ENVIRONMETRICS
Environmetrics 2007; **18**: 303–314
Published online 20 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.830

Semiparametric zero-inflated Poisson models with application to animal abundance studies

Monica Chiogna[1]*,† and Carlo Gaetan[2]

[1]*Dipartimento di Scienze Statistiche, Università di Padova, Italy.*
[2]*Dipartimento di Statistica, Venezia, Italy.*

SUMMARY

This paper describes a framework for flexibly modeling zero-inflated data. Semiparametric regression based on penalized regression splines for zero-inflated Poisson models is introduced. Moreover, an EM-type algorithm is developed to perform maximum likelihood estimation. As an illustration, a study of animal abundance is tackled. In fact, abundance often shows excess of zeroes and is a complicated function of the explanatory variables. In particular, the relationships between avian abundance and environmental variables indicating land use are tackled. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: generalized additive models; penalized regression splines; EM algorithm; land use

1. INTRODUCTION

Very often, environmental and ecological data exist only in the form of counts. Not uncommonly, such counts contain a number of zeroes which is greater than the number that would be predicted using standard, unimodal count distributions. An example of such data comes from animal abundance studies, where data for uncommon or rare species come in the form of counts of abundance that contain a high number of zeroes. Such data are generally referred to as zero-inflated (ZI) and require specialized methods for statistical analysis (see Martin *et al*. (2005) and Warton (2005) for two recent discussions on different models).

Generally speaking, ZI data are common in the applications in which it is possible to imagine that the data generating process moves back and forth between a first state where no events are observed and a second state where the events follow, say, a Poisson distribution, (see, Lambert, 1992). A natural way to model such data is to put a point mass ω at 0. That is, with probability ω, we sample a degenerate distribution at 0 and with probability $(1 - \omega)$ we sample, say, a Poisson(λ) distribution. Such models are called zero-inflated Poisson (ZIP) models in the literature. Cohen (1963) and Johnson and Kotz (1969) discuss ZIP models without covariates. Lambert (1992) employs the ZIP model in a regression set up.

*Correspondence to: M. Chiogna, Dipartimento di Scienze Statistiche, Università di Padova, Italy.
†E-mail: monica@stat.unipd.it

Received 4 September 2006
Accepted 19 December 2006

In this paper, we develop semiparametric modeling of ZIP data. This relates to the work of Barry and Welsh (2002), who nonparametrically model ZIP data by following a two-step process. First, they model whether the counts are zero or not. They then model the observed counts, conditional on the counts being greater than 0. In both steps, they use a Generalized Additive model. In our approach, the two-steps process is overtaken by contemporary modeling the two components of the mixture via nonparametric or semiparametric functions. Our proposal is to construct the nonparametric components of the regressor using penalized regression splines, so that inference can be accomplished by making use of the EM algorithm.

We shall illustrate this idea by analyzing abundance of a songbird, *Indigo Bunting*. Although motivated in the present paper by the need of modeling avian abundance, our proposed approach is entirely generic and can be useful in various applications dealing with an excess of zeroes. The structure of the paper is as follows. The problem and the data available for our illustrative example will be illustrated in Section 2. Section 3 will depict the state of art about parametric zero-inflated modeling, whereas Section 4 will present the extension to a semiparametric setting. The last two sections of the paper cover the example and summarize our findings.

2. MOTIVATING EXAMPLE

Land transformation is the most prominent component of human-induced global change. An important consequence of land use change is habitat change, which affects various ecological processes at various spatial scales. Songbirds have served as model organisms in a number of studies investigating the consequences of human-induced land transformation, as abundance of birds varies according to the proportion of suitable habitat in the landscape. In such studies, the relationship between songbirds abundance and local and landscape habitat variables is explored and used as a measure of human-induced changes to the environment.

The abundance data that we consider come from a study aimed at determining the landscape parameters and scales that best predict the number of neotropical migrant bird species along a gradient of anthropogenic disturbance in the Southern Piedmont region of the southeastern United States (Strathford and Robinson, 2005). The study species were surveyed using the protocol established by the North American Breeding Bird Survey. We quote the main traits of the study, as described in Strathford and Robinson (2005); we refer the reader to the above mentioned study for more details on the survey and the study area.

