Modelling zero-inflated continuous data: with applications to changes in the number of birds visiting gardens in winter

Ben Swallow*, 1, Steve Buckland 1, Ruth King 1, Mike Toms 2

1CREEM, School of Maths and Statistics, University of St Andrews

2British Trust for Ornithology, Thetford

ISEC, Montpellier, July 3rd 2014
Outline

1 Motivation
   - Statistical motivation
   - Ecological motivation

2 Modelling framework

3 Model implementation

4 Results
   - Standard model
   - Change-change model

5 Conclusions
Statistical motivation

- Zero-inflation common problem in ecology
- Most of the extensions to standard models to incorporate zero-inflation are with discrete distributions
- Analysis of ecological datasets often poses similar problems but with continuous data
- How do we deal with non-negative continuous data when it is zero-inflated?
Models to explain changes in ecological populations can take many forms.

Log-linear models are often used with environmental covariates to explain changes in observed counts. These covariates can enter the model in many different forms.

Are results consistent?
Modelling changes in ecological populations

- Models to explain changes in ecological populations can take many forms.
- Log-linear models often used with environmental covariates used to explain changes in observed counts.
Models to explain changes in ecological populations can take many forms.

Log-linear models often used with environmental covariates used to explain changes in observed counts.

These covariates can enter the model in many different forms.
Modelling changes in ecological populations

- Models to explain changes in ecological populations can take many forms.
- Log-linear models often used with environmental covariates used to explain changes in observed counts.
- These covariates can enter the model in many different forms.
- Are results consistent?
The Tweedie distributions

Versatile family of probability distributions belonging to the class of exponential dispersion models, whose variance is given by

$$\text{Var}[Y] = \varphi V(\mu)$$

If

$$y \sim \text{Tw}(\mu, \varphi, p)$$

then

$$\text{Var}[Y] = \varphi \mu^p$$

$\varphi > 0$ is the dispersion parameter and $p / \in (0, 1)$.

Distributions can be discrete or continuous.

Contain many standard distributions as special cases (e.g. Normal, Poisson, gamma, inverse Gaussian).

Compound Poisson-gamma distributions when $1 < p < 2$

Poisson-gamma compound

$$y_{ij} \sim \sum_{i=1}^{N_i} w_{ij}$$

s.t.

$$N_i \sim \text{P}(\lambda_i)$$

and

$$w_{ij} \sim \Gamma(\alpha, \beta_i)$$
The Tweedie distributions

- Versatile family of probability distributions belonging to the class of exponential dispersion models, whose variance $\text{Var}[Y] = \phi V(\mu)$
- If $y \sim Tw(\mu, \phi, p)$ then $\text{Var}[Y] = \phi \mu^p$
- $\phi > 0$ is the dispersion parameter and $p \notin (0, 1)$
- Distributions can be discrete or continuous
The Tweedie distributions

- Versatile family of probability distributions belonging to the class of exponential dispersion models, whose variance $\text{Var}[Y] = \phi V(\mu)$
- If $y \sim \text{Tw}(\mu, \phi, p)$ then $\text{Var}[Y] = \phi \mu^p$
- $\phi > 0$ is the dispersion parameter and $p \notin (0, 1)$
- Distributions can be discrete or continuous
- Contain many standard distributions as special cases (e.g. Normal, Poisson, gamma, inverse Gaussian)
- Compound Poisson-gamma distributions when $1 < p < 2$

Poisson-gamma compound

\[ y_{ij} \sim \sum_{i=1}^{N_i} w_{ij} \text{ s.t. } N_i \sim P(\lambda_i) \text{ and } w_{ij} \sim \Gamma(\alpha, \beta_i) \]
Songbird population change

Species
- Blackbird
- Blue tit
- Chaffinch
- Coal tit
- Collared dove
- Great tit
- Greenfinch
- House sparrow
- Robin
- Sparrowhawk
- Starling
## Sparrowhawk population change

<table>
<thead>
<tr>
<th>Year</th>
<th>Average count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970</td>
<td>0.00</td>
</tr>
<tr>
<td>1990</td>
<td>0.10</td>
</tr>
</tbody>
</table>

![Graph showing the population change of sparrowhawks from 1970 to 1990](image-url)
<table>
<thead>
<tr>
<th>Year</th>
<th>Average count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970</td>
<td>0.00</td>
</tr>
<tr>
<td>1990</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Sparrowhawk population change

Early 70s

Average count

0.00 0.10

1970 1990

Year
Garden bird data

- BTO’s Garden bird feeding survey (GBFS) monitors numbers of bird visiting feeding stations across the country
- Spans winter period (October to March) since 1970/71
- 26 weekly maxima of each species seen feeding on provisioned food
- Averaged over 26 weeks giving an (effectively) continuous variable
Garden bird data

