On a Mantel-Test Extension for k Series of Distances and Related k-Way Multivariate Methods

Roberto Santos*[1] and Didier G. Leibovici[2]

[1]School of Geography, University of Nottingham

[2]Nottingham Geospatial Institute, University of Nottingham, UK

*Email: Roberto.Santos@nottingham.ac.uk

Abstract

The Mantel test of correlation of two distance matrices is well-known and often used in ecology and landscape genetic. In the presence of multiple distances, a partial correlation tests has been used but is limited and do not offer a symmetric view. As distances are positives the basic formula can be extended to more than two series. Rescaling each distance matrix to have a maximum of 1 and minimum of zero (the diagonal) ensures fair contributions from each domain used in this k-Mantel test. This leads also to consider multiway arrays expressing this Mantel's extension approach either from using the cross-products of the distances or considering k-co-occurrences based on these distances. Simple histograms or k-way multidimensional methods can then depict structures and associations. The paper presents the k-Mantel randomisation testing and these various methods with analyses results for a crop genetic environment association study using data for 33 bambara groundnut landraces (a neglected and underutilised crop).

Keywords: Mantel test, distances, genetic information, landscape genetic, multiway methods.

1 Introduction

Producing food and energy for an increasing population is a major challenge for modern agriculture (Foley et al., 2011). Current global food system concentrates on 1) a small number of crops (in 2013, wheat (Triticum spp.), rice (Oriza sativa), maize (Zea mays) which are responsible for almost 50% of the total harvested area in the globe) (FAOSTAT, 2013), and 2) farming practices based on high inputs of fertilisers and water (Foley et al., 2011). Though with an estimated population of almost 9 billion people by 2050, these crops might be close to their maximum potential production, and that means that further advances in productivity might take a nonviable quantity of resources. Still, there are around 7000 edible plants available for cultivation. These plants, some of them known as neglected and underutilised crops are: usually of indigenous origin, mainly cultivated in marginal areas, with little or none inputs and particularly secure due to their adaptation to the local environment. In general, there is a lack of proper varieties of these crops, and they are mainly cultivated using landraces, a mixture of genotypes locally adapted, that contains high levels of genetic variation showing a large range of productivity. Breeding programs have been established to

*

develop varieties of these crops, narrowing the genetic variation and promoting desirable agricultural traits (Padulosi et al., 2012; Mayes et al., 2012; Galluzzi and López Noriega, 2014).

In this context, a better understanding of how distinct dimensions such as geographic, environmental and anthropogenic interact and affect the evolutionary and ecological processes that lead to differentiation and genetic variation in species, is needed. In landscape genetic, the Mantel test and the partial Mantel test (Diniz-Filho et al., 2013) are commonly used but relate to finding associations between two dimensions, e.g. genetic and geography. The paper proposes to extend the Mantel test to more than two set of distances, allowing to test interactions such as genetic, geography, environmental conditions.

The results are based on a study of 33 landraces of bambara groundnut, an underutilised crop mainly cultivated in SubSaharan Africa and Southeast Asia. The genetic dataset is composed of the genotyping information about presence or absence of 20 Single Sequence Repeat (SSR) molecular markers of 33 landraces of Bambara groundnut, published by Molosiwa et al. (2015). From this data for 128 plants, the allele frequency for each group of landraces was produced as well as a matrix of genetic distance among each pair of landraces using the Neis genetic distance method (Nei, 1972) and the R software package adegenet (Jombart, 2008). The geographic distances among populations were calculated using a least cost path analysis using the package gdistance and linguistic distance was calculated using a method adapted from (Wichmann et al., 2013). More details on the methods for distance calculation are available in Santos et al. (2016). A primary analysis led to create two groups of these 33 landraces based on k-means clustering after a principal component analysis (PCA) of the allele frequency data Santos et al. (2016) .

2 k-Mantel test

The Mantel test or Knox test (Knox, 1964; Mantel, 1967) of correlation of two distance matrices was originally used in the health domain for space-time clustering of epidemics. The approach found great popularity in landscape genetic (Diniz-Filho et al., 2013; Guillot et al., 2009) to test genetic- geographic interaction. Across all pairs comparisons, Knox was working on co-occurrences in space and time whilst Mantel was using standardised distances; their statistic is commonly based on the sum of the products for each pair of dissimilarity metrics in the the two dimensions (as the co-occurrence within a given threshold for Knox (1964), and as a centred and reduced distance for Mantel (1967)). This can be generalised to k dimensions, where all the distances matrices are normalised between 0 and 1 (as in the rest of the paper):