> Thirteen routes of 40–50 points (each point spaced at 0.8-km intervals) were selected along lightly traveled roads. Each route was surveyed twice between June 1 and July 17 and surveyed between 0.5 h before sunrise and 1100 h. All species heard or seen were recorded in 3 min visits to each counting point. Within 100th m, distance of the bird to the observer was estimated within 5 m. Beyond 100th m, distances were estimated to the nearest 50 m. Counts were done on days with little or no wind or rain. Points were not included in the analysis if noise levels from traffic were excessive. Circular buffers of 100-, 200-, and 1000 m around each georeferenced (15-m resolution) point were clipped from a 30-m resolution Landsat 7 TM LULC image taken March 2002. Landscape attributes, such as percent pine, in each buffer were estimated using FRAGSTATS software.

A total of 624 points were surveyed between June 1 and July 17 in years 2002 and 2003. Seventy-eight points were removed from the database because their buffers fell outside the study area. A map of the location of the countpoints is represented in Figure 1.

DOI: 10.1002/env

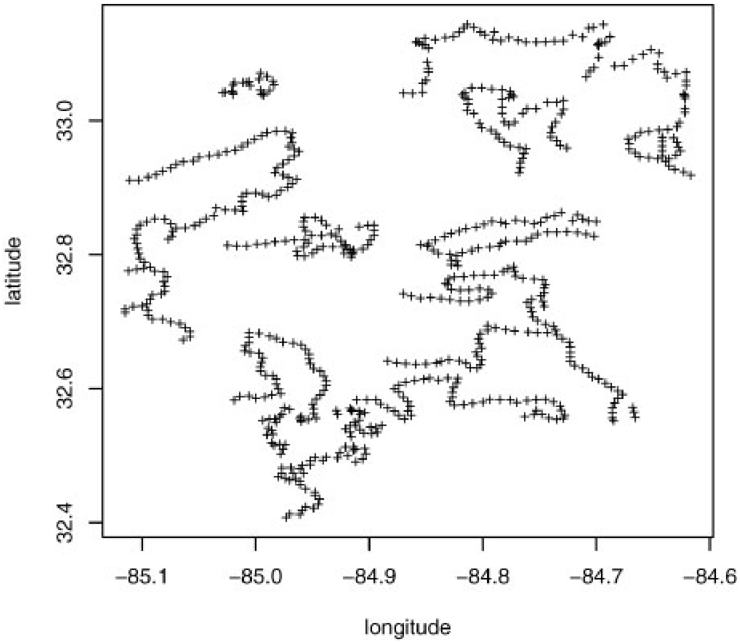

Figure 1. Map of the countpoints

We shall consider as an illustration a somewhat simpler formulation of the general problem tackled by Strathford and colleagues. Among the 13 species considered in the study, we shall concentrate on *Indigo Bunting*. Five land use predictors are considered for their likely relevance in acting on the distribution of individuals, that is, on species abundance. In particular, the following composition variables (in percentage) measured within 1 km around the count points are examined: natural woodland, denoted by M, open parks, hayfields, pasture, denoted by G, early successional forests, denoted by T, impervious surfaces (roads, housing), denoted by U, and pine plantations, denoted by P.

The observed distribution of the counts over the study period is shown in Figure 2. The observed frequency (in percentage) of each count is compared with the frequency expected under a standard Poisson model having mean equal to the observed mean of counts. The plot shows an evident lack of fit of the Poisson model for the first two counts, that is, for zero and one. This might be an indication of zero inflation, although a formal test is called for before deciding on significance of the departure of the data from the theoretical model.

The relationship between species abundance and landscape attributes has been explored using a generalized additive model and assuming a Poisson error distribution. All predictors were fitted using smooth terms, with the number of degrees of freedom chosen using a cross validation procedure. The analysis revealed statistical significance of all the variables apart from P; moreover, it showed the importance of flexibly modeling U, G, and T, as shown by Figure 3, which reports the estimated nonparametric components for the semiparametric model having a linear effect for M and nonlinear effects for U, G, and T.

