RESEARCH

Retrospective versus prospective designs for studies of crop-raiding by elephants in Kakum, Ghana

Richard F.W. Barnes,[1] Emmanuel Danquah,[2] Emmanuel M. Hema,[3] Umaru-Farouk Dubiure,[4] Mildred Manford,[5] Awo Nandjui[6] and Yaw Boafo[7]*

[1]Ecology, Behavior and Evolution Section, Biology Division, University of California at San Diego, La Jolla, CA 92093-0116, USA (Present address: Department of Medicine, University of California at San Diego, La Jolla, CA 92093-8651, USA)
[2]Department of Wildlife and Range Management, Faculty of Renewable Natural Resources, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[3]Université de Ouagadougou/CUP-D, laboratoire de Biologie et Ecologie Animales, 09 BP 848, Ouagadougou 09, Burkina Faso
[4]Wildlife Division, Accra, Ghana (*Present address*: Mole National Park, PO Box 8, Damongo, Ghana)
[5]2720 Foster Ave, Apt #1A, Brooklyn, NY 11210, USA
[6]21 BP 721, Abidjan 21, Côte d'Ivoire
[7]Wildlife Division, P.O. Box M239, Accra, Ghana
*corresponding author email: r1barnes@ucsd.edu

Abstract

Crop-raiding by elephants continues to increase as human populations in elephant ranges expand. The risk of crop losses can be reduced by identifying the most important farmland features that attract elephants. Risk factors vary from place to place and must be identified by site-specific studies. The most important risk factors include distance of farm to reserve boundary line, area under cultivation, number of crop types on farm and degree of each farm's isolation. Here we take the data from an earlier prospective study of crop-raiding around the Kakum Conservation Area in southern Ghana to illustrate a better method of analysis using a zero-inflated Poisson model. We then use the same data set to illustrate the advantages and drawbacks of a retrospective design. With a retrospective design a raided farm is matched with one or more intact farms at the end of the growing season. This method is cost-effective for field workers whose resources are limited because it does not require repeated visits to farms to monitor raids. The optimum sample size is about 30 raided farms that are each matched with at least two intact farms.

Additional key words: Ghana, elephants, crop-raiding, human-elephant conflict

Résumé

La maraude des cultures par les éléphants continuera à être exacerbée par l'expansion des populations humaines. On peut réduire le risque de pertes des cultures par l'identification des caractéristiques agricoles les plus importantes qui attirent les éléphants. Les facteurs de risque varient d'un endroit à un autre et doivent être identifiés par des études spécifiques au site. Ici, on prend les données d'une étude prospective antérieure

de maraude des cultures autour de la zone de conservation de Kakum au sud du Ghana pour illustrer une meilleure méthode d'analyse en utilisant un modèle de Poisson à inflation de zéros. Nous utilisons ensuite le même ensemble de données pour illustrer les avantages et les inconvénients d'une conception rétrospective. Avec un modèle rétrospectif une ferme endommagée est jumelée à une ou plusieurs fermes intactes à la fin de la saison de croissance. Cette méthode est rentable pour les travailleurs sur le terrain dont les ressources sont limitées. La taille optimale de l'échantillon est d'environ 30 fermes endommagées qui sont chacune jumelée avec au moins 2 fermes intactes.

Mots clés supplémentaires: Ghana, éléphants, destruction des cultures, conflits homme–éléphant

Introduction

People and elephants come into more frequent contact as human populations expand. The result is anger and despair as elephants ravage farmers' fields. Management for the benefit of farmers requires that one understands the farmland landscape features that attract elephants. The risk of crop raiding can be reduced by advising farmers how to modify their farming practices (Barnes et al. 2005). But every situation is different: while distance of farm to reserve boundary line, area under cultivation, number of crop types on farm and degree of farm's isolation may be risk factors common to all sites, wildlife managers will usually need to identify the most important variables that attract elephants to the farmland around their particular protected areas.

