Estimating cavity tree and snag abundance using negative binomial regression models and nearest neighbor imputation

GMUG Meeting
28 January 2009
Portland, OR

Bianca Eskelson*, Temesgen Hailemariam*, and Tara Barrett⁺

* Oregon State University, ⁺ PNW Research Station, Anchorage

Outline

- Introduction
- Objectives
- Methods
  - Data
  - Negative binomial regression models
  - Nearest neighbor imputation
- Results
- Summary

Introduction

- Significant structural components of forest ecosystems
- Snags created by episodic events and small-scale mortality
- Cavity development governed by stochastic processes
- Highly variable with many zero observations
- Difficult to model

Introduction (cont'd)

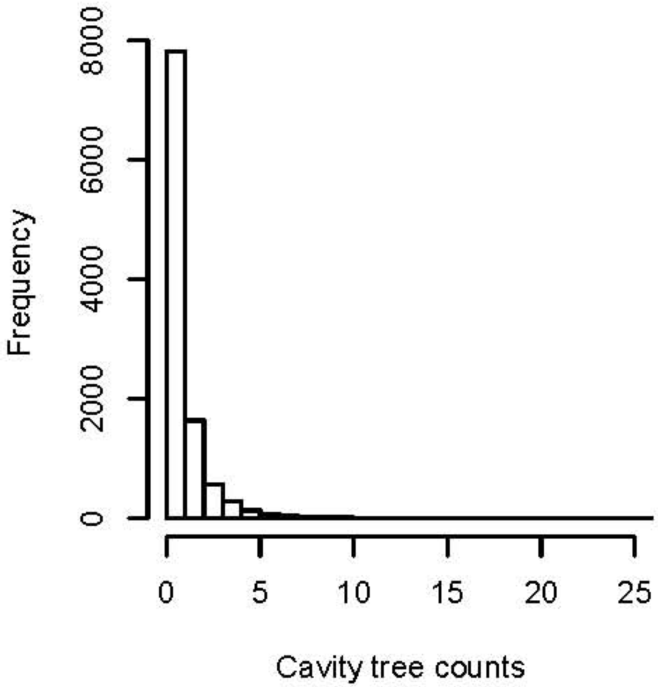
- Information on snag and cavity tree abundance is collected as part of FIA inventory in western US
- Interest to estimate snag and cavity tree abundance with variables that are readily available (e.g., forest cover maps, remotely sensed data)

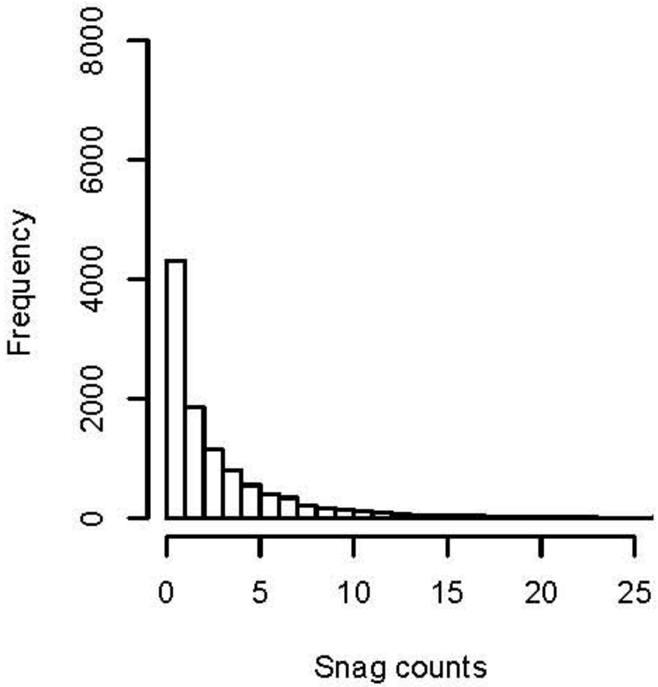
Objectives

- Estimate snag and cavity tree abundance with negative binomial (NB) regression models
  - NB
  - zero-inflated NB
  - zero-altered NB
- Estimate snag and cavity tree abundance with nearest neighbor (NN) imputation methods
  - MSN
  - randomForest
- Compare suitability and predictive abilities of NB models and NN methods