$$M_k = \sum_p d_p^1 d_p^2 ... d_p^k \qquad \text{Equation 1}$$

where p runs for all $(n(n-1/2))$ distinct pairs i, j of n objects and d_p^k is the distance between i, j of that pair for the dimension k, *e.g.* geographical, genetic, semantic; $d_p^k = d_k(x_i^k, x_j^k)/max_{lm}(d_k(x_l^k, x_m^k))$ in which $d_k(x_i^k, x_j^k)$ or $d_{\{i,j\}}^k$ is the distance in the domain k for the pair of object i and j based on vectors of observations x_i^k and x_j^k. As in Mantel (1967) a permutation test can be performed by operating random permutations $\sigma_j(.)$ in each dimension over the vector pairs:

$$M_k(\sigma_k(p)) = \sum_p d_{\sigma_1(p)}^1 d_{\sigma_2(p)}^2 d_{\sigma_k(p)}^k \qquad \text{Equation 2}$$

giving the null distribution of the M_k statistic under the hypothesis of exchangeability, enabling to test 'an association' across the dimensions. Table 1 gives some results on 2, 3 4 dimensions for the 33 landraces dataset.

dimensions	null distribution of M_k min / max	observed M_{k_0}	(%max)	p-value
$k = 2$ gen.geo	120.9/127.2	130.5	(102.6%)	1e-04
$k = 2$ geo.ling	210.2/214.2	217.7	(101.6%)	1e-04
$k = 2$ geo.env	97.1/108.3	118.8	(109.7%)	1e-04
$k = 2$ gen.env	123.1/130.5	129.6	(99.3%)	0.0017
$k = 3$ env.gen.ling	113.9/120.9	122.1	(101%)	1e-04
$k = 3$ geo.gen.ling	111.5/117.6	123.8	(105.3%)	1e-04
$k = 3$ geo.gen.env	51.8/59.1	67.2	(113.7%)	1e-04
$k = 4$ geo.gen.env.ling	47.5/54.8	63.9	(116.6%)	1e-04

Table 1: Mantel M_k associations for 2, 3, 4 dimensions p-values using 9999 permutations, the d_c is the SVD compromise of all distances

3 Mantel associations tensors

Multidimensional analysis based on distance or directly on the measurements are also very popular in ecology and landscape ecology (Guillot et al., 2009; Jombart, 2008). In this section it is proposed to build relevant multiway data, arrays of 2 or more dimensions (a tensor) that are relevant to this Mantel-Knox extension approach over k dimensions. Multiway methods such as Principal Tensor Analysis (PTAk), multiway correspondence analysis (FCAk) (Leibovici, 2010) or non-negative tensor factorisation (NNTF) (Shashua et al., 2006) can be used to extract associations of fuzzy clustering properties multiple across dimensions. From the series of distances, a k-way dissimilarity tensor can be 'built' in a symmetric way on the pairs as in the following equation:

$$(TM_k)_{p_1 p_2 \ldots p_k} = d^1_{p_1} d^2_{p_2} \ldots d^k_{p_k} \qquad \text{Equation 3}$$

which is a rank one tensor, so already being decomposed as would do a PTAk decomposition for example (looking for a sum of rank one tensors). One may look for a symmetrical approximation $(d^c \otimes d^c \ldots \otimes d^c)$ for which d^c realises a compromise distance across the dimensions 1 to k. The symmetrical tensor optimisation, $argmax_{d^c}(\prod_{j=1}^{k} d^c d^j)$, is finding the d^c maximising simultaneously, in fact, each Mantel's statistic with every dimension, and this is equivalent to the first component of a PCA of the series of distance vectors. Note the symmetric tensor decomposition is no longer a rank one decomposition because of the symmetric constraint. One may also look for the non-negative matrix factorisation (NMF) to get a symmetric tensor approximation or using the NNTF of TM_k.

k-way dissimilarity tensors on the n objects instead of the pairs can be also analysed to describe the interactions between the domains, up to $k = 4$ here: geographical (geo), genetic (gen), environmental (env) and linguistic (ling), for example:

$$(TM_k^{(d_1,d_2,d_3)})_{rst} = d^1_r(s,t) d^2_s(r,t) d^3_t(r,s) \qquad \text{Equation 4}$$

where here $d^1_r(s,t) = \sum_{i=s,t} d^1_{\{r,i\}}$ idem for $d^2_s(r,t)$ and $d^3_t(r,s)$. This is an asymmetric tensor where the ways or modes of the tensor are 'specialised' into one of the dimensions (1 to k). Besides

describing associations across the domains, a d^c can be built from computing the Euclidian distance after extracting meaningful components from a tensor decomposition such as PTAk. Figure 1 shows