- Blackbird
- Blue Tit
- Col. Dove
- Chaffinch
- Coal Tit
- Greenfinch
- Great Tit
- H. Sparrow
- Robin
- Starling

Frequency vs. Average count graphs for each bird species.
Standard model

\[ y_{i,t} \sim Tw(\mu_{i,t}, \phi, p) \]

\[
\log \left( \frac{\mu_{i,t}}{\mu_{i,t-1}} \right) = \alpha + \mathbf{x}_i^\top \beta + \mathbf{v}_{i,t}^\top \gamma + \epsilon_i
\]

\[ \epsilon_i \sim N(0, \sigma^2) \]

or

\[
\log \left( \frac{\mu_{i,t}}{\mu_{i,t-1}} \right) = \mathbf{v}_{i,t}^\top \gamma + \epsilon_i
\]

\[ \epsilon_i \sim N(\alpha + \mathbf{x}_i^\top \beta, \sigma^2) \]
GBFS dataset also contained site specific data including grid reference (converted to northing/easting) and classification of site as rural or suburban \{-1,1\}

5 km$^2$ resolution average monthly ground frost through UKCP09, matched to the site

Also included number of sparrowhawks, collared dove and year-lagged count of prey (to test for density dependence)
The change-change approach

\[ y_{i,t} \sim Tw(\mu_{i,t}, \phi, p) \]

\[ \mu_{i,t} = \mu_{i,1} \exp \left\{ \log \left( \frac{v_{i,t} + 1}{v_{i,1} + 1} \right)^\top \gamma + \epsilon_i \right\} \]  \hspace{1cm} (2)

\[ \epsilon_i \sim N(\alpha + x_i^\top \beta, \sigma^2) \]

or

\[ \log \left( \frac{\mu_{i,t}}{\mu_{i,t-1}} \right) = \log \left( \frac{v_{i,t} + 1}{v_{i,t-1} + 1} \right)^\top \gamma \]

\(^1\)Newson et al. (2010) *JAE* 47, 244–252
Outline

1 Motivation
   - Statistical motivation
   - Ecological motivation

2 Modelling framework

3 Model implementation

4 Results
   - Standard model
   - Change-change model

5 Conclusions
MCMC methods

- In classical framework, unbiased estimation of $p$ can be difficult and Zhang (2013)$^1$ showed Markov chain Monte Carlo (MCMC) methods provided consistently better estimates.
- Model fitted in hierarchical Bayesian framework using MCMC methods.
- Reversible jump MCMC used to estimate posterior model probabilities to inform covariate selection.
- Can also easily deal with missing data and covariate values.

---

1 Motivation
   • Statistical motivation
   • Ecological motivation

2 Modelling framework

3 Model implementation

4 Results
   • Standard model
   • Change-change model

5 Conclusions
Results: standard model

Table: Posterior means of regression parameters for model 1. Posterior means are given with marginal posterior probabilities below in brackets. Significant covariates are highlighted in green (+ve) or red (-ve).

