MODELING RATIOS WITH POTENTIAL ZERO-INFLATION TO ASSESS SOIL NEMATODE COMMUNITY STRUCTURE

Joanna Zbylut
Kansas State University

Leigh Murray
Kansas State University

S. H. Thomas
New Mexico State University - Main Campus

J. Beacham
New Mexico State University - Main Campus

J. Schroeder
New Mexico State University - Main Campus

See next page for additional authors

Recommended Citation

Zbylut, Joanna; Murray, Leigh; Thomas, S. H.; Beacham, J.; Schroeder, J.; and Fiore, C. (2014). "MODELING RATIOS WITH POTENTIAL ZERO-INFLATION TO ASSESS SOIL NEMATODE COMMUNITY STRUCTURE," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1007

Author Information

Joanna Zbylut, Leigh Murray, S. H. Thomas, J. Beacham, J. Schroeder, and C. Fiore

MODELING RATIOS WITH POTENTIAL ZERO-INFLATION TO ASSESS SOIL NEMATODE COMMUNITY STRUCTURE

Joanna Zbylut[1], Leigh Murray[1], S.H. Thomas[2], J. Beacham[2], J. Schroeder[2], C. Fiore[2]
[1]Department of Statistics, Kansas State University, Manhattan, KS 66506, [2]Department of Entomology, Plant Pathology and Weed Science, New Mexico State University, Las Cruces, NM 88003

Abstract

The southern root-knot nematode (SRKN) and the weedy perennials, yellow nutsedge (YNS) and purple nutsedge (PNS) are simultaneously-occurring pests in the irrigated agricultural soils of southern New Mexico. Previous research has characterized SRKN, YNS and PNS as a mutually-beneficial pest complex and has shown their enhanced population growth and survival when they occur together. In addition, it was shown that the density of nutsedge in a field could be used as a predictor of SRKN juveniles in the soil. In addition to SRKN, which is the most harmful of the plant parasitic nematodes, in southern New Mexico other species or categories of nematodes were identified and counted. Some of them are not as damaging to crop plants as SRKN, and some of them may be essential for soil health. The nematode species could be grouped into categories according to trophic level (what nematodes eat) and herbivore feeding behavior (how herbivore nematodes eat). Then three ratios of counts each were calculated for trophic and feeding behavior categories to investigate the soil nematode community structure. These proportions were modeled as functions of the weed hosts YNS and PNS by generalized linear regression models using the logit link function and three distributions: the Binomial, Zero-Inflated Binomial (ZIB) and Binomial Hurdle (BH). The latter two were used to account for potential high proportions of zeroes in the data. Formulas for the probability mass functions and moments were developed for the ZIB and BH. The SAS NLMIXED procedure was used to fit models for each of three sampling dates (May, July and September) in the two years of an alfalfa field study. General results showed that the Binomial generally provided the best fit, indicating lower zero-inflation than expected, but that ZIB and BH are often comparable. Importance of YNS and PNS predictors varied over sample dates and ratios. Specific results for one selected ratio illustrate the differences in estimated probabilities between Binomial, ZIB and BH distributions as YNS counts increase.

Keywords: Nematodes; Nutsedge; Community Ratios; Binomial; Zero-Inflated Binomial; Binomial Hurdle.

1. Introduction

Nematodes are usually microscopic worm-like organisms or roundworms, which can live in almost every habitat on earth. There are thousands (if not millions) of nematode species, most of them still not described by scientists. Many of them play critical ecological roles as decomposers and predators on microorganisms, but some are parasitic species, which affect humans directly or indirectly.

The most economically-damaging genera of plant-parasitic nematodes on horticultural and field crops are the root-knot nematodes, *Meloidogyne* spp. These species live and feed within plant roots most of their lives. The southern root-knot nematode, *Meloidogyne incognita* (SRKN) is one of the most damaging species of plant-parasitic nematode. Southern root-knot nematode is widely distributed and without proper management, can result in yield losses that exceed 40% in chile and 25% in cotton and many other annual New Mexico crops (Thomas et al. 1997). Southern root-knot nematode and the weedy perennials, yellow nutsedge (*Cyperus esculentus*, YNS) and purple nutsedge (*Cyperus rotundus*, PNS) are simultaneously-occurring pests in the irrigated agricultural soils of southern New Mexico. Previous research (Schroeder et al. 1994, Thomas et al. 1997, Schroeder et al. 2004, Thomas et al. 2004, Schroeder et al. 2005) has characterized SRKN, YNS and PNS as a mutually-beneficial pest complex, showing their enhanced population growth and survival when they occur together. Therefore the effectiveness of management practices that target the nematodes or nutsedge weeds alone is substantially reduced due to these mutually-beneficial interactions.

Other work (Fiore 2004, Fiore et al. 2009, Ou et al. 2008, Murray et al. 2012, Trojan et al. 2009) focused on developing an economic integrated pest-management strategy able to manage these three pests. Crop rotation with a non-dormant, SRKN-resistant alfalfa (*Medicago sativa* cultivar 'Mecca II'), which has aggressive growth and can successfully compete with nutsedge for light and other resources, was shown to provide simultaneous suppression of those pests (Fiore et al. 2009). It was also shown by monitoring the locations of pest population suppression and resurgence of each pest, that the density of nutsedge in a field could be used as a predictor of SRKN juveniles in the soil (Ou et al. 2008, Murray et al. 2012).

In addition to SRKN, other species or categories of nematodes were identified in the alfalfa field study (Trojan et al. 2009). Some of them are not as damaging to crop plants as SRKN, and some of them may be essential for soil health.

This paper is a follow-up to the modeling work by Ou et al. (2008) and Murray et al. (2012) to use counts of both SRKN and other nematode species or categories to examine the soil nematode community structure, especially as it relates to the presence of the plant hosts, YNS and PNS. In this paper, it is the ratios of counts that are of interest, not the absolute counts, because the ratios give information on competition and synergism between groups of nematodes.

2. Materials and Methods for the Nematode/Nutsedge Field Experiment

2.1 Description of field experiment

The data used in this paper and in Fiore (2004), Fiore et al. (2009), Ou et al. (2008) and Murray et al. (2012) came from a two-year alfalfa field experiment which was initiated in September 2004 at the Leyendecker Plant Science Research Center, New Mexico State University, near Las Cruces, NM, in soil which was infested with the SRKN/YNS/PNS pest complex. For complete information on the management of the field experiment, see Fiore (2004) and Fiore et al. (2009). To obtain data, researchers chose a 55x100 m rectangular section of a 1-ha alfalfa field with similar irrigation properties. Further, this section was split into a grid with a total of 1,375 plots of size 2 m x 2 m and was sampled six times in

- 2005: Sample 1 (May 19), Sample 2 (July 8), Sample 3 (September 16)
- 2006: Sample 1 (May 2), Sample 2 (July 25), Sample 3 (September 28).

