**LVI Process / kNN Upgrades**

**May 22 2012**

1. Report RMSE’s by attribute and class; also useful for looking at level of classification.

**April 15 2013**

List of issues

1. Over fitting
2. Multicollinearity
3. Ensemble modeling
4. Is there a relationship between distance and RMSE and if so what is it?
5. Is there a relationship between KHAT and RMSE?
6. Identifying frequency distributions of combinations of nearest neighbours.
7. Using classification to explore distribution of species
8. Using classification to identify outliers in the target dataset
9. Dictionary management – generation of dictionary
10. How are missing values accommodated?
11. Multicore processing
12. Clean up of code … eggs
13. PypeR

Other tests

1. Ground to photoplot comparison
2. Cost/benefit of ground plot sampling
3. LVI calls for ground plots

**March 22 2013**

Hello Temesgen

Thank you very much for spending time with me during this last week and thank you to your family for having me into your home – it was wonderful to meet them and the food was delicious.

As we discussed below are some questions that I think are worthy of addressing in your final report, and some additional ideas on next steps that I think have yet to be resolved.

Questions:

1. In your own words can you provide a list of what the current software  does and does/not currently do.
2. What improvements would you make to the analytical process and which ones do you see as high priority?
3. What improvements would you make to the software and which ones do you see as a high priority?
4. What improvements would you make to the documentation and which ones do you see as a high priority?
5. Is there anything in this work that you think is worthy of publication or that with some more work would be worthy of publication; if the latter what needs to be done to get it to that stage.
6. In your opinion, should the MoFLNRO (herein referred to as the MoF) continue to develop the kNN process and accompanying software starting with the process and software that has been provided to you.  If not what alternative do you suggest?

Next Steps (not in any order of priority)

1. The variable selection process should be wrapped up in a sampling, calibration, and testing routine where the calibration and testing datasets are separated.
2. There is a need to adopt/adapt/develop an estimate (and estimator) of variable importance.
3. Consideration should be given to ensemble modeling, i.e. using more than one model to estimate nearest neighbours.  Note that in this context the process of classification, discriminant analysis, nearest neighbor identification could be considered as process that is contained within a random forest  type of context; the two are not necessarily mutually exclusive.
4. The process for dealing with correlation amongst the variables should be automated.  In particular, consideration should be given to regressing each X-variable on the remaining X-variables within a dataset and removing those where the regression correlation coefficient is above (or below if negative) a certain user defined threshold.
5. The various routines should be wrapped up into single routines to reduce the number of steps.  This might also involve development of an easier to use – user-interface.
6. A process should be developed to better facilitate the construction of data dictionaries used in the process, particularly as it relates to the reference and target datasets.  In line with this topic the data dictionaries  that are to remain unchanged should be separated from those involving the target and reference datasets.
7. Consideration should be given to enabling multi-core or parallel processing for those (nearest neighbor combinatorial) routines that require large computer resources particularly as the number of observations and numbers of variables increase.
8. A procedure is needed to benchmark the target dataset vs the reference dataset with respect to the X-variable dataset with an emphasis of identifying those observations in the target dataset that are outside of the range contained within the reference dataset, and so too, those types or combinations that are under-represented in the reference dataset relative to those in the Target dataset.  This would be very useful in identifying locations suitable for establishment of additional samples so as to ensure adequate representation of the range of variation.  Note also that classification and/or principle Components analysis would be useful to assist in this endeavor.
9. The emphasis in this first project was placed on 1) developing a process that would be complete in terms of creating reference and target datasets and using these to create a map of desired Y-variables using kNN.  From a contractual point of view the emphasis was not placed on producing software that would enable implementation of the process from beginning to end, but it was a personal goal to do this as a means of helping the MoF become more self sufficient.  The vision was always that this would provide a foundation that could be extended in a number of different directions with time providing more options and basis for  comparisons of results with the goal of producing the best map possible.  If this work is to be continued then this should continue to be the basic design philosophy which by definition is modular in approach.
10. The manual should be re-organized.  Its current state is a result of having built the manual while at the same time building the code base – so that it has evolved as the project developed.  The content is mostly there but it should be rearranged so that the workflows are presented in a more orderly fashion.    There are also some issues around the manner of presentation that could be improved. Your suggestion that the introduction to the work plan be used as the basis for the introduction to this document will be implemented.  I will also endeavor to follow up on some of the other editorial remarks that you made during our session.
11. With a small degree of modification this software could be made completely generic such that it could be applied generally to processes involving classification, discriminant analysis (or alternatives for example involving log odds modeling, etc.), and nearest neighbor types of analyses either alone or combination.   It should be recognized that this process could be equally well applied to the notion of generating tree lists; I think it is important that the MoF recognize this.

I am sure that there is something that I have forgotten …  There is one other favour that I would like to ask of you.  You mentioned that you are interested in using the manual and routines as a training tool with your students.  I think this is a great idea but would ask that you give the Ministry of Forests and Tesera Systems credit for having built the tool in the first place.  Also if you find any problems or make any improvements or extend the work in any way then it would be appreciated if those could be provided to myself for further integration into the package.  So far this has all been done in an informal sort of way, but perhaps going forward from here we should be looking at a more formal open source kind of licensing and agreement.  I will look into that from my end pending your final recommendations that follow from your review.

I would like to end this on a personal note.  Firstly, I am a real believer in this kind of a review process and feel that regardless of the outcome, this kind of a process is vital to the development of strong program …. It is important that you conduct your review as an independent third party and that you provide an honest appraisal and set of recommendations that are in the best interest of moving the LVI process forward.   Secondly, I think that the effort to produce some software and an accompanying manual went beyond what the MoF contract called for, regardless of the outcome of this review.  It was done because I thought that just maybe I could do it and about half way through I realized how I might do a better job of it …. I am of the belief that in order for the Ministry to succeed they need to be familiar with the tools used to produce the products, not just the products themselves, and that there is no better way to do this other than to assemble the tools in such a way as to facilitate their use and interpretation by MoF staff without having to know the guts of what is going on under the hood.  This allows the Ministry to be the owner of the process in a way that would not be possible absent the software.

In closing a great deal of thanks goes to the MoF and Xiaoping in particular for giving me the opportunity to work on this project.  It has been a great opportunity to pursue work that I enjoy and care about.  Thanks also to you Temesgen for agreeing to do this review in spite of your very busy schedule.  This will help both Tesera and the MoF move forward from here.  Without this contribution the work would not be complete.

All the Best

Ian