DOI: 10.1002/env

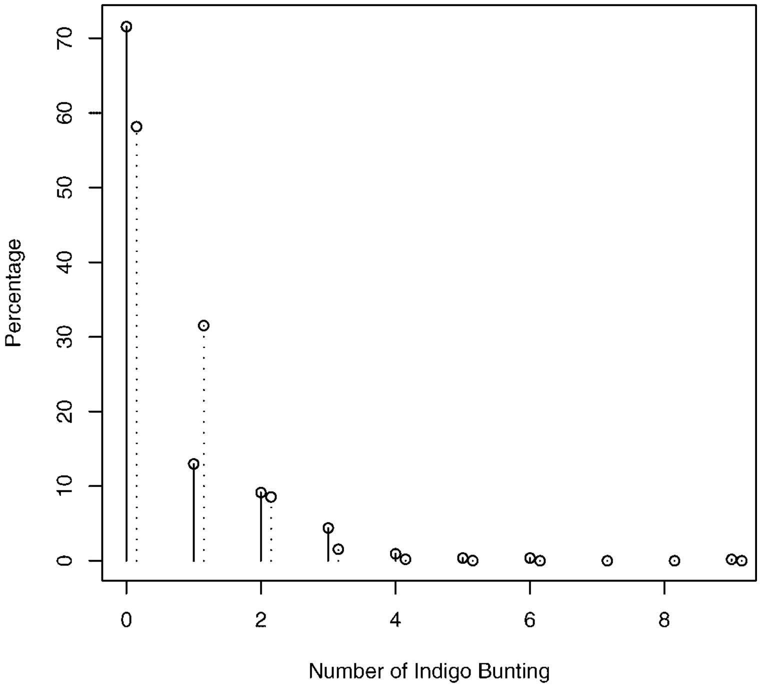

Figure 2. Observed (solid lines) and expected (dashed lines) frequencies, in percentage

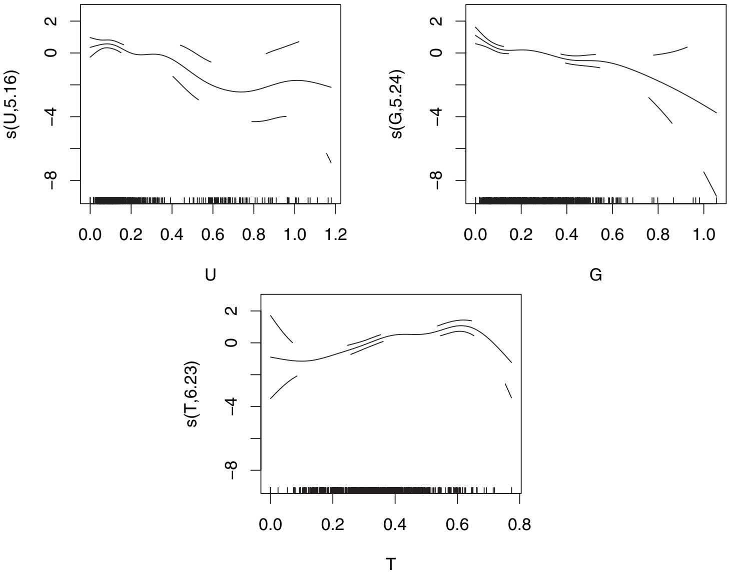

Figure 3. Estimated nonlinear effects for *U*, *G*, and *T*

DOI: 10.1002/env

3. ZERO-INFLATED POISSON MODELS

Approaches to zero-inflated models have been extensively reviewed by Ridout *et al.* (1998). In particular, it is important to emphasize the distinction between the two stage approach and the mixture approach. The former approach separates zeroes from nonzeroes and models first the presence/absence of counts, and then, conditional on presence, models the nonzero counts by a truncated discrete distribution. The latter approach classifies the zeroes into two different groups and creates a mixture between a degenerate distribution and a count distribution, or, more generally, between two count distributions.

The choice of the most appropriate approach for a particular problem is crucial. For example, the mixture assumption is often unrealistic, as there are no clear means of logically separating the zeroes into different classes. On the other hand, the mixture modeling approach unifies the zeroes on a common probability scale allowing for simultaneous inference.

In this paper, we shall primarily be concerned with the mixture approach, as this appears to reflect the data generating mechanism of the abundance data that we like to analyze. In what follows, we shall formally describe the generic structure of a ZIP regression model in the mixture approach.