Because of logistical constraints on personnel and time, N=80 2 m x 2 m plots were randomly chosen out of the 1,375 plots on each sample date. No plots were resampled within each year, but a few plots sampled in 2005 were resampled in 2006. Data were obtained from a 0.25 x 1 m quadrat put in the middle of a chosen 2 x 2 m plot and included visual counts of YNS and PNS shoots and counts of twelve categories or species of nematodes recovered from the soil. To obtain counts of nematode populations, ten 50-cm^3 soil cores were collected near nutsedge plants, if nutsedge plants were present in selected quadrat, or at random within the quadrat, if no nutsedge plans were present. Juvenile nematodes were extracted from the 500 cm^3 of soil by elutriation and processed using centrifugal flotation (Jenkins, 1964).

2.2 Characterizing the soil nematode community

The first step in characterizing the nematode community was to identify what nematodes are present. In this data, 12 categories or species of nematode were identified (Table 2.1). Unfortunately, for May and July of 2005, some species of nematodes were not identified separately, but were listed together as "other", as indicated in Table 2.1.

Table 2.1 Identified categories and species of nematodes in Nematode data.

	Nematode categories[1]	May05 & July05	Sep05, May06, July06 & Sep06
1	*Meloidogyne incognita*	√	√
2	*Trichodorus* spp.	√	√
3	*Tylenchorhynchus* spp.	√	√
4	*Pratylenchus* spp.	√	√
5	*Mesocriconema* spp.	√	√
6	Bacteriovores		√
7	Aphelenchoid		√
8	Dorylaimoid		√
9	*Hemicycliophora* spp.		√
10	Entomopathogenic		√
11	*Tylenchus* spp.		√
12	Monochoid		√

[1]Nematode categories not identified separately in May and June 2005 were pooled into "Other" category.

Second, nematode categories and species were grouped into categories according to trophic level (what nematodes eat) and herbivore feeding behavior (how herbivore nematodes eat).

Trophic categories were defined as follows (S. H. Thomas, personal communication):

• **Fungivore**: nematode species that feed exclusively on fungi present in the soil. These nematodes play an important role in soil health by taking nutrients extracted from organic matter by soil fungi and excreting them into the water in the soil pore spaces where such nutrients are available to plants and other organisms. Aphelenchoid is the unique type of esophagus found in most fungivore nematodes.

• **Bacteriovore**: nematode species that feed exclusively on bacteria present in the soil. These are also important in soil health for the same reason as fungivores, except these nematodes excrete compounds extracted from organic matter by bacteria.

• **Herbivore**: nematode species that feed on higher plants; any plant-parasitic nematode is a herbivore. Herbivores include the following species: *Meloidogyne, Trichodorus, Tylenchorhynchus*, *Pratylenchus, Mesocriconema* and *Hemicycliophora*.

• **Other**: this was a general category that was used to lump all nematodes that are not Herbivores ("other" = fungivore + bacteriovore + omnivore + predator = Bacterial Feeders + Aphelenchoid + Dorylaimoid + Entomopathogenic + *Tylenchus* + Monochoid).

Feeding behaviors of interest within the herbivore group in nematodes were defined as follows (S. H. Thomas, personal communication):

• **Endoparasites**: nematodes that live inside of the host plant. They can be:

  • **Sedentary Endoparasites**: nematodes that feed as stationary parasites inside the plant root. They invade root tissues after hatching and then each set up a permanent feeding location. They must transform root tissue of the host plant to support theirs sedentary lifestyle, which is the most damaging to the plant. *Meloidogyne incognita* (SRKN) is a typical sedentary endoparasite and is the only such species in this data;

  • **Migratory Endoparasites**: lesion nematodes that feed while migrating around inside the root. They do not transform any root tissue, but kill the cells they feed on and cause wound channels in the roots. They are considered intermediate in the amount of damage a nematode causes on the plant. *Pratylenchus* is an example of migratory endoparasite.

• **Ectoparasites**: nematodes that live on the surface of the host plant. They never enter a root, and only stick their stylet (a type of feeding tube) into the root from outside in the soil. Typically they are the least damaging to the host, because only the stylet is inserted into the root. In this data, Ectoparasites include the species *Trichodorus, Mesocriconem*, T*ylenchorhynchus*, and *Hemicycliophora*.

Third, Trojan et al. (2009) used the above trophic and herbivore feeding behavior categories to define three tropic ratios (Table 2.2) and three feeding behavior ratios (Table 2.3) which give definitions of the ratios as well as descriptions of their meaning in the context of the soil nematode community structure. These ratios were calculated for all six dates, except for the ratios BH and FB which could not be calculated for May and July 2005 due to species being pooled into "other", as mentioned previously (Table 1.2). Therefore, there were 4(dates) x 6(ratios) +2(dates) x 4 (ratios) = 32 date+ratio combinations to be modeled.

Table 2.2 Ratios and Their Interpretation for Trophic Levels (see Table 2.1)

Ratio Label	Ratio calculation[2]		Interpretation
	Numerator	Denominator	
BH = $\frac{B}{B+H}$	B = Bacteriovores	B + H = Bacteriovores + Herbivores where H = Herbivores = the sum of counts for following categories: 1. Meloidogyne; 2. Trichodorus, 3. Tylenchorhynchus; 4. Pratylenchus, 5. Mesocriconema; 9. Hemicycliophora	proportion of bacteriovores to plant-parasitic nematodes
FB = $\frac{F}{F+B}$	F = Fungivores = count of: 7. Aphelenchoid	F + B = Fungivores + Bacteriovores (as previously defined for BH)	proportion of the nematodes that are important in nutrient cycling are feeding on fungi
HT = $\frac{H}{Total}$	H = Herbivores (as previously defined for BH)	Total = total count of 12 categories or species of nematodes	proportion of the total soil nematode community that are plant-parasitic or herbivores

Table 2.3 Ratios and Their Interpretation for Herbivore Feeding Behaviors (see Table 2.1)

