Some General Observations about kNN

1. RMSE decreases exponentially with an increase in the number of variables and can be slightly U shaped.
2. Bias increases as RMSE decreases – a well known trend off but the bias in the LVI dataset was relatively small in most cases.
3. Clearly sampling intensity and range of variation matters – if certain species are not well represented then these may appear to be better than they really are in reality.
4. Each observation may be better represented in terms of distance by different variable sets – asymmetries may exist wherein one subpopulation is better represented by a certain set of variables while another subpopulation is better represented by another set of variables; this could be further extended to consider each of the y-variables separately.
5. In general the Euclidean distances and/or Mahalanobis distances produced the least bias results, but consistent with remarks above … with slightly higher RMSE’s. The latter two were nearly identifcal since they are both based on the use of discriminant function scores that have been normalized with respect to the mean and standard deviation and each of the discriminant functions are nearly if not completely orthoganal. These metrics are also consistent with MSN but based on maximizing the between-to-within group variance ratio’s associated with the eigen values and eigen vectors.
6. Tentaively – there the RMSE’s associated with a given set of nearest decreases and then increases as the number of groups increases from to 1 to n. In part this may because the number of classes also constrains the number of discriminant functions but this question needs further investigation.