates that $\log(p_{Success}/(p_{Failure})$ ***increases*** by 0.04, approximately, for each 1000 jets flown. The 95% confidence interval for the coefficient of the parameter is [-0.08,0.16].

Table 14. Logistic regression of Nesting Success on flight activity. The coefficient of Flight_Activity is not significantly different from 0 and positive. Testing that LLT has a negative impact, we do not reject the null hypothesis (p = 0.574).

	Coefficient	Confidence interval	s.e.	Wald z	P (2-sided)	P (1-sided)
Intercept	1.85	[1.07,2.63]	0.40	4.68	0.000	
Year=2000	-2.14	[-3.04,-1.24]	0.46	-4.65	0.000	
Year=2001	-2.42	[-3.32,-1.52]	0.46	-5.20	0.000	
Year=2002	-2.01	[-2.91,-1.11]	0.46	-4.36	0.000	
Year=2003	0.15	[-0.93,1.23]	0.55	0.28	0.783	
Year=2004	-0.59	[-1.55,0.37]	0.49	-1.19	0.234	
Year=2005	-2.54	[-3.46,-1.62]	0.47	-5.42	0.000	
Year=2006	-1.23	[-2.17,-0.29]	0.48	-2.57	0.010	
Year=2007	-0.22	[-1.28,0.84]	0.54	-0.41	0.683	
Flight_Activity=TRUE	0.04	[-0.39,0.47]	0.22	0.19	0.853	***0.574***

Table 15. Logistic regression of Nesting Success on flight intensity. The coefficient of Flight_Intensity is not significantly different from 0 and positive. Testing that LLT has a negative impact, we do not reject the null hypothesis with (p = 0.731)

	Coefficient	Confidence interval	s.e.	Wald z	P (2-sided)	P (1-sided)
Intercept	1.82	[1.06,2.58]	0.39	4.67	0.000	
Year=2000	-2.18	[-3.10,-1.26]	0.47	-4.68	0.000	
Year=2001	-2.46	[-3.38,-1.54]	0.47	-5.23	0.000	
Year=2002	-2.02	[-2.92,-1.12]	0.46	-4.39	0.000	
Year=2003	0.15	[-0.93,1.23]	0.55	0.28	0.783	
Year=2004	-0.57	[-1.53,0.39]	0.49	-1.15	0.249	
Year=2005	-2.50	[-3.42,-1.58]	0.47	-5.30	0.000	
Year=2006	-1.20	[-2.12,-0.28]	0.47	-2.53	0.012	
Year=2007	-0.19	[-1.23,0.85]	0.53	-0.35	0.727	
Flight_Intensity	0.04	[-0.08,0.16]	0.06	0.62	0.539	***0.731***

6.2 Nesting Productivity

The following table shows an analysis of the average number of fledglings observed.

Applying any multiple comparison correction, one would deduce that no significant difference has been found at the 0.05 threshold.

Without applying a multiple comparison correction, significant difference at the 0.05 threshold are observed in 2004 and in 2007. As noted, 2006 and 2007 are years during which there was no LLT activity in the LLTA. In fact, the average number of fledglings in the LLTA appears to diminish (relative to the control area) as the number of flights is reduced.

Table 16. Year-wise statistical analysis of the Nesting Productivity.

	Control Area			LLTA								
Year	n	Mean number of fledglings	SD	n	Mean number of fledglings	SD	LLTA Status	Difference LLTA Minus Control	SE	Relative difference	p value	Power for α=0.05 and LLT impact is −30%
1999	30	1.57	0.94	30	1.77	0.90	Activity	0.20	0.24	13%	0.799	0.45
2000	30	0.57	0.90	30	1.03	1.13	Activity	0.47	0.26	82%	0.959	0.20
2001	30	0.70	0.88	30	0.60	1.04	Activity	-0.10	0.25	-14%	0.344	0.24
2002	30	0.93	1.17	30	1.00	1.14	Activity	0.07	0.30	7%	0.588	0.28
2003	30	1.90	0.77	30	1.73	1.03	Activity	-0.17	0.24	-9%	0.241	0.57
2004	30	1.97	1.07	30	1.43	0.96	Activity	-0.53	0.26	-27%	0.024	0.60
2005	29	0.72	1.01	30	0.43	0.77	Activity	-0.29	0.23	-40%	0.109	0.22
2006	30	1.33	1.05	30	1.00	0.96	**No activity**	-0.33	0.26	-25%	0.102	0.38
2007	26	2.27	0.83	29	1.66	1.20	**No activity**	-0.61	0.28	-27%	0.017	0.80

We cannot carry out the proposed analysis of variance on the nesting productivity because the raw data are not available for years 1999-2002.