Ratio Label	Ratio calculation[3]		Interpretation
	Numerator	Denominator	
SM = $\frac{S}{S+M}$	S = Sedentary Endoparasites = the count of: 1. Meloidogyne	S + M = Sedentary Endoparasites+ Migratory Endoparastites where M = Migratory Endoparasites = the count of: 4. Pratylenchus	the proportion of sedentary endoparasites (most damaging nematodes) among all the endoparasites in a sample
SE = $\frac{S}{S+E}$	S = Sedentary Endoparasites (as previously defined for SM)	S + E = Sedentary endoparasites + Ectoparasites where E = Ectoparasites = the sum of counts for following categories/species: 2. Trichodorus; 3. Tylenchorhynchus, 5. Mesocriconema; 9. Hemicycliophora	proportion of sedentary endoparasites to ectoparasites (least damaging) in a sample
ME = $\frac{M}{M+E}$	M = Migratory Endoparasites (as previously defined for SM)	M + E = Migratory Endoparastites + Ectoparasites (as previously defined for SM & SE)	proportion of migratory endoparasites to ectoparasites in a sample

3. Statistical Analysis

Previous analyses (Ou et al. 2008, Murray et al. 2012) focused on modeling SRKN juvenile counts as predicted by YNS and PNS plant counts. Nematode counts are discrete non-negative integers, which often have a skewed frequency distribution and higher than expected zero-counts, under the assumption of a Poisson distribution. Therefore, these data were analyzed using the Poisson and additionally the Poisson with scale parameter (Ou et al., 2008) and additionally the Generalized Poisson, the Zero-Inflated Poisson and the Poisson Hurdle distributions (Murray et al., 2012).

This current work is a continuation of that research to assess the overall soil nematode community structure as measured by the three trophic-level proportions and three feeding behavior proportions. For these ratios, large number of zero counts in the ratio numerator may occur, meaning that the Binomial distribution may not fit well.

Therefore the objective of this work was to perform statistical modeling for the 32 date+ratio combinations using YNS and PNS counts as predictors and three probability distributions (with the logit link function):

- Binomial distribution
- Zero-Inflated Binomial distribution (ZIB)
- Binomial Hurdle distribution (BH).

3.1 Probability distributions for dealing with "too many" zeroes

Stroup (2012) provides general discussion for Zero-Inflated and Hurdle distributions to deal with excessive zero counts. Generally, both types of distributions use a mixture of a binary, on-off process and a discrete counting distribution. Both are two-component distributions with a zero counts component (the "off" phase) and a discrete counting distribution component, denoted f(y) for the "on" phase. The difference between the two is that in Zero-Inflated distributions there are two sources of zeroes because zeroes can occur in both the "off" phase and "on" phase, whereas in Hurdle distributions, there is only one source of zeroes, because zeroes occur only in the "off" part of the process. The parameter π_0 is called the inflation probability because, when the system is "off", zero counts are "inflated" or more frequent in comparison to the distribution that consists of only the discrete counting distribution (the "on" part).

The generic Zero-Inflated probability mass function (PMF) can therefore be defined as (Stroup, 2012):

$$\Pr(Y=y) = \begin{cases} \pi_0 + (1-\pi_0)\,\mathrm{f}(0) & for\ y = 0 \\ (1-\pi_0)\,\mathrm{f}(y) & for\ y = 1,2,\ldots,n \end{cases} \qquad \text{(Eq. 3.1)}$$

The generic Hurdle PMF can be defined as (Stroup, 2012):

$$\Pr(Y=y) = \begin{cases} \pi_0 & for\ y = 0 \\ (1-\pi_0)\,\frac{\mathrm{f}(y)}{1-\mathrm{f}(0)} & for\ y = 1,2,\ldots,n \end{cases} \qquad \text{(Eq. 3.2)}$$

In this work, f(y), the discrete PMF for the "on" part of the process, was the Binomial PMF because we were modeling ratios of discrete counts. We used Stroup's generic PMFs to obtain formulas for PMFs, means and variances of the ZIB and BH. Table 3.1 gives the PMFs, means and variances for the Binomial (with random variable Y), ZIB (with random variable Z and BH (with random variable H) distributions for Y=Z=H=0, 1, 2, …, n, where π_p is used to denote the Binomial probability and π_0 the zero-inflation probability.

Table 3.1 Comparison of the Probability Mass Functions and Moments for the Binomial, ZIB and BH Distributions

	Binomial	ZIB	BH
PMF and Parameter Space	Pr (Y=y) $= \binom{n}{y} (\pi_p)^y (1-\pi_p)^{n-y}$, y = 0, 1, 2, …, n	Pr(Z=z) $= \pi_0 + (1-\pi_0)\Pr(Y=0)$ $= \pi_0 + (1-\pi_0)(1-\pi_p)^n$ for z=y = 0 or $= (1-\pi_0)\Pr(Y=y)$ $= (1-\pi_0)\binom{n}{y}(\pi_p)^y(1-\pi_p)^{n-y}$ for z= y =1, 2, … , n	Pr(H=h) $= \pi_0$ for h= y = 0 or $= (1-\pi_0)\frac{\Pr(Y=y)}{1-\Pr(Y=0)}$ $= (1-\pi_0)\frac{\binom{n}{y}(\pi_p)^y(1-\pi_p)^{n-y}}{1-(1-\pi_p)^n}$ for h = y =1, 2, … , n
Mean	$E(Y) = n\pi_p$	$E(Z) = E(Y)(1-\pi_0)$ $= n\pi_p(1-\pi_0)$	$E(H) = E(Y)\frac{(1-\pi_0)}{1-(1-\pi_p)^n}$ $= (n\pi_p)\frac{(1-\pi_0)}{1-(1-\pi_p)^n}$
Variance	$V(Y) = n\pi_p(1-\pi_p)$	$V(Z) = E(Z^2) - [E(Z)]^2$ $=\{(1-\pi_0)[V(Y) + (E(Y))^2]\} - [E(Z)]^2$ $=\{(1-\pi_0)[n\pi_p(1-\pi_p) + (n\pi_p)^2]\}$ $- [n\pi_p(1-\pi_0)]^2$	$V(H) = E(H^2) - [E(H)]^2$ $=\{\frac{(1-\pi_0)}{1-(1-\pi_p)^n}[V(Y) + (E(Y))^2]\} - [E(H)]^2$ $=\{\frac{(1-\pi_0)}{1-(1-\pi_p)^n}[n\pi_p(1-\pi_p) + (n\pi_p)^2]\}$ $- [(n\pi_p)\frac{(1-\pi_0)}{1-(1-\pi_p)^n}]^2$
Reduce to Binomial	•	when $\pi_0 = 0$	when $\pi_0 = 0$

[Note that "n" refers to the denominator count of a ratio, not the sample size N=80 of the number of quadrats being measured in each sample date.]