Data

- Washington, Oregon, California
- 10,607 stands





Data (cont'd)

Explanatory variables:

- Average stand age (years)
- Midpoint of five height classes (m)
- Slope (%), aspect (%), elevation (m)
- Midpoint of seven site classes (m^3/ha/yr)
- % conifer basal area
- Forest type groups: Douglas-fir, fir/spruce/mountain hemlock, other conifers, hardwoods
- Owner groups: Forest Service, other federal, state and local government, private

Negative binomial

$$P(Y = y) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(y+1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}$$

- *y* and *μ* > 0
- *random variable y = count response*
- *Γ = gamma*
- *α* represents the degree of overdispersion
- regression model obtained by relating mean *μ* to a vector of explanatory variables ***x***: $\mu = e^{x^T \beta}$

Zero-inflated NB

$$P(Y=y)=\begin{cases}\pi+(1-\pi)*f_{count}(0;x,\beta) & if\ \ y=0\\(1-\pi)*f_{count}(y;x,\beta) & if\ \ y=1,2,....\end{cases}$$

- probability of belonging to the point mass component: $\pi = f_{zero}(0;z,\gamma)$
- probability of belonging to the count distribution: $(1-\pi)$
- $f_{count}(0;x,\beta)=\left(1-(1+\alpha\mu)^{-(1/\alpha)}\right)$
- $f_{count}(y;x,\beta)$ = NB probability function

Zero-altered NB

- probability of a zero count: $f_{zero}(0; z, \gamma)$

  → determined by logistic regression
- zero-truncated NB: $f_{zt}(y; x, \beta) = \dfrac{f_{count}(y; x, \beta)}{1 - f_{count}(0; x, \beta)}$

- combined model:

$$P(Y = y) = \begin{cases} f_{zero}(0; z, \gamma) & if \;\; y = 0 \\ (1 - f_{zero}(0; z, \gamma)) * f_{zt}(y; x, \beta) & if \;\; y = 1,2,.... \end{cases}$$

NN imputation

- Donor-based methods

- **X-variables:** measured on all units
- **Y-variables:** forest attributes measured on subset of units

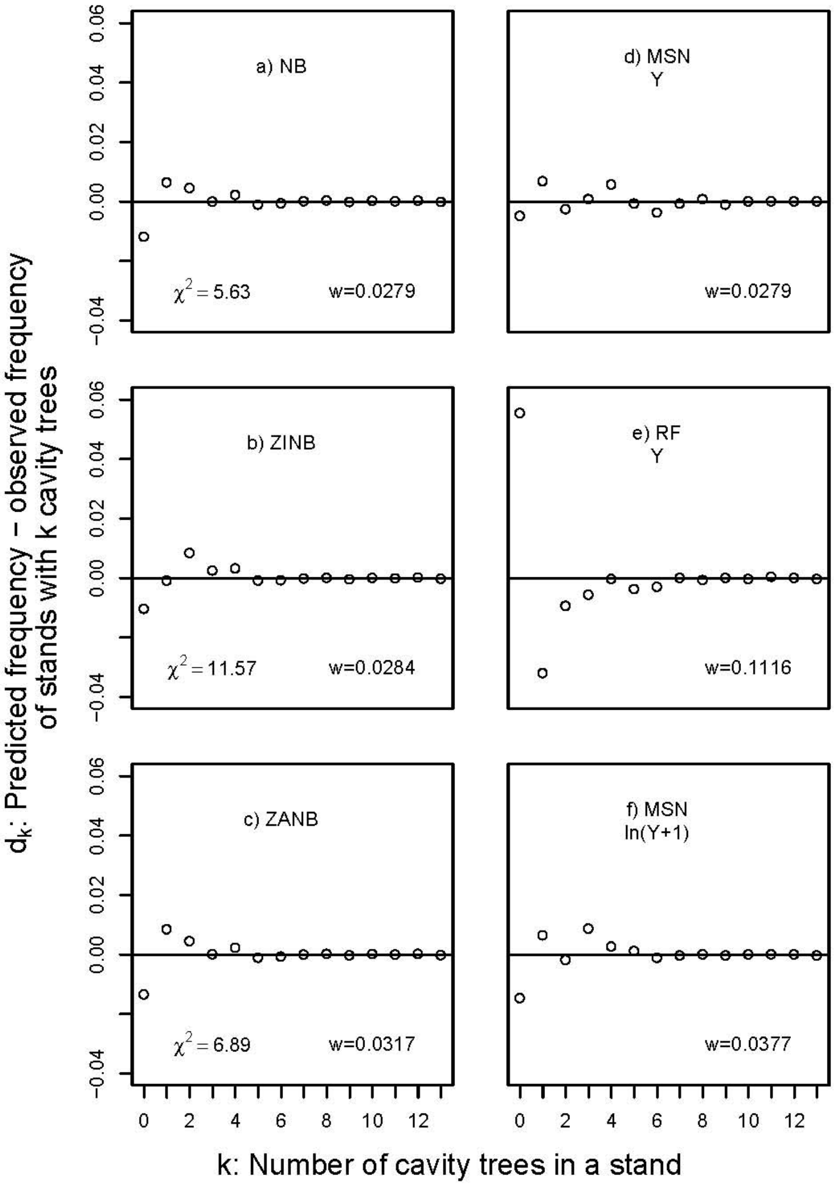
- **Target data:** units with X-variables only
- **Reference data:** units with X- and Y-variables