The ZIP distribution is defined as

$$p(y;\omega,\lambda)=\begin{cases}\omega+(1-\omega)\exp(-\lambda) & y=0\\ (1-\omega)\exp(-\lambda)\lambda^{y}/y! & y=1,2,\ldots\end{cases} \tag{1}$$

where $0\leq\omega\leq 1$ and $\lambda>0$. In this formulation, the quantity ω plays the role of a mixing proportion. In fact, we can see Equation (1) as a special case of the 2-component finite mixture model:

$$p(y;\omega,\lambda)=(1-\omega)p_0(y;\lambda)+\omega p_1(y),$$

where $p_0(y;\lambda)=\exp(-\lambda)\lambda^{y}/y!$, $p_1(y)=\chi_{\{0\}}(y)$ and χ_A is the indicator function of the set A.

Let (y_i,x_i), $i=1,\ldots,n$, be a random sample of size n from the model (1), with x_i being a vector of $p+q$ covariates. In the standard formulation of ZIP regression models, both ω and λ are considered to be functions of the covariates. More formally, $\omega=\omega_i=\omega(x_i)$ and $\lambda=\lambda_i=\lambda(x_i)$, $i=1,\ldots,n$. The parametric approach traces back to Lambert (1992), who proposed the use the logit and log-linear link to model ω_i and λ_i, respectively, that is,

$$\log\left(\frac{\omega_i}{1-\omega_i}\right)=\sum_{j=1}^{q}\beta_j x_{ij},\quad \log\lambda_i=\sum_{j=q+1}^{p+q}\beta_j x_{ij},\quad i=1,\ldots,n, \tag{2}$$

where β_j, $j=1,\ldots,p+q$, are unknown parameters to be estimated. In this formulation, the first q covariates of x_i are not necessarily different from the last p covariates.

The finite mixture form of the distribution suggests that maximum likelihood estimation can be accomplished by employing the EM algorithm (Dempster *et al.*, 1977; Lambert, 1992). In fact, let Z_i, $i=1,\ldots,n$, be n unobserved Bernoulli variables such that $Pr(Z_i=1)=\omega_i$, indicating from which component the unit sample (y_i,x_i) comes. The 'complete-data' log-likelihood is given by

$$l^{C}(\beta)=\sum_{i=1}^{n}\left\{z_i\log\left(\frac{\omega_i}{1-\omega_i}\right)+\log(1-\omega_i)+(1-z_i)\log p_0(y_i;\lambda_i)\right\}.$$

DOI: 10.1002/env

The EM algorithm maximizes the log-likelihood by iteration of two steps: the E-step and the M-step.

At stage k, given a current estimate $\beta^{(k-1)}$, the E-step computes the conditional expectation of $l^C(\beta)$ given $(y_1, \dots, y_n)^T$, that is, the quantity

$$S(\beta, \beta^{(k-1)}) = E\{l^C(\beta)|y_1, \dots, y_n; \beta_{k-1}\}.$$

The M-step chooses $\beta = \beta^{(k)}$ so as to maximize $S(\beta, \beta^{(k-1)})$.

The E-step is simple because the random variable Z_i given y_i is a Bernoulli random variable with probability of success

$$w(y_i; \beta) = \frac{\omega_i p_1(y_i)}{(1-\omega_i)p_0(y_i; \lambda_i) + \omega_i p_1(y_i)}.$$

In the M-step, it is easy to see that maximizing the quantity

$$\sum_{i=1}^{n} \left\{ w(y_i; \beta^{(k-1)}) \log\left(\frac{\omega_i}{1-\omega_i}\right) + \log(1-\omega_i) \right\}$$

is equivalent to maximizing a weighted log-likelihood of a binomial regression model. On the other side, maximizing the quantity

$$\sum_{i=1}^{n} (1 - w(y_i; \beta^{(k-1)})) \log p_0(y_i; \lambda_i)$$

is equivalent to maximizing a weighted log-likelihood of a Poisson regression model. This allows to use standard routines for GLMs to perform the maximization task and reduces the computational effort required by the M-step of the algorithm quite substantially. This offers an additional strong motivation for adopting such maximization strategy.