To compare the mean formulas, define the following multipliers

A = $(1-\pi_0)$

and

B = $(\frac{1}{1-(1-\pi_p)^n})$.

Note that A = $(1-\pi_0)$ is always smaller than 1 but that B = $(\frac{1}{1-(1-\pi_p)^n})$ is smaller or greater than 1, depending on n and π_p.

It is easily seen that the Binomial mean E(Y), is always greater than the ZIB mean E(Z), since E(Z) = A * E(Y). However, the Binomial mean E(Y) is *generally* greater than the BH mean E(H), but not always. Note that

E(H) = A * B * E(Y) = B * E(Z).

This says that the Binomial mean is greater than the BH mean when the multiplier B is big enough to compensate for a small multiplier A. Only in the case of small n (n=1 or n ≤ 3 in some cases) or when $\pi_0 = \pi_p = 0.50$ for n=1, is the BH mean E(H) greater than or equal to the Binomial mean E(Y). These patterns can be seen in Table 3.2 which shows calculated means for the Binomial, ZIB and BH distributions for combinations of (π_0 =0.25, 0.50) and (π_p =0.33, 0.50) and for sample sizes of n=1, 3, 5, 10 and 100.

With respect to variances, we note that the variance of the Binomial V(Y) is smaller than variance of ZIB V(Z) or BH V(H) for large n but not necessarily for small n. Zbylut (2014) gives a table (not included here due to space limitations) showing the calculated variances for the same combinations as in Table 3.2.

Graphs of the PMFs for Binomial, ZIB and BH PMFs for the same four combinations of (π_p, π_0) are given in Figure 3.1 for n=5. In every case, the probability of zero counts is much greater for ZIB and BH than for the Binomial. For the Binomial, P(Y=0) will only the highest probability if π_p is much smaller than the examples given here, whereas the ZIB P(Z=0) and the BH P(H=0) is generally the highest probability (except for BH with $\pi_0 = 0.25$ and $\pi_p = 0.33$). Also, in general, probabilities are lower for the highest number of successes and higher for zero successes when comparing the ZIB and BH to the Binomial.

4. Results of Model Fitting

4.1 Fitting the models

Because the ZIB and BH PMFs are not available in the SAS GENMOD and GLIMMIX procedures, the SAS (ver. 9.3) NLMIXED procedure was used to fit models based on maximum likelihood. SAS NLMIXED code for fitting all models is given in Zbylut (2014) Appendix A, while Appendix A of this paper gives NLMIXED code for one ratio and on regression model.

Four regression models were fitted for the Binomial probability π_p using YNS and PNS as predictors:

- Intercept-only model (with parameter denoted β_0),
- YNS-only model (with YNS count as predictor and parameters denoted β_0, β_1),
- PNS-only model (with PNS count as predictor and parameters denoted β_0, β_2),
- Full model (with YNS, PNS, YNS^2, PNS^2 and YNS*PNS as predictors and parameters denoted β_0, β_1, β_2, β_{11}, β_{22}, β_{12}).

Table 3.2 Comparison of means for Binomial E(Y) versus the ZIB E(Z) and BH E(H) distributions

ZIB $\pi_0 = 0.25$ & $\pi_p = 0.33$						
π_0	0.25					
π_p	0.33	E(Y)	x multipA	= **E(Z)**		**E(Y)**
n	**1**	0.330	0.75	**0.2475**	<	0.330
n	**3**	0.990	0.75	**0.7425**	<	0.990
n	**5**	1.650	0.75	**1.2375**	<	1.650
n	**10**	3.300	0.75	**2.4750**	<	3.300
n	**100**	33.000	0.75	**24.7500**	<	33.000

BH $\pi_0 = 0.25$ & $\pi_p = 0.33$							
π_0	0.25						
π_p	0.33	E(Y)	x multipA	x multipB	= **E(H)**		**E(Y)**
n	**1**	0.330	0.75	3.03	**0.7500**	>	0.330
n	**3**	0.990	0.75	1.43	**1.0619**	>	0.990
n	**5**	1.650	0.75	1.16	**1.4307**	<	1.650
n	**10**	3.300	0.75	1.02	**2.5210**	<	3.300
n	**100**	33.000	0.75	1.00	**24.7500**	<	33.000

ZIB $\pi_0 = 0.50$ & $\pi_p = 0.33$						
π_0	0.50					
π_p	0.33	E(Y)	x multipA	= **E(Z)**		**E(Y)**
n	**1**	0.330	0.50	**0.1650**	<	0.330
n	**3**	0.990	0.50	**0.4950**	<	0.990
n	**5**	1.650	0.50	**0.8250**	<	1.650
n	**10**	3.300	0.50	**1.6500**	<	3.300
n	**100**	33.000	0.50	**16.5000**	<	33.000

BH $\pi_0 = 0.50$ & $\pi_p = 0.33$							
π_0	0.50						
π_p	0.33	E(Y)	x multipA	x multipB	= **E(H)**		**E(Y)**
n	**1**	0.330	0.50	3.03	**0.5000**	>	0.330
n	**3**	0.990	0.50	1.43	**0.7079**	<	0.990
n	**5**	1.650	0.50	1.16	**0.9538**	<	1.650
n	**10**	3.300	0.50	1.02	**1.6806**	<	3.300
n	**100**	33.000	0.50	1.00	**16.5000**	<	33.000

ZIB $\pi_0 = 0.50$ & $\pi_p = 0.50$						
π_0	0.50					
π_p	0.50	E(Y)	x multipA	= **E(Z)**		**E(Y)**
n	**1**	0.500	0.50	**0.2500**	<	0.500
n	**3**	1.500	0.50	**0.7500**	<	1.500
n	**5**	2.500	0.50	**1.2500**	<	2.500
n	**10**	5.000	0.50	**2.5000**	<	5.000
n	**100**	50.000	0.50	**25.0000**	<	50.000

BH $\pi_0 = 0.50$ & $\pi_p = 0.50$							
π_0	0.50						
π_p	0.50	E(Y)	x multipA	x multipB	= **E(H)**		**E(Y)**
n	**1**	0.500	0.50	2.00	**0.5000**	=	0.5000
n	**3**	1.500	0.50	1.14	**0.8571**	<	1.5000
n	**5**	2.500	0.50	1.03	**1.2903**	<	2.5000
n	**10**	5.000	0.50	1.00	**2.5024**	<	5.0000
n	**100**	50.000	0.50	1.00	**25.0000**	<	50.0000