The objective of this type of study is to identify risk factors. These are the variables most strongly associated with incidents of crop raiding. An earlier paper described how the design of studies of crop raiding should take account of the aims of the study and the resources available (Barnes 2008). A second paper described simple methods for analyzing the data from such studies (Barnes 2009). That paper was aimed particularly at researchers who need to analyse their data before leaving the field site. In the present paper we expand upon the analysis described by Barnes et al. (2005) for risk factors for farms around the Kakum Conservation Area (KCA) in southern Ghana. In that study we observed a cohort of farms from the beginning to the end of the growing season. That was a prospective study because we identified the farms at the beginning of the field work and followed them forward throughout the length of the growing season (Gordis 2004). Here we show that a zero-inflated model provides a better fit to the data compared to our earlier analysis, because most farms are unlikely to be raided. That means that the outcome variable (number of raiding incidents on each farm) has many values that are zero (that is, there were many farms where no incidents were recorded). We then use the same data to simulate a retrospective study. A retrospective study is one in which the researcher identifies the farms at the end of the study period and then looks backward to collect data on what has happened during that period (Gordis 2004). We argue that a retrospective design is a more cost-effective use of resources than a prospective design. We also discuss the question that vexes most field workers: how large a sample is needed to achieve the statistical power that will reveal important effects?

Study site and methods

The Kakum Conservation Area (KCA) lies in the forest zone in south-west Ghana. Agriculture is the predominant activity in communities surrounding KCA, resulting in a landscape mosaic of cultivation, farm bush, secondary forest and swampland. Main cash crops cultivated include cocoa, oil palm, and citrus, whilst major food crops are maize, cassava, plantain, cocoyam, yam, rice and vegetables. Though the system of farming is rain-fed shifting cultivation, farming activity is done throughout the year, resulting in an all-year-round occurrence of crop raids. Kakum was the site for the largest and longest study of crop raiding ever undertaken in West Africa (Barnes et al. 2003, 2005, 2006; Boafo, et al. 2004). In that study a prospective design was chosen for the 2001 crop growing season whereby farms were identified at the beginning of the season and monitored for nine months until the end of season. Ten communities around the conservation area were selected at random. Two hundred and three farms in those communities were registered and each was visited at regular intervals. All crop raiding incidents on each farm were recorded during the growing season. In addition, a series of measurements was made on each farm: distance from

the KCA boundary, area of farm, distance from the next nearest farm, number of crop types and area of each type of crop.

In this study we used SAS 9.2 (SAS Institute Cary, North Carolina) to fit zero-inflated Poisson and conditional logistic regression models (Hosmer and Lemeshow 2000; Shoukri and Chaudhary 2007).

Prospective design

The number of crop raiding incidents was not normally distributed. Rather, there were many farms were untouched and each of these had a value of zero for the number of raids. Other farms suffered just one raid, a few suffered two or three, and a tiny handful suffered four or five (Fig. 1). Furthermore, all these numbers were integers. These features—a large number of zeroes plus integers that are positive—are typical of count data and the usual methods of ordinary least squares (OLS) regression should not be applied (Hilbe 2015). Instead, we used Poisson regression models, which are part of the family of generalized linear models (McCullough and Nelder 1989). These are the appropriate models to use with count data (McCullough and Nelder 1989; Manly 2001; Hilbe 2015).

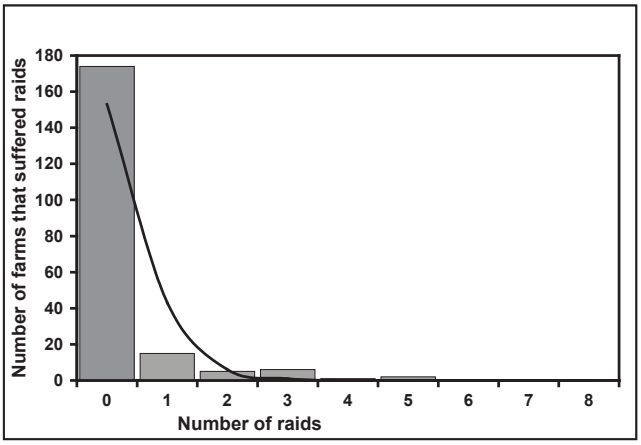

Figure 1. Frequency histogram of elephant crop-raiding incidents on a sample of 203 farms around the Kakum Conservation Area in the growing season of 2001. The curve shows the fitted Poisson distribution.

