Nearest Neighbour Methods for Imputing Missing Data Within and Across Scales

Valerie LeMay, PhD, Forest Resources Management Department, Univ. of BC, Vancouver, Canada
Valerie.LeMay@ubc.ca
and
H. Temesgen, PhD, Dept. of Forest Resources, Oregon State University, Corvallis, OR
Hailemariam.Temesgen@oregonstate.edu

Presented at the "Evaluation of quantitative techniques for deriving National scale data for assessing and mapping risk workshop", Denver, CO, July 26-28, 2005

Abstract

Commonly, forested lands are divided into polygons based on forest type. Information for each polygon often includes variables that are measured on aerial photographs (e.g., species composition, height class), and additional variables derived from the aerial attributes using yield or other models (e.g., estimated volume per ha). For detailed information, such as the amount of coarse woody debris, stand structure, or tree-lists (stems per ha by species and diameter), ground sampling of every polygon is usually not possible. However, this information would be useful to represent the current inventory, and as model inputs to project future conditions. In stands that are sampled, detail at lower scales is often also of interest, but this may be available only for some of the sampled stands because of high measurement costs. Also, for recently cut stands, particularly, partially cut stands, estimates of future regeneration are needed as inputs to growth models.

For estimating variables of interest, imputation approaches are used as alternative to regression approaches. Imputation involves substituting, plausible measurements from one or more selected units with similar characteristics to units lacking these measures (Rubin 1987, Ek *et al.* 1997, McRoberts 2001). Data with all variables measured are termed "reference data", whereas data with some variables missing are termed "target data". If only one selected unit is used in the substitution, the variability of the missing variables as represented in the reference data will be preserved in the estimates imputed to the target data (Moeur and Stage 1995, Ek *et al.* 1997, Haara *et al.* 1997, Maltamo and Kangas 1998, Moeur 2000, LeMay and Temesgen 2005). This differs from regression approaches where averages, conditional on the values of the predictor variables, are used as the estimates for the missing variables in the target data. Since imputation involves searching the reference dataset for a "match" to the target dataset, this can be computer intensive. However, software has been developed by Moeur and Stage (1995), and updated by Moeur (2000). The most recent version of this software (MSN version 2.12; 2003[1]) is quite easy to use.

We have used imputation methods to: 1) estimate regeneration after partial cutting in mixed-species and/or uneven-aged stands (Hassani et al. 2004); 2) estimate tree-lists from aerial variables (forest cover) (LeMay and Temesgen 2001; Temesgen *et al.* 2003); 3) estimate wildlife trees (dead standing, or recently dead and fallen tree (Temesgen and LeMay 2001), and 4) to estimate stand level ground variables from aerial variables (LeMay and Temesgen 2005). We also have used simulation to compare different imputation methods (measures of similarity and numbers of reference plots used in imputation) for different sampling intensities for the reference data (LeMay and Temesgen 2005).

In this presentation, we present background information on imputation methods and demonstrate how these methods are employed to generate tree-lists from aerial attributes for non-sampled polygons, and improve forest inventories, analyses, and management. Examples are then given from our work using data from multi-species and multi-aged stands from southeastern British Columbia.

Keywords: Regeneration, nearest neighbour imputation, partial cutting, non-sampled polygons, tree-lists, snags, forest cover

[1] **http://forest.moscowfsl.wsu.edu/gems/msn.html**

References Cited

Ek. A.R., Robinson, A.P., Radtke, P.J., and Walters, D.K. 1997. Development and testing of regeneration imputation models for forests in Minnesota. For. Ecol. & Manage. 94: 129-140.
Haara, A., M. Maltamo, and T. Tokola. 1997. The k-nearest neighbor method for estimating basal area diameter distribution. Scand. J. For. Res. 12: 200-208.
Hassani, B.T., V. LeMay, P.L. Marshall, H. Temesgen, and A-A. Zumrawi. 2004. Regeneration imputation models for complex stands of southeastern British Columbia. For. Chron. 80(2): 271-278.
LeMay, V. and H. Temesgen. 2001. Nearest neighbor methods for generating tree-lists from aerial data. Invited paper presented at the IUFRO 4.11 Forestry biometry, modelling, and information science conference, The University of Greenwich, School of Computing and Mathematical Sciences, June 26-29, 2001, Greenwich, England. 23 pp. http://cms1.gre.ac.uk/conferences/iufor/proceedings/
LeMay, V. and H. Temesgen. 2005. Comparison of nearest neighbor methods for estimating basal area and stems per ha using aerial auxiliary variables. Forest Science 51(2): 109-119.
Maltamo, M. and A. Kangas. 1998. Methods based on k-nearest neighbor regression in the prediction of basal area diameter distribution. Can. J. For. Res. 28:1107-1115.
McRoberts, R.E., M.D. Nelson, and D.G. Wendt. 2002. Stratified estimation of forest area using satellite imagery, inventory data, and the k-Nearest Neighbors technique. Remote Sensing of Environment 82: 457-468.
Moeur, M. 2000. Extending stand exam data with most similar neighbor inference. pp. 99-107 in Proceedings of the Society of American Foresters National Convention. Sept. 11-15, 1999, Portland, Oregon.
Moeur, M. and A.R. Stage. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. For. Sci. 41: 337-359.
Rubin, D.B. 1987. Multiple imputations for non-response in surveys. John Wiley & Sons, New York. 258 pp.
Temesgen, H. and V.M. LeMay. 2001. Generating tree-lists from aerial attributes for sustainable resource management. In Proceedings of the IUFRO conference on Forest Modelling, August 12 to 17, 2001, Vancouver, BC, pp. 409-422.
Temesgen, H., V.M. LeMay, K.L. Froese, and P.L. Marshall. 2003. Imputing tree-lists from aerial attributes for complex stands of south-eastern British Columbia. For. Ecol. & Manage. 177: 277-285.

Biographical Sketches:

Valerie LeMay, PhD, RPF, Associate Professor, Biometrics/Mensuration, Dept. of Forest Resources Management, 2045-2424 Main Mall, University of British Columbia, , Vancouver, BC V6T 1Z4 Canada Phone: 604-822-4770 FAX: 604-822-9106 EMAIL: Valerie.LeMay@ubc.ca

1981	BSc (Forestry), University of Alberta
1982	MSc (Forest Measurements, University of Alberta
1982– 1986	Measurement analysis forester, Alberta Forest Service, Edmonton, AB
1985	Sessional lecturer (1/2 time leave from Alberta Forest Service for 1 term), Forest measurements
1988	Forest biometrics lecturer, LORAX forestry, Surrey, BC
1989	PhD (Forest Biometrics) University of British Columbia
1989 – present	Faculty member in Forest Resources Management Dept., UBC
1997 – 1998	Acting Dept. Head, Forest Resources Management Dept., UBC

H. Temesgen, PhD, RPF (BC), Assistant Professor, Biometrics/Mensuration, Dept. of Forest Resources, 237 Peavy Hall, Oregon State University, Corvallis, OR 97331-5703 USA Phone: 541-737-8549 FAX: 541-737-3039 Hailemariam.temesgen@oregonstate.edu

1985	BSc, Plant Sciences, Alemaya University Of Agriculture
1986-1989	Assistant Lecturer, Alemaya University Of Agriculture
1992	MSc, Quantitative Silviculture, Lakehead University
1996	Sessional lecturer, University of British Columbia, Forest measurement
1999	PhD, Forest Biometrics, University of British Columbia
1999-2003	Research Associate in Forest Biometrics and Measurement, University of British Columbia
2003 – present	Assistant Professor in Forest Biometrics and Measurement at Oregon State University