ZIB $\pi_0 = 0.25$ & $\pi_p = 0.50$						
π_0	0.25					
π_p	0.50	E(Y)	x multipA	= **E(Z)**		**E(Y)**
n	**1**	0.500	0.75	**0.3750**	<	0.500
n	**3**	1.500	0.75	**1.1250**	<	1.500
n	**5**	2.500	0.75	**1.8750**	<	2.500
n	**10**	5.000	0.75	**3.7500**	<	5.000
n	**100**	50.000	0.75	**37.5000**	<	50.000

BH $\pi_0 = 0.25$ & $\pi_p = 0.50$							
π_0	0.25						
π_p	0.50	E(Y)	x multipA	x multipB	= **E(H)**		**E(Y)**
n	**1**	0.500	0.75	2.00	**0.7500**	>	0.5000
n	**3**	1.500	0.75	1.14	**1.2857**	<	1.5000
n	**5**	2.500	0.75	1.03	**1.9355**	<	2.5000
n	**10**	5.000	0.75	1.00	**3.7537**	<	5.0000
n	**100**	50.000	0.75	1.00	**37.5000**	<	50.0000

Figure 3.1 Graphs of Binomial, ZIB and BH PMFs for 4 combinations of π_p and π_0 & n=5

The zero inflation probability π_0 was modeled as an intercept-only model with parameter denoted α_0.

Note that for the fitted regression models, the formulas in Table 3.1 are modified by substituting the appropriate formula for π_0 and π_p. For example, for the regression model fitting the intercept and YNS,

$\pi_0 = \exp(\alpha_0)/[1+ \exp(\alpha_0)]$

and

$\pi_p = \exp(\beta_0 + \beta_1 YNS)/[1+ \exp(\beta_0 + \beta_1 YNS)]$.

In total, there were 32 (date+ratio) x 4 (regression models) x 3 (distributions) = 384 models. Summary Tables for all 384 models are given in Appendix B of Zbylut (2014).

In the following sections, we discuss results based only on statistical model-fitting and comparing the estimated probabilities for the three distributions. Interpretation of the biological and ecological implications of these results will be addressed in a future paper targeted at the disciplines of nematology and weed science.

4.2 Overall results

For each date+ratio combination, Akaike's Information Criterion (AIC) was used to determine the "best" model with minimum AIC. Table 4.1 lists the combination of (regression model + PMF) with the smallest value of AIC for each date+ratio combination. From Table 4.1, it is seen that the Binomial provided the best fit for 20 date+ratios out of 32, ZIB was the best for 10, with BH being best for only 2. The dominant position of Binomial model may indicate lower zero-inflation than was expected, based on previous work modeling only SRKN counts using members of the Poisson family of distributions (Ou et al. 2008, Murray et al. 2012). The importance of YNS and PNS predictors varied over date+ratio combination, since the YNS-only model was the best predictor for 11 date+ratio combinations, the Full model was best for 9, the Intercept-only model was the best for 8, and the PNS-only model was the best for 4.

4.3 Best regression + PMF models for all ratios in September 2005 and 2006

In this section, we examine results more closely for the two September sample dates, which were of specific interest to the nematologist and weed scientist co-authors because both the nutsedges and nematodes will then be at maximum levels, having experienced a full growing season.

Table 4.2 shows the smallest values of AIC in bold for the best combination of regression model + PMF for September. However, instead of just noting the PMF with the lowest AIC, we now look at how comparable the other two PMFs are to the best PMF, based on AIC. Burnham and Anderson (2010) suggest comparing AIC by calculating the difference between AIC values for the best model (i.e. with minimum AIC) and another model, denoted model *i*. This AIC difference is defined as $\Delta i = AIC i - AIC min$. Further, if $\Delta i \leq 2$, Burnham and Anderson suggest that the level of empirical support of model *i* is substantial in comparison to the best model, meaning the two models have comparable support. When $2 < \Delta i < 10$, model *i* is considerably worse than the best model in explaining some substantial variation in the data. Models with $\Delta i \geq 10$ have no support and might be omitted from further consideration.

It can be seen in Table 4.2, that although the Binomial is the best with smallest AIC value in most cases, the ZIB and/or BH are often comparable. For the BH and HT ratios in both 2005 and 2006, the ZIB and BH are both comparable to the Binomial ($\Delta i \leq 2$). For the 2006 FB ratio, ZIB is comparable to the best Binomial ($\Delta i = 1.6$), while BH is much worse ($\Delta i = 17.6$). For the

2005 SM ratio, ZIB is also comparable to the best Binomial (Δi = 2), while BH could be excluded (Δi = 76.8). For the 2006 SE ratio, BH is comparable to the best ZIB (Δi = 0.5), while Binomial is much worse (Δi = 20.7). For the 2005 ME ratio, ZIB is comparable to the best Binomial (Δi = 0.6), while BH could be omitted (Δi = 14.1). For the 2005 FB ratio, the ZIB was the best, with Δi = 3.1 and 5.2 for Binomial and BH, respectively, meaning that they are worse than ZIB, but should not be excluded from consideration. For the 2006 SM ratio, BH and Binomial models very poorly approximate the data compared to the best ZIB, with their respective Δi both being greater than 10.

To summarize the statistical aspects of model fitting for September 2005 and 2006, in most cases, the Binomial typically gave the best fit, but in many cases ZIB and BH were comparable models. This suggests that there was some zero-inflation, but that the zero-inflation was not as strong as was seen in the previous analyses of SRKN counts (Ou et al, 2008, Murray et al. 2012).

4.4 Illustrating differences in estimated probabilities resulting from the fitted regression models and the Binomial, ZIB and BH distributions for the SM ratio in September 2005 and 2006

To illustrate specific differences in the estimated probabilities between the three distributions and how the probabilities change as the predictor value changes, we present two examples, the SM ratio for September 2005 and 2006, using the regression model with YNS as the predictor (Figures 4.1 and 4.2). We note that the regression model with YNS was the best for SM in September 2006. In comparison, the best regression model for the September 2005 SM ratio was the intercept-only model with AIC=64.9, but the second best was YNS model with AIC=65 (Zbylut, 2014, Appendix B). Therefore the two PMFs have comparable support.