- **Similarity metric** determines similarity between target and reference data

NN imputation (cont’d)

- **RF:** randomForest (Crookston and Finley 2008)
- **MSN:** Most Similar Neighbor (Moeur and Stage 1995)
- **MSN ln(Y+1):** MSN using ln(Y+1) as Y-variable

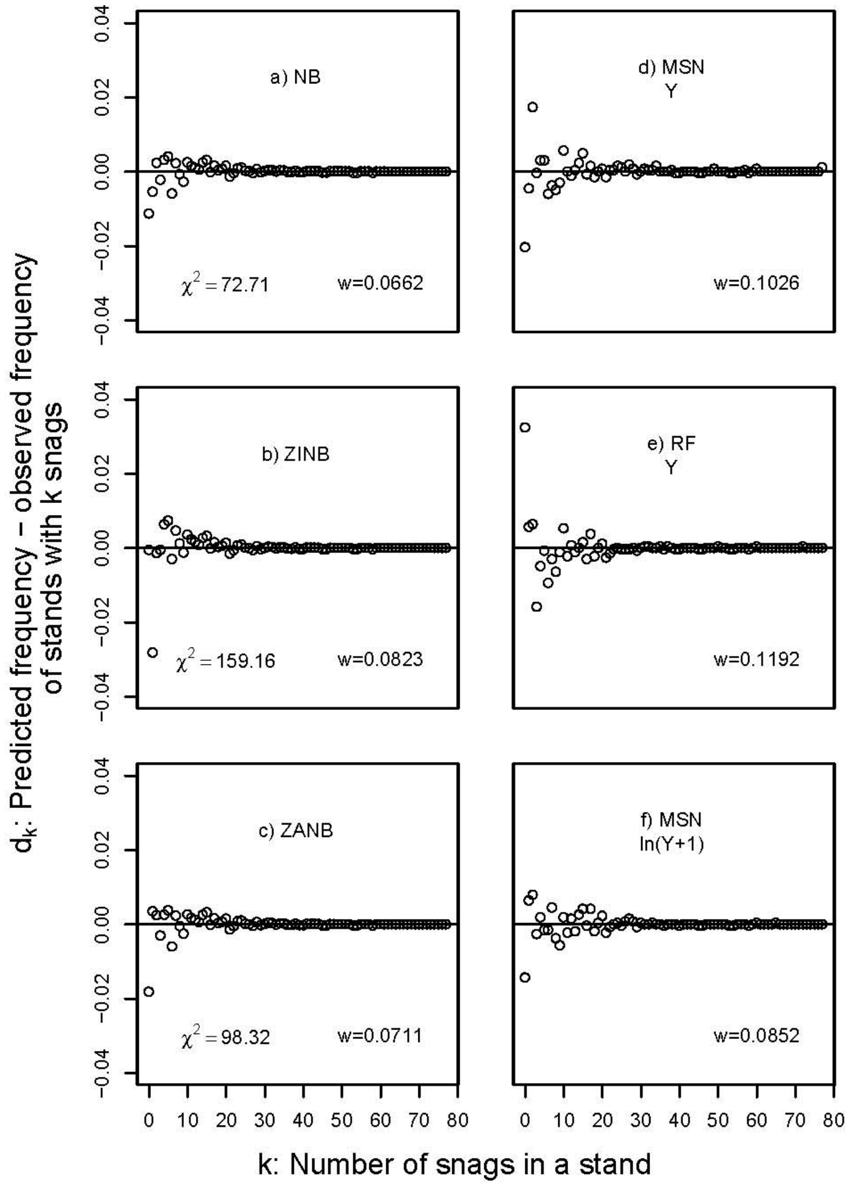
Results

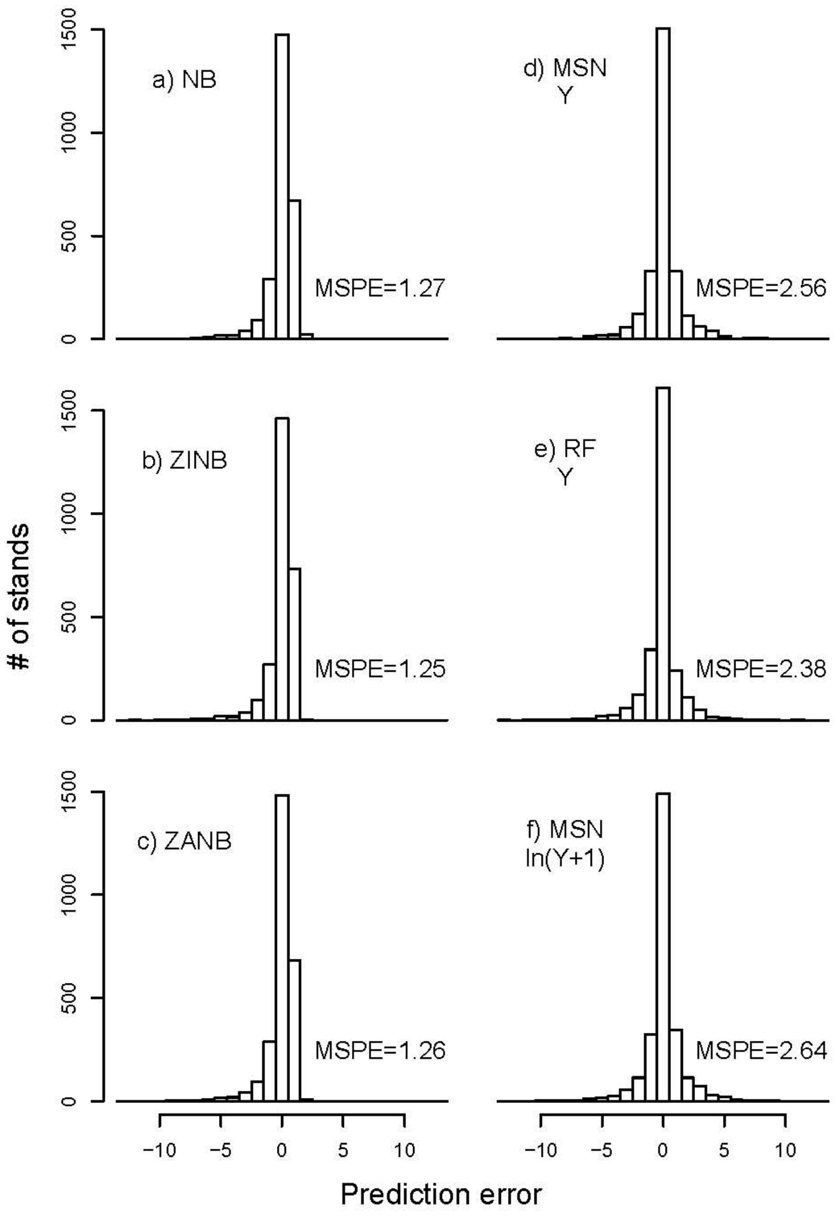


- Predictions bad for stands with small cavity tree counts
- NB models: almost perfect predictions for stands with counts > 5
- NN imputation: almost perfect predictions for stands with counts > 7

Results

- predictions bad for stands with small snag counts
- d_k smaller for NB models than for NN imputation methods