4. SEMIPARAMETRIC ZIP MODELS

We extend model (2) by allowing semiparametric regression terms for both components of the model, that is,

$$\log\left(\frac{\omega_i}{1-\omega_i}\right) = \sum_{j=1}^{q} h_j(x_{ij}), \quad \log \lambda_i = \sum_{j=q+1}^{p+q} h_j(x_{ij}), \quad i = 1, \dots, n,$$

where each h_j can be either a linear effect or an unspecified ('non-parametric') function. In order to estimate the nonparametric components of the regressors, we resort on penalized regression splines. Following Wood (2000) and Wood and Augustin (2002), we choose $\{x^*_{jk}, k = 1 \dots, m_j\}$, that is, a set of points (knots) in the range of x_j, and let $b_{jk}(x) = |x - x^*_{jk}|^3$ for $k = 1, \dots, m_j$, with $b_{j,m_j+1}(x) = 1$ and $b_{j,m_j+2}(x) = x$. The function $h_j(x_j)$ is represented by a linear combination of the basis functions

DOI: 10.1002/env

$b_k(x_j)$,

$$h_j(x_j) = \sum_{k=1}^{m_j+2} \alpha_{jk} b_{jk}(x_j), \qquad j = 1, \ldots, p, \tag{3}$$

where the α_{jk} are unknown coefficients. The function $h_j(x_j)$ is a natural cubic spline if the linear constraints

$$\sum_{k=1}^{m_j} \alpha_{jk} = 0, \qquad \sum_{k=1}^{m_j} \alpha_{jk} x^*_{jk} = 0, \tag{4}$$

are satisfied. Estimating h_j is now equivalent to estimating the parameters α_{jk} subject to linear constraints. It is straightforward to see that this brings the problem back to a standard parametric model fitting problem, once the basis functions $b_k(x_j)$ have been chosen. We control the number of the basis functions, that is, the number of knots, by the penalty function

$$J(h_j) = \int [h''_j(x)]^2 dx = \alpha_j^T H_j \alpha_j, \tag{5}$$

where $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{j,m_j+2})^T$ and $H_j = (h^j_{kl})_{k=1,m_j+2,\ l=1,m_j+2}$ with $h^j_{kl} = \int b_{jk}(x) b_{jl}(x) dx$.

Using the basis (3), the penalty (5) and rearranging the parameters vector in a unique vector α, the optimization problem becomes:

$$\text{maximize} \qquad l(\alpha) - \sum_{j=1}^{m+q} \psi_j \alpha_j^T H_j \alpha_j$$

$$\text{subject to} \qquad C\alpha = 0,$$

where $l(\alpha)$ is the log-likelihood function and the rows of C represent the linear constraints (4), the identifiability constraints being $\sum_{i=1}^n h_j(x_{ij}) = \sum_{i=1}^n \sum_{k=1}^{m_j+2} \alpha_{jk} b_{jk}(x_{ij}) = 0$ (see e.g., Wood and Augustin, 2002).

We can re-write the constrained maximization problem as an unconstrained one. Suppose that C is a $c \times d$ matrix, $c < d$. By a QR or a Householder decomposition, it is possible to find an orthogonal matrix Q such that

$$CQ = [0_{c,d-c}, T],$$

where T is a $c \times c$ matrix and $0_{c,d-c}$ is a $d \times (d-c)$ matrix. We partition Q into two parts: $Q = [U, V]$. U is a $c \times (d-c)$ matrix and V is a $d \times c$ matrix, such that $CV = 0$, $CV = T$. Setting $\alpha = U\theta =$

DOI: 10.1002/env

$[U_1, \ldots, U_{m+q}][\theta_1^T, \ldots, \theta_{m+q}^T]^T$, the objective function can be re-written as

$$pl(\theta) = l(\theta) - \sum_{j=1}^{m+q} \psi_j \theta_j^T U_j^T H_j U_j \theta_j. \tag{6}$$

The EM algorithm still applies with a slight modification in the M-step, where we maximize

$$S(\theta, \theta^{(k-1)}) - \sum_{j=1}^{m+q} \psi_j \theta_j^T U_j^T H_j U_j \theta_j.$$