Estimated probabilities were calculated using the PMF formulas in Table 3.1, with the fitted regression models substituted for π_0 and π_p (Zbylut 2014, Appendix B). Note that distributions graphed in Figure 3.1 are equivalent to the case where the best model is an intercept-only model for both the Binomial π_p and the zero-inflation π_0. In contrast, the regression model with YNS as the predictor for π_p model is given in Figures 4.1 and 4.2 to show how the estimated probabilities of the numerator counts change as YNS increases from 0 to 8. The ratio denominator count n is set to 5, so that the numerator count y is from 0 to 5. In reality, the denominator can obviously be different from 5, but multiple figures would need to be shown for varying values of n.

The September 2005 SM ratio (Figure 4.1) is a case where the modeled zero-inflation is either essentially non-existent (ZIB) or moderate (BH), and all three distributions have a negative estimated slope for YNS for the regression model for π_p. The zero-inflation parameter estimate is $\widehat{\alpha_0}$ = -19.31, which makes the estimated zero-inflation probability $\widehat{\pi_0}$ = 4.0966E-9. Therefore for the September 2005 SM ratio, the ZIB probabilities are almost exactly equal to the Binomial probabilities, as are the estimates for β_0 and β_1. For BH, $\widehat{\alpha_0}$ = .3023, resulting in $\widehat{\pi_0}$ = .5750, which doesn't depend on the value of YNS and hence P(H=0) is constant as YNS increases.

In all three PMFs, the negative $\widehat{\beta_1}$ means that the probability of a high (low) numerator count decreases (increases), as YNS increases. Thus for the Binomial and ZIB, for YNS=0, the estimated P(Y=5)=P(Z=5) are both 0.3558 and P(Y=1)=P(Z=1) are 0.0049 with the mode occurring at Y=Z=4, whereas for YNS=8, the estimated P(Y=5)=P(Z=5) is 0.0020 and P(Y=1) = P(Z=1) is 0.3700 with the mode occurring at Y=Z=1. The BH shows the same pattern although at a different scale due to the large estimated probability at H=0 (.5750) leaving only

Table 4.1 Combination of regression + PMF models with the smallest AIC (the best models) across all sample dates

B=Binomial, Z=ZIB & H=Hurdle Model, 1=Sample 1 (May), 2=Sample 2 (July), 3=Sample 3 (September)

	BH ratio				**FB ratio**				**HT ratio**						**SM ratio**						**SE ratio**						**ME ratio**					
	2005	2006			2005	2006			2005			2006			2005			2006			2005			2006			2005			2006		
	3	1	2	3	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Intercept										B						B	B				B			Z	B	Z		B				
YNS		B			Z	B			H			Z						Z	H	Z									B	Z		Z
PNS								B							B							B					B					
Full	B		B	B			B				B		B	B									Z								Z	

Table 4.2 Models with AIC (the best models) for September Samples (Sample 3). Bolded AIC values are smallest (best).

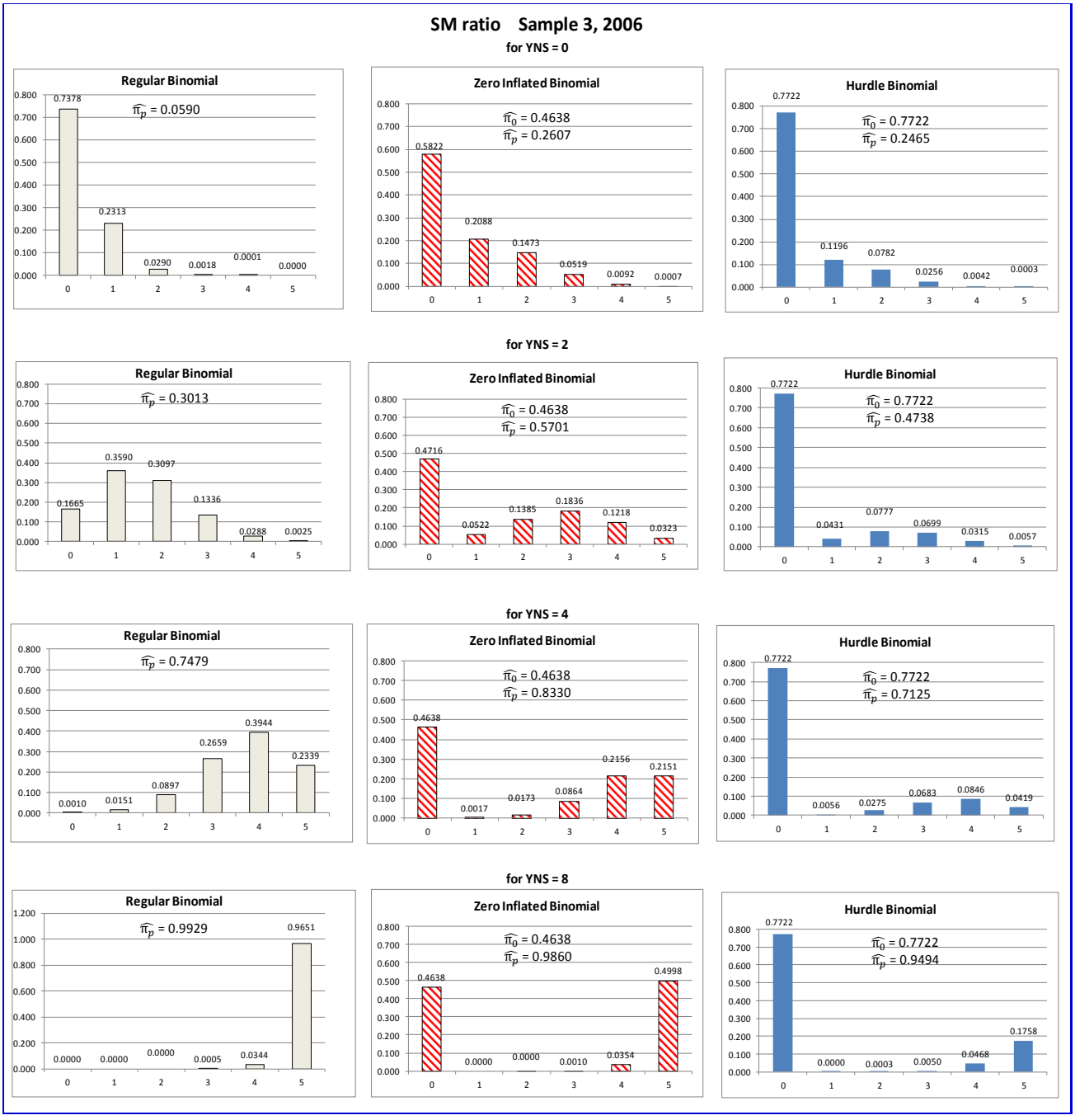
		BH ratio		**FB ratio**		**HT ratio**		**SM ratio**		**SE ratio**		**ME ratio**	
		2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006
Intercept	B							**64.9**			143.6		
	ZIB							66.9			**122.9**		
	BH							141.7			123.4		
YNS	B			290.3					131			**83.1**	671.4
	ZIB			**287.2**					**99.4**			83.7	**523.9**
	BH			292.4					120.3			97.2	531.2
PNS	B				**236.6**								
	ZIB				238.2								
	BH				254.2								
Full	B	**367**	**668.1**			**371.7**	**691.3**			182.7			
	ZIB	369	670.1			373.7	693.3			**172.2**			
	BH	369	670			372.7	693.3			175.6			