Ordinary least squares models have the form Y = a + bX, where Y is the dependent variable and X is the independent variable, and the residuals are normally distributed. The Poisson regression model has the form:

$$\ln Y = a + bX$$

which is the same as:

$$Y = \exp(a + bX)$$

The residuals are Poisson distributed and the model is fitted by maximum likelihood (McCullough and Nelder 1989; Hilbe 2015). The logarithmic relationship means that predicted values of Y can never be negative (i.e. you cannot have a negative number of crop raiding incidents) which is sometimes a problem with OLS models. In the earlier paper we fitted a Poisson model to the KCA data (Barnes et al., 2005). Counts on each farm of the numbers of crop raiding incidents were the response or dependent variable. The independent variables were those likely to attract elephants into the farmlands. While the Poisson model was the correct one to fit to these data, we now realize that we fitted the wrong kind of Poisson model. Of the 203 farms monitored during the course of the growing season, 174 (or 86%) were not afflicted by elephants (Fig. 1). This number of zeros is significantly greater than that expected from a Poisson distribution ($\chi 2 = 46.28$, df = 3, $p < 0.001$). The zero-inflated Poisson (ZIP) regression model is the appropriate method to use when you have a count data with a disproportionate number of zeros (Afifi et al. 2007).

The zero-inflated Poisson regression model is a two-part process. First, a logistic regression estimates the probability (p_0) of the count being zero rather than Poisson distributed. This is the zero-inflated part of the method. Then a Poisson regression model is fitted to the counts with probability $1 - p_0$ (Afifi et al. 2007).

To illustrate this example, we used only two independent variables: the distance (in metes) of the farm from the KCA boundary, and the number of crop types grown on each farm. The fitted ZIP model is shown in Table 1 (note that some farms were dropped because of missing values). Taking the inflate part of the table (the upper half of Table 1), the coefficient for the distance to boundary was 0.003. This value

is significant (p = 0.004). Because this is a logistic regression, the exponent of the coefficient is the odds ratio (Hosmer and Lemeshow 2000; Afifietal. 2007). The odds ratio is therefore exp (0.003) = 1.003. This means that as one moved away from the KCA the odds increased that the farm will be a zero (that is, it will be intact). Specifically, the odds of the farm remaining intact increased by a factor of 1.003 (0.3%) for each metre further from the park boundary. In other words, the risk of raid decreased significantly with increasing distance from the KCA boundary. On the other hand, the number of crop types had no effect on the risk of raid, since in this case p = 0.434 (Table 1).

Table 1. Results of the zero-inflated Poisson regression model applied to farms around the Kakum Conservation Area (n = 199 farms). SE: standard error

Stage	Parameter	Estimate	SE	p
Inflate portion	Intercept	0.140	1.017	0.890
	Distance to boundary	0.003	0.001	0.004
	Number of crop types	–0.153	0.195	0.434
Poisson portion	Intercept	–0.958	0.801	0.232
	Distance to boundary	0.000	0.000	0.928
	Number of crop types	0.351	0.135	0.009

The lower half of Table 1 deals with the Poisson part of the model, that is, the number of incidents (farms with a value of 1, 2, 3, 4 or 5 raids). We find that distance from the boundary was not associated with the number of raids (p = 0.928). In contrast, the number of crop types was strongly associated (p = 0.009) with the number of raids. The coefficient is 0.351. For a Poisson regression the exponent of the estimate is a measure of the risk ratio, the greater risk associated with a one unit increase in the independent variable. Here exp (0.351) = 1.421 which means that each time you increased the variety of crops on your farm by adding another crop type, you increased the risk of further raids by 42%.

This prospective design was expensive in terms of personnel and resources, because each farm had to be visited regularly throughout the study period. We found that only 29 out of 203 farms were raided by elephants. In other words, this was not an efficient way of finding raided farms. A retrospective design would have been far more efficient.

Retrospective design

Selection of controls

With a retrospective design, at the end of the growing season (or the year, or some clearly defined period of interest) one looks backwards in time (Barnes 2008). One then identifies a sample of farms that was raided during the growing season. Next one selects a sample of intact farms (the controls) for comparison.

Using the data from the KCA study, let us assume that we identified 29 raided farms at the end of the growing season. Before we sample a section of the intact farms there is an issue that we have not yet addressed: the 203 farms were not distributed randomly across the landscape. Rather, they fell in 10 randomly-distributed communities or clusters. The characteristics of the different communities may influence the probability of crop raids within each community. However, we can remove the variation due to the communities by matching each raided farm against one or more intact farms in the same community.

In the first step of this analysis, we matched each raided farm with one randomly selected intact farm from the same community. There was one intact farm that lacked a measurement of distance to the park boundary, so it was dropped. There were no other intact farms that had been measured in that particular community and so that left 28 raided and 27 intact farms in this sample for the analysis (Tables 2a and 3a).

In the second step we intended to match each raided farm with two intact farms from the same community. However, we realized that in some communities we had not measured enough intact farms. Therefore in communities with many intact farms we randomly selected up to three intact farms to match to each raided farm. With each raided farm matched with one, two or three intact farms we had a sample of 57 intact farms to the 28 raided ones, a ratio of almost 2:1 for the analysis (Tables 2b and 3b).