A simple application of the Jensen inequality shows that the sequence $pl(\theta^{(k)})$ is not decreasing. We denote by $\hat{\theta}$ the final estimate when a convergence criterion is met. The estimated covariance matrix for $\hat{\theta}$, $V(\hat{\theta})$, can be obtained using the same argument as in Wood (2004) by considering the negative of the Hessian of the penalized likelihood (6), namely

$$V(\hat{\hat{\theta}}) = -\frac{\partial^2}{\partial\theta\partial\theta^T} l(\theta) + \sum_{j=1}^{m+q} \psi_j U_j^T H_j U_j.$$

5. THE APPLICATION

Although the exploratory analysis in Section 2 suggested a lack of fit of the Poisson distribution on the smallest counts, the question remains whether such departure is due to zero inflation or not. To test such hypothesis, we resorted on the score test (van den Broek, 1995)

$$T = \left\{ \sum_{i=1}^{n} \frac{\chi_{\{0\}}(y_i) - \hat{p}_i}{\hat{p}_i} \right\}^2 \Bigg/ \left\{ \sum_{i=1}^{n} \frac{1 - \hat{p}_i}{\hat{p}_i} - n\bar{y} \right\},$$

where $\hat{p}_i = \Pr\{Y_i = 0\} = \exp(-\hat{\lambda}_i)$ and λ_i is the mean estimate under the Poisson GAM model. Under the null hypothesis, T would have an asymptotic χ^2 distribution with 1 degree of freedom (van den Broek, 1995), although the asymptotic result needs to be checked, as regularity conditions fail. Simulation studies might be performed to investigate the finite sample properties of the test statistic. The observed value was $T_{\text{obs}} = 65.133$, leading to the rejection of the null hypothesis.

Having tested the presence of zero inflation, we proceeded with the search for a good model. Semiparametric models assuming a ZIP error distribution were fitted using R routines (R Development Core Team, 2005) written by ourselves. Fitting of these models required specification of two predictors, the first controlling the mixing proportion ω and the second regulating the mean λ of the Poisson counts.

In an initial evaluation, all predictors were included in the two regressors using smooth terms. It is important to remark that, when modeling with basis functions, it is possible to control smoothness of the nonparametric components by selecting an appropriate number of basis functions for each term. Nevertheless, the control can be difficult. A sufficiently high number of basis functions for each unknown

Table 1. Final model: estimates for the parametric components

Component	Term	Estimate	SE
ω	Constant	-0.1005850	0.2532877
	G	1.9153162	0.8275954
λ	Constant	0.6270252	0.1310143
	U	-3.3491786	0.8695267

underlying function h_j allows to closely approximate the function, at the risk, although, of overfitting the data. On the other hand, a small number of basis functions likely avoids overfitting, but it can achieve a very poor approximation of the true h_j.

The penalized approach allows to alleviate the problem of selecting the number of basis functions by allowing to use a relatively high number of basis functions and avoiding overfitting by applying a penalty to the objective function. In this perspective, we selected a relatively high number of basis functions (degrees of freedom) for each smooth term in the model, that is, 200, controlling overfitting by regulation of the associated penalties. Control of the penalties was achieved through a grid search by monitoring the value of the penalized likelihood.

By implementing a backward selection method, terms were removed from the intermediate models whose confidence region for the smooth functions everywhere included zero. If the estimated smooth functions were straight lines, so that it is was not possible to simplify the model by further smoothing, linear terms were introduced in the models. The decision about removing linear terms was based on approximated tests on the coefficients.

The final model is summarized in Table 1, which shows the statistically significant parametric terms in each component of the model.

Both regressors included also a nonparametric term for the effect of the early transitional habitat; the corresponding estimated functions are shown in Figure 4.