Figure 4.1 Estimated probabilities for SM ratio, September, 2005, YNS as predictor and denominator count=5, calculated using following estimated regression coefficients:

Binomial (AIC=65)	ZIB (AIC=67)	BH (AIC=143)
$\widehat{\beta_0}$ = 1.4714*	$\widehat{\alpha_0}$ = -19.3131	$\widehat{\alpha_0}$ =0.3023
$\widehat{\beta_1}$ = -0.297	$\widehat{\beta_0}$=1.4714**	$\widehat{\beta_0}$=1.4126**
	$\widehat{\beta_1}$=-0.297	$\widehat{\beta_1}$=-0.1536

Figure 4.2 Estimated probabilities for SM ratio September 2006, YNS as predictor and denominator count=5, calculated using following estimated regression coefficients:

Binomial (AIC=131)	ZIB (AIC=99)	BH (AIC=120)
$\widehat{\beta_0}$ = -2.7692**	$\widehat{\alpha_0}$ = -0.1452	$\widehat{\alpha_0}$ =1.2205**
$\widehat{\beta_1}$ = 0.9641**	$\widehat{\beta_0}$=-1.0425**	$\widehat{\beta_0}$=-1.1176**
	$\widehat{\beta_1}$=0.6624*	$\widehat{\beta_1}$=0.5063



0.4250 total probability on H=1, 2, 3, 4, and 5.

The increase seen for the estimated P(Z=0) for the ZIB as YNS increases is due solely to the negative YNS slope and not $\widehat{\pi_0}$. This suggests that zero-inflation will be difficult to detect in the ZIB if the regression coefficient is negative, unless the zero-inflation probability π_0 is "large". This happens because the negative slope will naturally shift probability from high counts to low counts of the numerator as the predictor increases, effectively disguising any zero-inflation. Because zero-inflation in the BH does not depend on the predictor, it may be better at picking up zero-inflation than the ZIB in such cases. In this specific case, however, the BH AIC is much larger than that for the Binomial and ZIB, indicating that it is likely that there is no zero-inflation for the SM ratio in September 2005.

The September 2006 SM ratio (Figure 4.2) is a case where the modeled zero-inflation is either low (ZIB) or high (BH), and all three distributions have a positive estimated YNS slope for the regression model for π_p. For the ZIB, the zero-inflation parameter estimate is $\widehat{\alpha_0}$ = -0.1452, giving $\widehat{\pi_0}$ =0.4638, while for the BH, $\widehat{\alpha_0}$ = 1.2205, giving $\widehat{\pi_0}$ =0.7722 =P(H=0), which again does not depend on YNS.

In all three PMFs, the positive $\widehat{\beta_1}$ produces a picture opposite to that seen in Figure 4.1, that is, probability increases (decreases) for higher (lower) numerator counts as the YNS predictor increases. Therefore for the ZIB, the estimated P(Z=0) decreases from 0.5822 when YNS=0 to 0.4638 when YNS=8. In comparison, the estimated P(Z=5) is 0.0007 when YNS=0 and increases to 0.4998 when YNS=8.

In the case where the estimated slope of the predictor is positive, it should be easy to detect moderate to strong zero-inflation when the value of the predictor takes on large values in the data in both the ZIB and BH.

Obviously the ability to detect and interpret zero-inflation will be much more complicated when there is more than one predictor variable and when the estimated coefficients differ in sign.

5. Summary

This paper presented formulas for PMFs and moments for the Zero-inflated Binomial (ZIB) and Binomial Hurdle (BH) and compared them to the Binomial when no regression model is fitted for the Binomial parameter π_p.

A real example was given using six nematode community ratios based on soil nematode counts. For these ratios, the Binomial PMF provided the best fit for 20 sample dates out of 32 (Table 4.1), but in several cases the ZIB and BH were comparable to the Binomial, based on very small differences in AIC values. The dominant position of Binomial model may indicate lower zero-inflation than was anticipated based on previous research that modeled SRKN counts using the Poisson family of distributions (Ou et al. 2008, Murray et al. 2012). Having many ratios based on small denominator counts (as there were for several dates and ratios) would likely make it difficult to detect the zero-inflation unless it was extreme. (In particular, if the denominator count is 1, the ZIB and BH reduce to the Binomial, albeit with different π_p.) This result also points up the difference in detectability of zero-inflation when dealing with counts versus ratios.

The importance of YNS and PNS predictors varied over sample dates and ratios, a finding similar to that Ou et al. (2008) and Murray et al. (2012), where the best Poisson regression models varied over dates. In addition, regression models from Ou et al. (2008) and Murray et al. (2012) found positive relationships between nutsedge density and SRKN counts,

while is this work some negative relationships were seen. This difference may be attributable to using information from more nematode species (with different feeding and reproductive strategies) rather than just SRKN and/or, again, the difference between modeling ratios versus counts. This is an issue of biology or ecology.

Finally, this discussion about modeling results leaves open the question about what model(s) the nematologists and weed scientists think are appropriate from a biological or ecological standpoint. See Murray et al. (2012) for a discussion of when particular Poisson-family distributions might be logically appropriate versus statistically "best".

As mentioned earlier in this work, we discuss results based only on statistical model-fitting. Interpretation of the biological and ecological implications of these results to issues of soil health and nematode community structure will be addressed in a future paper targeted at the disciplines of nematology and weed science

6. References

Burnham, K. P., Anderson D. R. (2010). Model Selection and Multimodel Inference. A Practical Information – Theoretic Approach. Springer.

Fiore, C. (2004). A sustainable approach to nematode and nutsedge management in chile using nematode-resistant alfalfa as a rotation crop. M.S. thesis, New Mexico State University, Las Cruses, NM.

Fiore, C., Schroeder, J., Thomas, S., Murray, L. & Ray, I. (2009). Root-knot nematode-resistant alfalfa suppresses subsequent crop damage from nutsedge-nematode pest complex, Agronomy Journal 101: 754-762.