6.3 Nest Occupancy Rate

While we consider the nest occupancy rate difficult to interpret and its estimates possibly biased due to the sampling method, we do carry out the statistical analysis for completeness purposes.

In the year-wise analysis, none of the differences are statistically significant. In the logistic regression, the null hypothesis is not rejected and the coefficient of the appropriate variable is compatible with a **positive** effect of the LLT on nest occupancy rates.

Table 17. Year-wise permutation test for Nest Occupancy Rate. The alternative hypothesis states that the Occupancy rate is lower in the LLTA.

	Control Area		LLTA			
Year	Sample size	Nest Occupancy Rate	Sample size	Nest Occupancy Rate	LLTA Status	p-value
1999	98	0.60	99	0.58	Activity	0.409
2000	82	0.39	77	0.45	Activity	0.837
2001	74	0.47	75	0.41	Activity	0.285
2002	112	0.31	81	0.42	Activity	0.954
2003	53	0.60	40	0.75	Activity	0.957
2004	44	0.73	44	0.75	Activity	0.686
2005	40	0.73	49	0.63	Activity	0.243
2006	52	0.62	61	0.51	**No activity**	0.170
2007	53	0.53	58	0.50	**No activity**	0.457

The logistic regression follows. The coefficient of ***Flight_Activity=TRUE*** and that of ***Flight_Intensity*** are both positive and the null hypothesis is not rejected in either analysis. The details are in the following tables.

Table 18. Logistic regression of Nest Occupancy on flight activity. The coefficient of Flight_Activity is not significantly different from 0 and positive. Testing that LLT has a negative impact, we do not reject the null hypothesis (p = 0.**716).**

	Coefficient	Confidence interval	s.e.	Wald z	P (2-sided)	P (1-sided)
Intercept	0.31	[0.00,0.62]	0.16	1.92	0.055	
Year=2000	-0.67	[-1.10,-0.24]	0.22	-3.12	0.002	
Year=2001	-0.59	[-1.02,-0.16]	0.22	-2.68	0.007	
Year=2002	-0.94	[-1.35,-0.53]	0.21	-4.49	0.000	
Year=2003	0.34	[-0.17,0.85]	0.26	1.30	0.195	
Year=2004	0.68	[0.13,1.23]	0.28	2.41	0.016	
Year=2005	0.36	[-0.17,0.89]	0.27	1.35	0.177	
Year=2006	-0.08	[-0.57,0.41]	0.25	-0.30	0.761	
Year=2007	-0.25	[-0.74,0.24]	0.25	-1.02	0.309	
Flight_Activity=TRUE	0.11	[-0.14,0.36]	0.13	0.79	0.431	***0.716***

Table 19. Logistic regression of Nest Occupancy on flight intensity. The coefficient of Flight_Intensity is not significantly different from 0 and positive. Testing that LLT has a negative impact, we do not reject the null hypothesis (p = 0.826).

	Coefficient	Confidence interval	s.e.	Wald z	P (2-sided)	P (1-sided)
Intercept	0.31	[0.02,0.60]	0.15	2.01	0.044	
Year=2000	-0.71	[-1.14,-0.28]	0.22	-3.23	0.001	
Year=2001	-0.62	[-1.05,-0.19]	0.22	-2.79	0.005	
Year=2002	-0.95	[-1.36,-0.54]	0.21	-4.56	0.000	
Year=2003	0.34	[-0.17,0.85]	0.26	1.30	0.195	
Year=2004	0.70	[0.15,1.25]	0.28	2.46	0.014	
Year=2005	0.40	[-0.13,0.93]	0.27	1.48	0.140	
Year=2006	-0.08	[-0.55,0.39]	0.24	-0.32	0.747	
Year=2007	-0.26	[-0.73,0.21]	0.24	-1.05	0.295	
Flight_Intensity	0.03	[-0.03,0.09]	0.03	0.94	0.347	***0.826***

7 Conclusion

7.1 Interpretation of the current results

The data from the 1999-2007 survey leads us to conclude that the LLT activity is very unlikely to have a negative effect on the nesting success or the nesting productivity of the osprey. The variations between the LLTA and the control area are not statistically significantly different; furthermore, our point estimate of the impact is essentially null relative to the natural variation.

The same statistical conclusion applies to the nest occupancy rate. However, we consider that the interpretation of any difference between nest occupancy rates is difficult, especially in the long term, and that the current sampling process is very likely to introduce some bias in the estimation.