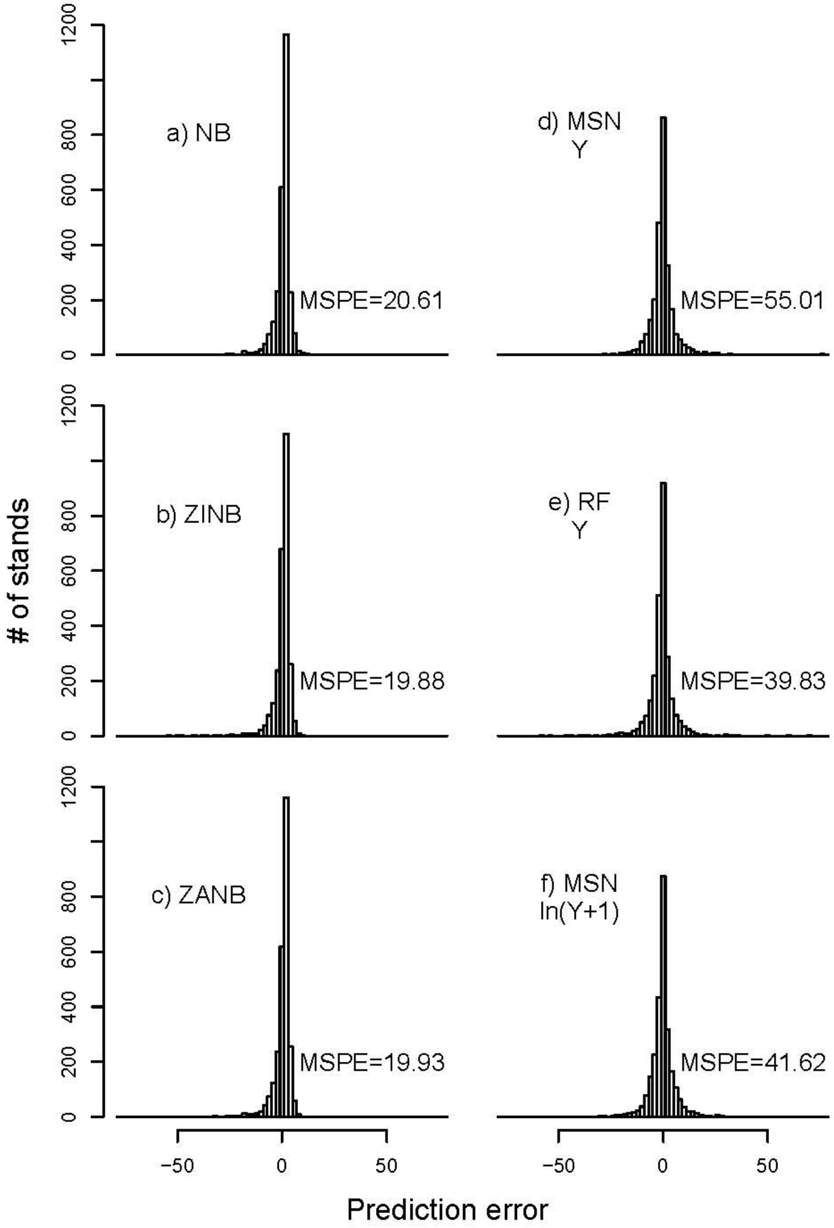


Results

- Prediction error of NN imputation methods covered whole range (-13, 13) → large MSPE
- NB models: no large overpredictions → smaller MSPE
- MSPE of NB models smaller than MSPE of NN imputation methods

Results

- Prediction errors for NN imputation methods cover full range
- NB models: no large overpredictions
- ZINB model smallest MSPE



Summary

- NB, ZINB, and ZANB models provided good estimates for overall cavity tree and snag abundance

  → prefer NB because of simplicity

- In terms of MSPE, NB regression models performed better than NN imputation methods

References

- Crookston, N.L. and A.O. Finley. 2008.yaImpute: An R package for kNN imputation. J. Stat. Soft. 23(10):1-16.
- Moeur, M. and Stage, A.R. 1995. Most similar neighbor: An improved sampling inference procedure for natural resource planning. For. Sci. 41(2):337-359.
- Zeileis, A., C. Kleiber, and S. Jackman. 2008. Regression models for count data in R. J. Stat. Software, 27(8):1-25.

Questions?