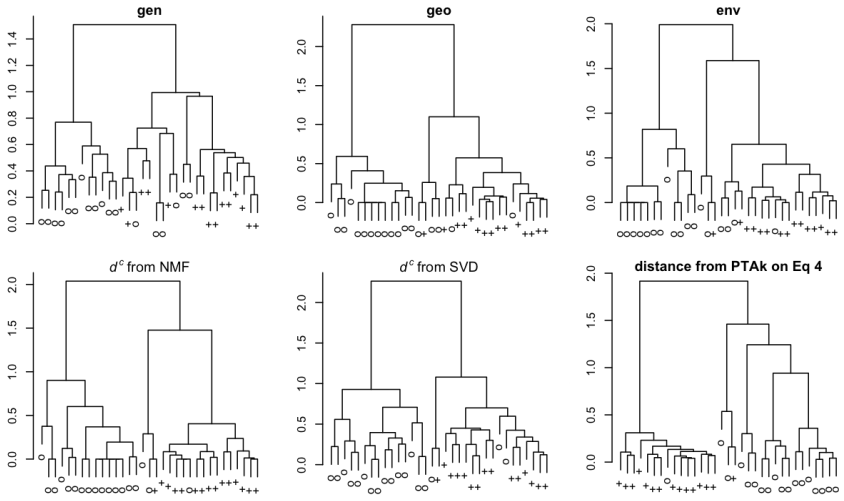

Figure 1: Example of d^c best symmetric dissimilarity approximation for the TM_k with product of the dimensions - 'gen', 'geo','env'- on pairs of 33 landraces labelled using gr: comparison of the dendrogram (Ward) for each distances and d^c compromises using SVD, NMF and PTAk on $TM_k^{(d_1,d_2,d_3)}$.

this best compromise association of the pairs across the dimensions for different methods. If the two groups o and + are relatively well separated in all dendrograms, the d^c ones are rendering more the compactness as well as classifying better these two groups. The landrace '792', a "+" is 'classified' with the "o" and "1276", a "o" is 'classified' withthe "+" on the NMF and PTAk dendrograms.

4 Mantel co-occurrences tensor

Co-occurrences and higher ($k > 2$) order co-occurrences have been used as approaches bringing extra constraints in analysing proximities which are relevant to spatial, spatio-temporal stucturing, *e.g.* clusters, outbreaks but also into multi-domains (Leibovici et al., 2014). k-across co-occurrences of order 2 can be defined, *i.e.* a pair of objects co-occurring in each of the dimension simultaneously:

$$(M_{koo})_p = 1_{d_p^1<\alpha_1} 1_{d_p^2<\alpha_2} ... 1_{d_p^k<\alpha_k} \qquad \text{Equation 5}$$

where the α_1, ... α_k are chosen thresholds of co-occurrence in each domain. Looking at its distribution over the pairs for a grouping factor (on the objects), *i.e.* proportion of M_{koo}'s, depicts the homogeneity of group associations from comparing the distributions when p is in one group or across two groups as in Figure 2, using as thresholds (*e.g.* α_1) the median of the distances for the dimensions. The "++" are more likely k-cooccurrent than the "oo" (using a 2x2 chi-square), so a more homogeneous group. Besides some "o" landing with the "+" the dendrograms on the compromises in Figure 1 was also showing this (NMF and PTAk) as the height of the jumps were on average lower for "+" (particular visible for the PTAk compromise d^c).

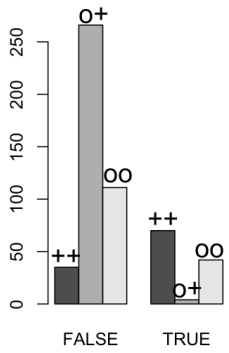

Figure 2: M_{koo}-histograms between groups and within groups for the 33 bambara landraces grouping and using 3-across co-occurrences: gen, geo and env: 116 out of 528 pairs (22%) are co-occurrent pairs (TRUE).