<table>
<thead>
<tr>
<th>Species</th>
<th>Intercept</th>
<th>North</th>
<th>East</th>
<th>Rur/sub</th>
<th>Dens. dep.</th>
<th>S/hawk</th>
<th>C. dove</th>
<th>Frost</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collared Dove</td>
<td>0.0185</td>
<td>-0.0016</td>
<td>0.0099</td>
<td>-0.0091</td>
<td>-0.0466</td>
<td>-0.0043</td>
<td>-</td>
<td>0.0096</td>
<td>0.0129</td>
</tr>
<tr>
<td>Blackbird</td>
<td>0.0110</td>
<td>0.0049</td>
<td>-0.0069</td>
<td>-0.0077</td>
<td>-0.0328</td>
<td>0.0103</td>
<td>0.0153</td>
<td>0.0119</td>
<td>0.0032</td>
</tr>
<tr>
<td>Robin</td>
<td>0.0044</td>
<td>0.0012</td>
<td>-0.0046</td>
<td>-0.0095</td>
<td>-0.0291</td>
<td>0.0049</td>
<td>0.0030</td>
<td>0.0009</td>
<td>0.0013</td>
</tr>
<tr>
<td>Blue tit</td>
<td>-0.0205</td>
<td>-0.0011</td>
<td>-0.0044</td>
<td>-0.0074</td>
<td>-0.0142</td>
<td>-0.0077</td>
<td>-0.0017</td>
<td>0.0077</td>
<td>0.0019</td>
</tr>
<tr>
<td>Coal tit</td>
<td>-0.0118</td>
<td>0.0125</td>
<td>-0.0168</td>
<td>0.0006</td>
<td>-0.0243</td>
<td>0.0135</td>
<td>0.0069</td>
<td>0.0034</td>
<td>0.0040</td>
</tr>
<tr>
<td>Great tit</td>
<td>-0.0099</td>
<td>-0.0007</td>
<td>-0.0010</td>
<td>-0.0090</td>
<td>-0.0188</td>
<td>0.0000</td>
<td>-0.0018</td>
<td>0.0045</td>
<td>0.0018</td>
</tr>
<tr>
<td>House sparrow</td>
<td>-0.0541</td>
<td>-0.0118</td>
<td>-0.0262</td>
<td>-0.0116</td>
<td>-0.0003</td>
<td>-0.0369</td>
<td>0.0005</td>
<td>0.0404</td>
<td>0.0106</td>
</tr>
<tr>
<td>Starling</td>
<td>-0.0594</td>
<td>-0.0080</td>
<td>-0.0008</td>
<td>0.0008</td>
<td>-0.0014</td>
<td>-0.0332</td>
<td>0.0080</td>
<td>0.0429</td>
<td>0.0082</td>
</tr>
<tr>
<td>Chaffinch</td>
<td>0.0045</td>
<td>0.0041</td>
<td>-0.0141</td>
<td>-0.0155</td>
<td>-0.0249</td>
<td>0.0035</td>
<td>0.0088</td>
<td>0.0290</td>
<td>0.0063</td>
</tr>
<tr>
<td>Greenfinch</td>
<td>-0.0243</td>
<td>0.0027</td>
<td>-0.0200</td>
<td>-0.0145</td>
<td>-0.0199</td>
<td>-0.0044</td>
<td>-0.0051</td>
<td>0.0104</td>
<td>0.0102</td>
</tr>
</tbody>
</table>
**Results: change-change model**

**Table:** Posterior means of regression parameters for model 2. Significant covariates are highlighted in green (⁺ve) or red (⁻ve).

<table>
<thead>
<tr>
<th>Species</th>
<th>Intercept</th>
<th>North</th>
<th>East</th>
<th>Rur/sub</th>
<th>S/hawk</th>
<th>C. dove</th>
<th>Frost</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collared Dove</td>
<td>-0.0057</td>
<td>-0.0080</td>
<td>0.0051</td>
<td>-0.0121</td>
<td>-0.0031</td>
<td>-</td>
<td>-0.0270</td>
<td>0.0106</td>
</tr>
<tr>
<td>Blackbird</td>
<td>0.0042</td>
<td>-0.0039</td>
<td><strong>-0.0125</strong></td>
<td>-0.0006</td>
<td>0.0006</td>
<td><strong>0.0070</strong></td>
<td><strong>-0.0109</strong></td>
<td>0.0020</td>
</tr>
<tr>
<td>Robin</td>
<td>-0.0017</td>
<td>0.0010</td>
<td>-0.0021</td>
<td>-0.0007</td>
<td>0.0002</td>
<td>0.0004</td>
<td>-0.0001</td>
<td>0.0008</td>
</tr>
<tr>
<td>Blue tit</td>
<td>-0.0262</td>
<td>0.0038</td>
<td>-0.0032</td>
<td>-0.0050</td>
<td><strong>-0.0096</strong></td>
<td>-0.0009</td>
<td>-0.0096</td>
<td>0.0015</td>
</tr>
<tr>
<td>Coal tit</td>
<td>-0.0235</td>
<td><strong>0.0163</strong></td>
<td>-0.0025</td>
<td>0.0016</td>
<td>0.0050</td>
<td>0.0099</td>
<td>0.0004</td>
<td>0.0027</td>
</tr>
<tr>
<td>Great tit</td>
<td>-0.0186</td>
<td>0.0021</td>
<td>-0.0019</td>
<td>-0.0009</td>
<td>-0.0009</td>
<td>-0.0005</td>
<td>0.0003</td>
<td>0.0013</td>
</tr>
<tr>
<td>House sparrow</td>
<td>-0.0572</td>
<td>-0.0085</td>
<td><strong>-0.0272</strong></td>
<td>-0.0068</td>
<td><strong>-0.0400</strong></td>
<td>-0.0023</td>
<td><strong>-0.0508</strong></td>
<td>0.0132</td>
</tr>
<tr>
<td>Starling</td>
<td>-0.0612</td>
<td><strong>-0.0318</strong></td>
<td>-0.0094</td>
<td>0.0005</td>
<td><strong>-0.0229</strong></td>
<td>0.0015</td>
<td><strong>-0.0886</strong></td>
<td>0.0117</td>
</tr>
<tr>
<td>Chaffinch</td>
<td>0.0006</td>
<td>-0.0054</td>
<td>-0.0047</td>
<td><strong>-0.0097</strong></td>
<td>-0.0006</td>
<td>0.0010</td>
<td><strong>-0.0272</strong></td>
<td>0.0054</td>
</tr>
<tr>
<td>Greenfinch</td>
<td>-0.0287</td>
<td>0.0014</td>
<td><strong>-0.0148</strong></td>
<td>-0.0014</td>
<td>-0.0014</td>
<td>-0.0060</td>
<td>-0.0106</td>
<td>0.0110</td>
</tr>
</tbody>
</table>
**Results**