Table 2. Estimates of median distance from boundary and median number of crop types for raided and intact farms

(a) The experiment with one control per raided farm (n = 28 raided and 27 intact farms)

	Raided farms	Intact farms	Wilcoxon rank-sum test[1] z	p
Median distance to boundary (m)	324.5	690.0	2.64	0.008
Median number of crop types	3.0	2.0	1.63	0.103

(b) The experiment with two controls per raided farm (n = 28 raided and 57 intact farms)

	Raided farms	Intact farms	Wilcoxon rank-sum test z (p)
Median distance to boundary (m)	324.5	596.0	3.21 (0.001)
Median number of crop types	3.0	2.0	2.33 (0.020)

Table 3. Estimates from the conditional logistic regression models

(a) The experiment with one control per raided farm (n = 28 raided and 27 intact farms). CI: confidence intervals

	Odds ratio	95% CI for odds ratio	Wald p
Distance to boundary (m)	0.997	0.994, 0.999	0.021
Number of crop types	1.372	0.824, 2.282	0.224

(b) The experiment with two controls per raided farm (n = 28 raided and 57 intact farms)

	Odds ratio	95% CI for odds ratio	Wald p
Distance to boundary (m)	0.997	0.994, 0.999	0.021
Number of crop types	1.372	0.824, 2.282	0.224

With matched samples, conditional logistic regression, designed to account for the matching was applied (Hosmer and Lemeshow 2000; Woodward 2005).

Results

With one control per raided farm, the intact farms were on average twice as far from the KCA boundary as raided farms (Table 2a). The conditional logistic regression is another way of expressing this relationship (Table 3a): the odds ratio of 0.997 means that as one moved away from the boundary the odds of being raided diminished. On the other hand, the difference in number of crop types was not significant (Tables 2a and 3a).

With two controls per raided farm, we see again that raided farms were significantly closer to the park boundary. The conditional logistic regression tells us that, after adjusting for the number of crop types, the odds of suffering a raid decreased by a factor of 0.995 for each meter further from the park boundary (Table 3b). This time raided farms had significantly more types of crops than intact farms (Table 2b). After adjusting for distance from the boundary, the odds of being raided increased by 1.619 times for each extra crop planted on the farm (Table 3b).

The results shown in Tables 2 and 3 make the point that one control was insufficient with this type of design. With this number of raided farms, in order to detect a significant difference one needs at least two controls for each raided farm.

Discussion

We adopted the prospective design during a workshop involving the senior staff of the KCA and the research team. This design required regular monitoring of every farm. But, at a time when the local authorities were given to understand that the landscape was being ravaged by marauding elephants, in the end only 15% of the farms were recorded as raided. While it was a relief to learn that the problem was not as bad as claimed, regular monitoring of 203 farms was not a cost-effective way of deploying our resources.

In contrast, the retrospective method is far more cost-effective because repeated visits to each farm are not required; each farm need be visited only once, on the day when all the data are collected. On the other hand, it does not provide the same information. For example, the ZIP model from the prospective design indicates that distance from the boundary determines the risk of whether or not a farm is raided while the number of crop types determines whether a farm is raided once, twice, thrice or more times. The retrospective method shows that both distance from the park and number of crop types influence the risk of being raided.

The data collection methods were not designed for a retrospective analysis or a matched analysis. Rather, we have taken advantage of this large data set to illustrate the retrospective approach. The retrospective experiment indicated that if you have 28 raided farms then one control is insufficient. At least two were required to detect a significant effect of both distance from the boundary and number of crops. In other words, the statistical power was increased by adding more controls. The optimum ratio of controls to raided farms is 4:1 (Woodward 2005). Nevertheless two or perhaps three controls are probably more practical options for fieldworkers with limited resources.

We conclude that when planning a study of crop raiding the retrospective design will be the most cost-effective method. One should aim to have 30 raided farms, each matched against at least two intact farms. A total of 90 farms (that is 30 cases and 60 controls) will require intense work at the end of the growing season, but anything less may lack the statistical power to detect important effects. On the other hand, a prospective design will provide more information. For example, prospective surveys show the proportion of farms that are afflicted by elephants. Therefore prospective surveys conducted in different years will yield information on whether or not the problem is getting better or worse. This is something you cannot estimate from the retrospective design used here. Nevertheless the prospective design is very much more expensive.

