

To: Xiaoping Yuan, Forest Analysis and Inventory Branch
PO Box 9212 Stn Prov Gov, 7th Floor 727 Fisgard Street
Victoria, BC V8W 9C2

March 31, 2013

Re: Review of the LVI Software and its manual entitled "LVI Linear Discriminant Analyses" by Ian Moss, PhD, RPF, Tesera Systems Inc.

This letter will serve as a progress report for the evaluation of the above-captioned software and its manual prepared by Ian Moss, PhD, RPF of Tesera Systems Inc.

I have briefly examined the overall architecture and structure of the LVI software by examining the order of scripts and description of the scripts in the accompanying manual, and by running selected R codes. This letter briefly summarizes my preliminary findings.

In the final report, I will direct my review and remarks in the following seven major headings.

- 1.0 General comments
- 2.0 Contents of current software (what it does and does/not currently do?)
- 3.0 Evaluation/Tests/Validation of routines or codes
 - 3.1. Software
 - 3.1.1. Test data
 - 3.1.2. Technical analysis of components and assessments
 - 3.1.3. Logic and consistency
 - 3.1.4. Adaptability/Accessibility/Modularity
 - 3.2. Manual
 - 3.2.1. Contents
 - 3.2.2. Layout
 - 3.2.3. Technical
- 4.0 Statement of findings
 - 4.1. Strengths of the software and its manual
 - 4.2. Weaknesses of the software and its manual
- 5.0 Future enhancement for the process and its accompanying software
 - 5.1. Technical and analytical issues
 - 5.1.1. Classification
 - 5.1.2. Discriminant Analysis
 - 5.1.3. Nearest Neighbor Selection
 - 5.2. Software
 - 5.3. Manual/Documentation
 - 5.4. Publication in peer-reviewed journal
- 6.0 Suggested priorities for implementing the process and accompanying software
 - 6.1. Short term
 - 6.2. Long term
- 7.0 Overall Recommendation

Sincerely,

Temesgen Hailemariam, PhD, RPF

Evaluation of the LVI software and its manual entitled “LVI Linear Discriminant Analyses” by Ian Moss, PhD, RPF, Tesera Systems Inc

Temesgen Hailemariam, PhD, RPF

Statements of Preliminary Findings

Forest resource management is currently faced with a myriad of challenges. Solutions to these challenges require quantitative approaches in order to reach sound and broad-based solutions. Forestry has become increasingly quantitative in its approaches to research and sustainable management. Rising forest values increase the demand for accuracy and precision in quantitative approaches for management prescriptions and projected outcomes. Both trends have magnified the importance precise classification of categories and imputation. The LVI software and its manual outlined selected classification, discriminant analysis, and nearest neighbor selection approaches to improve future forest inventory and mapping accuracy that are critical in tackling these challenges.

Dr. Moss has produced professional and functional software that handles the difficult problems of classification, discriminant analysis, and nearest neighbor identification. The thinking behind the software and its manual, and the approaches taken to improve procedures for estimating polygon attributes and to impute multivariate response variables are well developed—indeed commendable.

Some of the strengths of the software and its manual include:

1. Codes are written in R and Python, which are freely available and widely used by the remote sensing and forest biometrics community.
2. Clear and complete directions are available for installation. Hence, practitioners/analysts are able to install software with ease.
3. The structure of the software is generic and modular. It can be easily modified to include other types of analysis (e.g., Random Forest, Systems of Equations, etc.) in the future.
4. CSV files are selected as standard data format. This makes it easier for practitioners and analysts to transfer and share files.
5. The methods used are straightforward, and the statistical analyses employed are appropriate. The *Fuzzy c-means classification* used in the software is an efficient method for cluster analysis of remotely sensed image data. The procedures used for variable selection and assigning weights for identifying nearest neighbors are lucid and enlightening.
6. The example data used in the analyses are reasonable. The Input data included Landsat 5 TM imagery, VRI, and Predictive Ecosystem Mapping. The variable selection routines used to examine over 300 combinations of variables are logical and efficient.
7. After examining different distance metrics and different sets of Y and X variable sets, the codes or routines provide data sets that can be used to display maps.

8. The manual is logically arranged. The contents of the manual are clearly and sufficiently identifiable in each of the seventeen chapters. The chapters compose clear and coherent topics.
9. Building on each preceding chapter, the contents of the manual motivate practitioners/analysts to learn and master classification and imputation concepts. The approach taken in the software can easily be extended to improve accuracy of different tree-list generation methods at varying scales.

The software and its manual are at their inception stages, and will evolve over time. Future enhancement of the software and its manual and the imputation process need to consider the following.

1. The number of R routines is unnecessarily large. It could be condensed by writing functions or by developing user-interface. Data dictionary is lacking.
2. Correlation coefficients between or among Xs are manually calculated. It should be automated.
3. Current version of the software does NOT include ensemble modeling including Random Forests.
4. While the codes or routines provide data up to display of maps, post-classification to display maps are not yet possible.
5. Over fitting is well known problem in statistics and in developing models using remotely sensed data. Lacking from the exposition is the efficiency of the selected variable selection procedures: a) in mitigating the problem of over fitting; and b) in comparison to other variable selection strategies (e.g., optimization-based variable selection methods). Some analysis or discussion on these topics will shed some light to practitioners and analysts.
6. Bootstrapping as a potential method to impute categorical response variables was not considered.
7. While the manual is complete and comprehensive, the structure of the document could be improved by reorganizing and rewriting some of the sections. The manual is lengthy and could be condensed. The presentation, in places somewhat tedious owing to the subject matter, is generally unambiguous and informative. However, Chapters 10 through 16 need some reorganization and additional thoughts. I will have some editorial and technical comments in the final report.

In sum, the classification, discrimination, and imputation questions addressed in the software could also be looked through a parametric lens (e.g., logistic regression or stepwise variable selection, etc.). It is refreshing that Dr. Moss did not focus on parametric methods that have been examined for several decades. I commend him for staying away from the beaten path and for using *Fuzzy c-means classification* to classify and then using discriminant analysis and k-NN methods for imputing polygon attributes. The software and its manual will help to improve procedures for estimating polygon attributes and to impute multivariate response variables. I suggest the author consider publishing part of the software and associated expositions in peer-reviewed journal and contribute to the existing body of knowledge.

Sincerely,

Temesgen Hailemariam, PhD, RPF