We consider that the power of the statistical test carried out using the full database is sufficient to support these conclusions.

We consider that the most obvious sampling bias (i.e. toward nests close to the military base) would likely enhance the LLT impact, if there was any.

Other sources of evidence appear to support this conclusion. None of the studies carried out before 1999 seem to have indicated an LLT impact. Evidence from other sources (highway construction, power line maintenance – see IEMR Web site) indicates that ospreys can tolerate a fairly high level of disturbance during their nesting activity. The nest productivity also appears to be within the range reported for other North American sites (0.56 to 1.3 according to D.M.Bird, IEMR Web site).

Given the monitoring data and other observations, we conclude that any LLT impact on osprey reproduction is very likely to be much smaller than the natural variation in reproductive success.

Given that the 1999-2007 Monitoring Program is an observational study, one must consider the possibility that some hidden biases cancel quite exactly the impact of the LLT activity. It is unlikely that such biases, if they exist, can be discovered by accumulating more data using the same observational design, even if the sample sizes were increased.

Based on the evidence examined, we expect that continuation of the current Monitoring Program will NOT produce further information on LLT impact, should there be any.

7.2 Special circumstances

While we consider that the results of the monitoring program should be analyzed in a single model and that the data obtained from 1999 to 2007 do not show any LLT effect, we realize that special circumstances (e.g. introduction of a radically new activity) may warrant a one-year program. In this case, we recommend the use of Fisher's exact test and of the t-test for data with unequal variances.

7.3 Periodical monitoring

Periodical monitoring (e.g. monitoring, say, every 5 years) may come under consideration. We consider that such a monitoring program would give unreliable results due to the natural population fluctuation.

7.4 Database organization

Should the Monitoring Program be continued, we recommend that the information be kept in a single data base or spreadsheet so that consistency is insured and multiyear analyses can be carried out efficiently. In particular, the 1999-2007 data should be transfered in a single database.

7.5 Statistical tests

Should the Monitoring Program be continued, we recommend that the following statistical methods be applied.

7.5.1 Single-year analyses

Assuming that the sample size remains greater than or equal to 30 for each area, we recommend that single-year comparison of the nesting success rates be carried out using an exact one-tailed permutation test (the one-tailed version of Fisher exact test).

The same test should be used to compare the nest occupancy rates if such comparison is continued.

Assuming that the sample size remains greater than or equal to 30 for each area, we recommend that a single-year comparison of the nesting productivity be carried out using the t-test for population with unequal variances.

In all cases, the power of the test should be supplied.

If a multi-year sequence of single year comparisons is considered, a correction for multiple comparisons should be applied.

7.5.2 Multi-year analyses and intensity impact

We recommend that the multi-year analysis of the nesting success rate be carried out by a logistic regression, the year and the presence/absence of low level flights being used as nominal explanatory variables.

We recommend that the multi-year analysis of the Nest Productivity be carried out by analysis of variance, the year and the presence/absence of low level flight intensity being used as nominal explanatory variables.

We recommend that the multi-year analysis of the nesting success rate be carried out by a logistic regression, the year being used as a nominal explanatory variable and a measure of low level flight intensity as a covariate, if such a measure is available.

We recommend that the multi-year analysis of the nest productivity be carried out by analysis of variance, the year being used as a nominal explanatory variable and a measure of low level flight intensity as a covariate.

7.5.3 Probability of type I and type II error

Given the consequences of a type I error, we recommend that the significance level $\alpha = 0.05$ be used.

The choice of probability β of type II error must depend on the alternative hypothesis selected. If the alternative hypothesis corresponds to a relatively small LLT impact (e.g. a 30% reduction in success rate), we recommend that β be between 0.05 and 0.10. If the alternative hypothesis corresponds to a very important LLT impact (e.g. an impact leading to extinction of the local population), we recommend that β be at most 0.05 and preferably 0.01.

7.5.4 Confidence intervals

We recommend that, in future analyses, confidence intervals for differences in parameters between groups be obtained and accompany results of statistical tests.

8 Acknowledgments

The author thanks Dr. David Schneider for his insightful review of two drafts of the document.

References

Rosenbaum, Paul R. *Observational Studies*, Springer, 2nd ed., 2002, 375 pp.

Good, Phillip I. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, Springer, 3rd ed., 2004, 376 pp.

Stewart-Oaten, A., Bence, J. R., Osenberg, C. W. “Assessing effects of unreplicated perturbations: No simple solutions”, *Ecology 73*, 1992, pp. 1396-1404.