Jenkins, W. R. (1964). A rapid centrifugal flotation technique for separating nematodes from soil. Plant Disease Report 48:692.

Murray, L., Thomas, S., Schroeder, J., Kreider, S., Ou, Z., Trojan, J., Fiore, C. (2012). Modeling the root-knot nematode / nutsedge pest complex: perspectives from weed science, nematology, and statistics. Proceedings of the 23rd Annual Kansas State University Conference of Applied Statistics in Agriculture: 128-151. Manhattan, KS, May 1-3, 2011.

Ou, Z., Murray, L., Thomas, S., Schroeder, J. & Libbin, J. (2008). Nutsedge counts predict *Meloidogyne incognita* juvenile counts in an integrated management system, Journal of Nematology 40: 99-108.

Schroeder, J., Kenney, M. J., Thomas, S. H. & Murray, L. (1994). Yellow nutsedge response to southern root-knot nematodes, chile peppers, and metolachlor, Weed Science 42:534-540.

Schroeder, J., Thomas, S. H. & Murray, L. W. (2004). Root-knot nematodes affect annual and perennial weed interactions with chile pepper, Weed Science 52:28-46.

Schroeder, J., Thomas, S. H. & Murray, L. (2005). Impacts of crop pests, and their management, on weeds, Weed Science 53:918-922.

Stroup, W. (2012). Generalized Linear Mixed Models. Modern Concepts, Methods and Applications. Chapman & Hall/CRC Texts in Statistical Science.

Thomas, S. H., Schroeder, J., Kenney, M. J. & Murray, L. (1997). Meloidogyne incognita inoculum source affects host suitability and growth of yellow nutsedge and chile pepper, Journal of Nematology 29: 404-410.

Thomas, S. H., Schroeder, J. & Murray, L. (2004). Cyperus tubers protect Meloidogyne incognita from 1,3-dichloropropene, Journal of Nematology 36:131-136.

Trojan, J., Thomas, S., Schroeder, J., Murray, L. & Schmidt, N. (2009). Soil nematode community structure in an alfalfa, chile, and cotton crop sequence. Annual meeting, Society of Nematologists, Burlington, VT, July 12-15, Journal of Nematology 41.

Zbylut, J. (2014) Modeling proportions to assess soil nematode community structure in a two year alfalfa crop. M.S. Report, Kansas State University, Manhattan, KS.

Appendix A – SAS Code: HT ratio, full quadratic model

SAS NLMIXED code is shown for the Binomial, ZIB and BH, the HT ratio, and the full quadratic regression model with predictors YNS, PNS, YNS^2, PNS^2 and YNS*PNS and their respective parameters denoted $\beta_0, \beta_1, \beta_2, \beta_{11}, \beta_{22}, \beta_{12}$ transliterated to b's and with the zero inflation probability π_0 modeled as an intercept-only model with parameter denoted α_0 (transliterated to a_0).

```
*************************************************************;
*HT ratio                                                   ;
*NLMIXED:                                                   ;
*   Binomial distribution                                   ;
*   LOG link                                                ;
*   Full model: YNS & PNS (b0, b1, b11, b2, b22, b12)       ;
*************************************************************;
title 'Binomial - Full Model - HT ratio';
proc nlmixed data=all;
by Year Sample;
    *Set FULL MODEL;
      parms b0=0   b1=0   b11=0
              b2=0   b22=0  b12=0;
      * Ratio HT;
      num=HTnum;
      den=HTdenom;
      * Binomial;
      LinpredBin=b0+b1*yns+b2*pns+b11*yns2+b22*pns2+b12*ynspns;
      Pi_p = 1 / (1+ exp(-linpredBin));
      if num = 0 then
            ll = den*log(1-Pi_p);
            else ll =  num*log(Pi_P) + (den-num)*log(1-Pi_P) + lgamma(den+1)
- lgamma(num+1) - lgamma(den-num+1);
      model num ~ general(ll);
      run;

*************************************************************;
*HT ratio                                                   ;
* NLMIXED:                                                  ;
*   ZIB distribution                                        ;
*   LOG link                                                ;
*   Full model: YNS & PNS (a0, b0, b1, b11, b2, b22, b12)   ;
*************************************************************;
title 'ZIB - Full Model - HT ratio';
proc nlmixed data=all;
by Year Sample;
      parms a0=0
              b0=0   b1=0   b11=0
              b2=0   b22=0  b12=0;
      * Ratio HT;
```

```
      num=HTnum;
      den=HTdenom;
    * ZIB model: linear predictor for the inflation prob;
    LinpredZero=a0;
      Pi_0=1/(1+exp(-linpredZero));
      * Binomial;
      LinpredBin=b0+b1*yns+b2*pns+b11*yns2+b22*pns2+b12*ynspns;
      Pi_p = 1 / (1+ exp(-linpredBin));
      * Log-likelihood for ZIB;
      if num = 0 then
            ll = log(Pi_0 + (1-Pi_0)*((1-Pi_p)**den));
            else ll = log(1-Pi_0) + num*log(Pi_P) + (den-num)*log(1-Pi_P) +
lgamma(den+1) - lgamma(num+1) - lgamma(den-num+1);
      model num ~ general(ll);
      run;

*************************************************************;
*HT ratio                                                   ;
*NLMIXED:                                                   ;
*   BH distribution                                         ;
*   LOG link                                                ;
*   Full model: YNS & PNS (a0, b0, b1, b11, b2, b22, b12)   ;
*************************************************************;
title 'BH - Full Model - HT ratio';
proc nlmixed data=all;
by Year Sample;
      parms a0=0
              b0=0   b1=0   b11=0
              b2=0   b22=0  b12=0;
      * Ratio HT;
      num=HTnum;
      den=HTdenom;
    * BH model: linear predictor for the inflation prob;
    LinpredZero=a0;
      Pi_0=1/(1+exp(-linpredZero));
      * Binomial;
      LinpredBin=b0+b1*yns+b2*pns+b11*yns2+b22*pns2+b12*ynspns;
      Pi_p = 1 / (1+ exp(-linpredBin));
      * Log-likelihood for Binomial Hurdle;
      if num = 0 then
            ll = log(Pi_0);
            else ll = log(1-Pi_0) + num*log(Pi_p) + (den-num)*log(1-Pi_p) +
lgamma(den+1) - lgamma(num+1) - lgamma(den-num+1) - log(1-((1-Pi_p)**den));
      model num ~ general(ll);run;
```