![Graph showing relative predation risk for various bird species](image)

**Figure:** Combined models (red); standard model only (black); change model only (green)

Conclusions

- Overall increase in sparrowhawks may have led to decreases in the species most susceptible to predation
- Our modelling approach consistent with expected predation effect
- Using a change-change approach can highlight interesting additional impacts on populations whilst remaining consistent in predation effects
- The Tweedie distributions are a highly flexible class of distributions - don’t require strong assumptions *a priori*
- Can accommodate both discrete or continuous data, zero-inflation, over-dispersion
Acknowledgements
Priors

\[ \sigma^2 \sim \Gamma^{-1}(0.001, 0.001) \]
\[ \alpha; \beta; \gamma_{2:4} \sim N(0, 10^{-2}) \]
\[ \gamma_1 \sim HN(0, 10^{-2}) \]
\[ \phi \sim U(0, 5) \]
\[ p \sim U(1, 2) \]
\[ \mu_0 \sim U(0, 200) \]
Posterior distributions

Joint posterior

\[
\pi(\theta, \epsilon, \sigma^2, \mu_0 | y) \propto \mathcal{T}_W (y | \theta, \epsilon, \sigma^2, \mu_0) f(\epsilon | \sigma^2) p(\theta) p(\mu_0) p(\sigma^2)
\]
Posterior distributions

Joint posterior

\[ \pi(\theta, \epsilon, \sigma^2, \mu_0|y) \propto Tw(y|\theta, \epsilon, \sigma^2, \mu_0) f(\epsilon|\sigma^2)p(\theta)p(\mu_0)p(\sigma^2) \]

Posterior conditional for \(\sigma^2\)

\[ \pi(\sigma^2|\epsilon) \sim \Gamma^{-1} \left( \frac{n_{site}}{2} + \alpha_{\sigma}, \frac{\sum_{i=1}^{n_{site}} \epsilon_i^2}{2} + \beta_{\sigma} \right) \]
# Parameter estimates

**Table:** Variance parameter estimates for two model specifications

<table>
<thead>
<tr>
<th>Species</th>
<th>$\phi_{std}$</th>
<th>$p_{std}$</th>
<th>$\phi_{ch}$</th>
<th>$p_{ch}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collared Dove</td>
<td>0.5393</td>
<td>1.3280</td>
<td>0.5980</td>
<td>1.3533</td>
</tr>
<tr>
<td>Blackbird</td>
<td>0.2040</td>
<td>1.2434</td>
<td>0.2065</td>
<td>1.2781</td>
</tr>
<tr>
<td>Robin</td>
<td>0.0638</td>
<td>1.0581</td>
<td>0.0680</td>
<td>1.0673</td>
</tr>
<tr>
<td>Blue tit</td>
<td>0.1683</td>
<td>1.4424</td>
<td>0.1660</td>
<td>1.4728</td>
</tr>
<tr>
<td>Coal tit</td>
<td>0.3323</td>
<td>1.2650</td>
<td>0.3498</td>
<td>1.2742</td>
</tr>
<tr>
<td>Great tit</td>
<td>0.1994</td>
<td>1.1796</td>
<td>0.2050</td>
<td>1.1867</td>
</tr>
<tr>
<td>House sparrow</td>
<td>0.6758</td>
<td>1.3445</td>
<td>0.7433</td>
<td>1.3708</td>
</tr>
<tr>
<td>Starling</td>
<td>0.6982</td>
<td>1.3875</td>
<td>0.7367</td>
<td>1.4093</td>
</tr>
<tr>
<td>Chaffinch</td>
<td>0.3815</td>
<td>1.3634</td>
<td>0.4082</td>
<td>1.3962</td>
</tr>
<tr>
<td>Greenfinch</td>
<td>0.5212</td>
<td>1.4127</td>
<td>0.5380</td>
<td>1.4294</td>
</tr>
</tbody>
</table>