Farms around the KCA were scattered through the bush which, being secondary growth (or "farmbush"), was very dense and for the most part impenetrable. One could not walk easily from farm to farm. Hence taking a completely random sample of farms around the KCA would have been very time consuming. In order to cut travel costs it was more practical to select a number of villages at random, and then select farms around each village. The same is probably true of most study sites in the forest zone. However, all the farms around a particular village may share a characteristic

that influences their risk of attack by elephants. These farms may be more alike than a sample of farms selected completely at random from the whole study area. In other words, there is a degree of correlation between the farms within each village or cluster. There are therefore two sources of variation: variability between farms in the same village, and variability between villages (Shoukri and Chaudhary 2007). Failing to account for this will underestimate the true standard error of the regression coefficient. This increases the risk of Type I error, the risk of rejecting the null hypothesis when it is true (Shoukriand Chaudhary 2007), or recording a significant effect when in fact there is none. Matching is the simplest way to deal with this problem. However, there may be situations where you cannot find enough intact farms to match with the raided ones in each cluster. In that case, one can take a random sample of intact farms from the communities. Then one should use hierarchical models—also called mixed-effects models or random-effects models—that are designed to account for the variation between farms and between communities (Shoukri and Chaudhary 2007). However, this type of analysis becomes quite complicated especially if there are numerous covariates.

Carefully planned studies will show which features of the farming landscape are most likely to draw elephants. Then wildlife managers can advise farmers on how to reduce the attractiveness of their farms. At Kakum it was obvious before we started that the farms adjacent to the boundary were most likely to be raided. But with the present analysis we know that, once you have adjusted for proximity to the park, the number of crop types is an important predictor too.

Acknowledgements

The field work formed part of the Elephant Biology and Management Project organized by Conservation International and the Ghana Wildlife Division. The field work was funded by Conservation International, the Center for Applied Biodiversity Science, the United States Fish & Wildlife Service (African Elephant Conservation Fund), the Smart Family Foundation and the Betlach Family Foundation. Brent Bailey inspired us throughout. The study would not have been possible without the effort and cooperation of the farmers in the field.

References

Afifi AA, Kotlerman JB, Ettner SL, Cowan M. 2007. Methods for improving regression analysis for skewed continuous or counted responses. *Annual Review of Public Health* 28:95–111.

Barnes RFW. 2008. The design of crop–raiding studies. *Gajah* 28:4–7.

Barnes RFW. 2009. The analysis of data from studies of crop-raiding. *Gajah* 30:19–23.

Barnes RFW, Boafo Y, Nandjui A, Farouk UD, Hema EM, Sanquah E, Manford M. 2003. An overview of crop raiding by elephants around the Kakum Conservation Area: Part 2: Technical report. Elephant Biology & Management Project, Africa Program, Conservation International, Washington DC.Unpublished.

Barnes RFW, Dubiure UF, Danquah E, Boafo Y, Nandjui A, Hema EM, Manford M. 2006. Crop-raiding elephants and the moon. *African Journal of Ecology* 45:112–115.

Barnes RFW, Hema EM, Nandjui A, Manford M, Dubiure U–F, Danquah E, Boafo Y. 2005. Risk of crop raiding by elephants around the Kakum Conservation Area, Ghana. *Pachyderm* 39:19–25.

Boafo Y, Dubiure, U–F, Danquah E, Manford M, Nandjui A, Hema EM, Barnes RFW, Bailey B. 2004. Long-term management of crop raiding by elephants around Kakum Conservation Area in southern Ghana. *Pachyderm* 37:68–72.

Gordis L. 2004. *Epidemiology*, 3rd ed. Elsevier Saunders, Philadelphia, Pennsylvania.

Hilbe JM. 2015. *Modeling count data*. Cambridge University Press, New York.

Hosmer DW, Lemeshow S. 2000. *Applied logistic regression*. John Wiley & Sons, New York.

Manly BFJ. 2001. *Statistics for environmental science and management*. Chapman & Hall/CRC, Boca Raton, Florida.

McCullagh P, Nelder JA. 1989. *Generalized linear models*, 2nd ed. Chapman & Hall, New York.

Shoukri MM, Chaudhary MA. 2007. A*nalysis of correlated data* with SAS and R. Chapman & Hall/CRC, Boca Raton, Florida.

Woodward M. 2005. *Epidemiology: Study design and data analysis*. Chapman & Hall/CRC, Boca Raton, Florida.