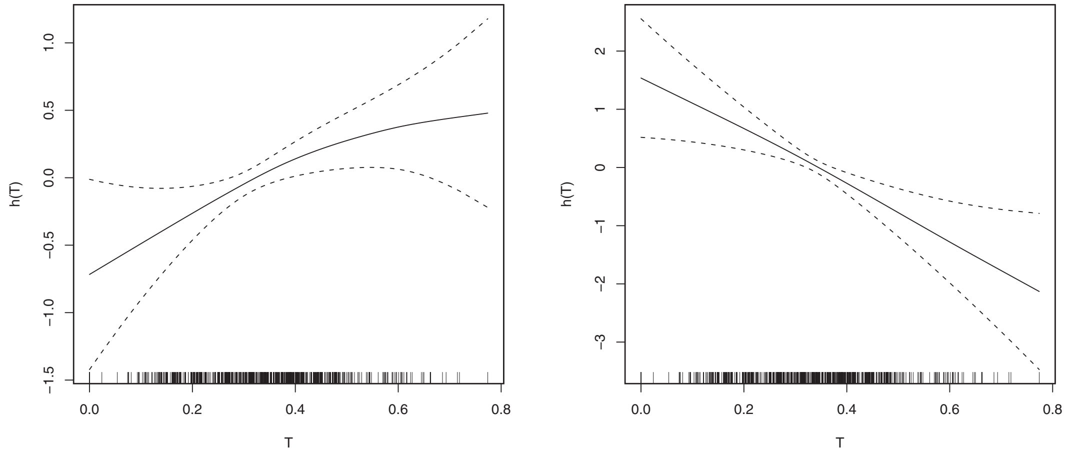

Figure 4. Estimated nonparametric effects of T in the λ component (left plot) and in the ω component (right plot) with approximated 95% confidence regions

DOI: 10.1002/env

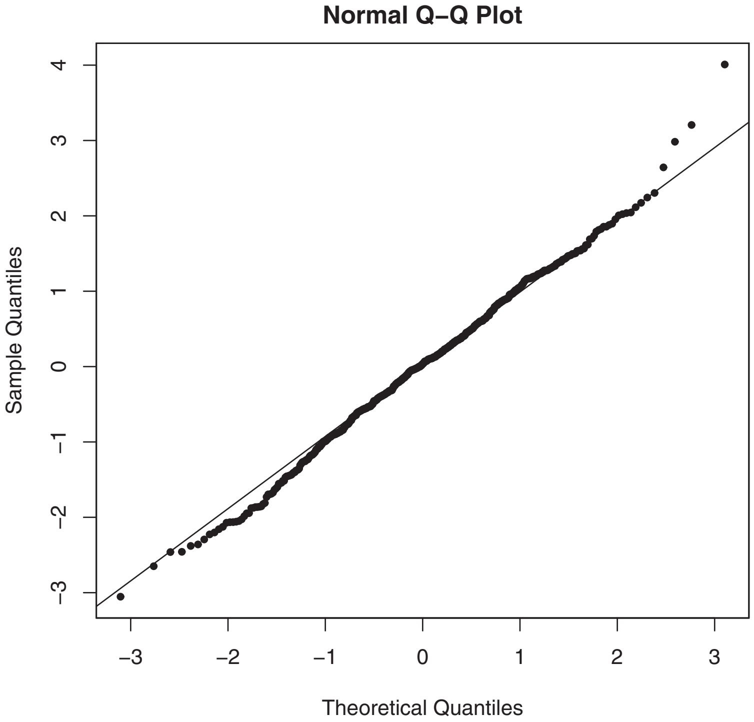

Figure 5. Diagnostic quantile plot

To assess the overall adequacy of the model, graphical residual analysis might be carried out. However, computation of residuals deserves some care. We considered the following randomized residuals (Dunn and Smyth, 1996)

$$v_i = \Phi^{-1}(r_i),$$

where $r_i = (1 - u_i)F(y_i - 1; \hat{\omega}_i, \hat{\lambda}_i) + u_i F(y_i; \hat{\omega}_i, \hat{\lambda}_i)$, u_i is a uniform random variable, Φ is the standard normal distribution function and $F(y; \omega, \lambda)$ is the ZIP distribution function. Note that randomization allows to achieve continuous residuals even if the response variable is discrete. It is easily seen that the randomized residuals are exactly normal, apart from sampling variability in the estimated parameters. Moreover, they are the only useful residuals for count data when the response takes on only a small number of distinct values. An examination of Figure 5 revealed no evident departure from normality.

Overall, results provided strong support about the importance of urban habitat and early transitional habitat as determinants of *Indigo Bunting* species richness. In particular, richness is higher in zones with less impervious surfaces (U) and more early successional forests (T). Although *Indigo Bunting* is primarily a forest bird, the results seem to confirm the preference of *Indigo Bunting* for edge habitat, where brushy fields meet the forest. The probability of being in an area not visited by *Indigo Bunting* increases with the increase of open parks, and decreases by increasing the coverage of early successional habitat. In conclusion, species richness appears greater along bordered than nonbordered transects.

DOI: 10.1002/env

6. CONCLUSIONS

In this paper, we have provided an introduction to the framework required for semiparametric analysis of ZI data, and illustrated this framework with an ecological example.

The extension of parametric models to a semiparametric setting based on penalized regression splines has proved to be effective and efficient. Although not developed here, it is worth noting that the proposed approach can be extended to spatial analysis by exploiting the interplay between spatial processes on $\mathbb{R}^d$ and thin plate splines (see, e.g., Nychka, 2000). The EM algorithm provided a relatively straight-forward mean to carry out inference. We have implemented an ad hoc backward selection modeling strategy, which has proved to be useful. Although such model selection, which combines a direct grid search over some penalty space of values and ad hoc evaluation of significance of the variables, is probably valuable in other practical modeling situations, it has to be said that it can become very costly as the number of smoothing terms increases.

Results from our analysis are consistent with those from various recent studies that have highlighted the need for procedures capable of fitting nonlinear relationships when analyzing environmental and ecological datasets. With respect to the study of abundance of *Indigo Bunting*, the semiparametric setting improves ecological interpretability of the role played by the regressors. In particular, our analysis provides clear support for the relevance of the early transitional habitat in explaining patterns of the species' richness in a gradient of anthropogenic disturbance. Use of these results in the evaluation of human-induced changes to the environment remains an open problem, whose solution would probably require consideration of time-varying nonlinear effects, for which appropriate statistical ZIP models need to be developed.

ACKNOWLEDGEMENTS

This paper was supported by MIUR under grant 134337, 2002 and grant 137478, 2004. The authors thank Jeffrey A. Strathford (Department of Biological Sciences, Auburn University, Auburn) for kindly providing the data and sharing with us his studies on the theme.

REFERENCES

Barry SC, Welsh A. 2002. Generalized additive modeling and zero inflated count data. *Ecological Modelling* **157**: 179–188.

Cohen AC. 1963. Estimation in Mixtures of discrete distributions. *Proceedings of the International symposium on Discrete Distributions*, Montreal, Quebec; 373–378.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological* **39**: 1–22.

Dunn PK, Smyth GK. 1996. Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5**: 236–244.

Johnson NL, Kotz S. 1969. *Distributions in Statistics: Discrete Distribution*, Houghton Mifflin: Boston, Massachusetts.

Lambert D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**: 1–14.

Martin TG, Wintle BA, Rhodes JR, Kuhnert PM, Field SA, Low-Choy SJ, Tyre AJ, Possingham HP. 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters* **8**: 1235–1246.

Nychka D. 2000. Spatial process estimates as smoothers. In *Smoothing and Regression. Approaches, Computation and Application*, Schimek MG (ed) John Wiley: New York; 393–424.

R Development Core Team. 2005. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing: Vienna, Austria.

Ridout MS, Demetrio CGB, Hinde JP. 1998. Models for count data with many zeros. *Proceedings of XIXth International Biometric Society Conference*, IBC98, Cape Town; 179–192.

Strathford JA, Robinson WD. 2005. Distribution of neotropical migratory bird species across an urbanizing landscape. *Urban Ecosystems* **8**: 59–77.

van den Broek J. 1995. A score test for zero inflation. *Biometrics* **51**: 738–743.

DOI: 10.1002/env

Warton DI. 2005. Most multivariate abundance data do not have extra zeros, compared to the negative binomial. *Environmetrics* **16**: 275–289.

Wood S, Augustin N. 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling* **157**: 157–177.

Wood SN. 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B, Methodological* **62**: 413–428.

Wood SN. 2004. On confidence intervals for GAMs based on penalized regression splines. *Technical Report 04-12*, Department of Statistics-University of Glasgow, Scotland.