Using a similar approach to section 3, k-tensor symmetrical (Equation 6 and 7), and asymmetrical ones (Equation 8) can bring insight to the data from k-co-occurrences:

$$(TM_{koo})_{rst} = 1_{max_{rst}(d^1_{\{l,m\}})<\alpha_1} + 1_{max_{rst}(d^2_{\{l,m\}})<\alpha_2} + 1_{max_{rst}(d^3_{\{l,m\}})<\alpha_3} \quad \text{Equation 6}$$

where here $(TM_{koo})_{rst}$ expresses a dissimilarity to co-occurrence over the three spaces.

$$(TM_{koo})_{rst} = max_{rst}(d^1_{\{l,m\}}) + max_{rst}(d^2_{\{l,m\}}) + max_{rst}(d^3_{\{l,m\}}) \quad \text{Equation 7}$$

where here $(TM_{koo})_{rst}$ expresses a dissimilarity to co-occurrence over the three spaces (sum or product across the three spaces could be used).

$$(TM_{koo})^{(d_1,d_2,d_3)})_{rst} = max(d^1_{\{r,s\}}, d^1_{\{r,t\}}) + max(d^2_{\{s,r\}}, d^2_{\{s,t\}}) + max(d^3_{\{t,r\}} d^3_{\{t,s\}}) \quad \text{Equation 8}$$

5 Conclusion

The principle within the Mantel correlation testing from distance matrices can be easily extended to multiple dimensions as well as producing dissimilarity tensors either from second order properties or higher-order such as the k-co-occurrences. Many more second order dissimilarity tensors, k-across co-occurrences and k-co-occurrences over k dimensions tensors producing dissimilarities across or over multiple dimensions can be analysed from existing multiway method, giving different insights. Illustrative analysis using the 33 bambara groundnut landraces data example gave some encouraging results that will be extended for the conference presentation.

6 References

Diniz-Filho, J. A. F., T. N. Soares, J. S. Lima, R. Dobrovolski, V. L. Landeiro, M. P. d. C. Telles, T. F. Rangel, and L. M. Bini
2013. Mantel test in population genetics. *Genetics and molecular biology*, 36(4):475–485.

FAOSTAT, D.
2013. Food and agriculture organization of the united nations. statistical database.

Foley, J. A., N. Ramankutty, K. A. Brauman, E. S. Cassidy, J. S. Gerber, M. Johnston, N. D. Mueller, C. OConnell, D. K. Ray, P. C. West, et al.
2011. Solutions for a cultivated planet. *Nature*, 478(7369):337–342.

Galluzzi, G. and I. López Noriega
2014. Conservation and use of genetic resources of underutilized crops in the americasa continental analysis. *Sustainability*, 6(2):980–1017.

Guillot, G., R. Leblois, A. Coulon, and A. C. Frantz
2009. Statistical methods in spatial genetics. *Molecular Ecology*, 18(23):4734–4756.

Jombart, T.
2008. adegenet: a r package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11):1403–1405.

Knox, G.
1964. The detection of space-time interactions. *Applied Statistics*, 13:25–29.

Leibovici, D. G.
2010. Spatio-temporal multiway decompositions using principal tensor analysis on k-modes: The r package ptak. *Journal of Statistical Software*, 34(10):1–34.

Leibovici, D. G., C. Claramunt, D. Le Guyader, and D. Brosset
2014. Local and global spatio-temporal entropy indices based on distance-ratios and co-occurrences distributions. *International Journal of Geographical Information Science*, 28(5):1061–1084.

Mantel, N.
1967. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27:209–220. Mantel N, Valand RS (1970) A technique of nonparametric multi- variate analysis. Biometrics, 26, 547558.

Mayes, S., F. Massawe, P. Alderson, J. Roberts, S. Azam-Ali, and M. Hermann
2012. The potential for underutilized crops to improve security of food production. *Journal of experimental botany*, 63(3):1075–1079.

Molosiwa, O. O., S. Aliyu, F. Stadler, K. Mayes, F. Massawe, A. Kilian, and S. Mayes
2015. Ssr marker development, genetic diversity and population structure analysis of bambara groundnut [vigna subterranea (l.) verdc.] landraces. *Genetic Resources and Crop Evolution*, 62(8):1225–1243.

Nei, M.
1972. Genetic distance between populations. *The American Naturalist*, 106(949):283–292.

Padulosi, S., N. Bergamini, T. Lawrence, et al.
2012. On farm conservation of neglected and underutilized species: status, trends and novel approaches to cope with climate change: Proceedings of an international conference, frankfurt, 14-16 june, 2011. *Bioversity International, Rome*.

Santos, R., A. Algar, R. Field, and S. Mayes
2016. Integrating giscience and crop science datasets: a study involving genetic, geographic and environmental data. Technical report, PeerJ Preprints.

Shashua, A., R. Zass, and T. Hazan
2006. Multi-way clustering using super-symmetric non-negative tensor factorization. In *European conference on computer vision*, Pp. 595–608. Springer Berlin Heidelberg.

Wichmann, S., A. Müller, A. Wett, V. Velupillai, J. Bischoffberger, C. Brown, et al.
2013. The asjp database (version 16). available online at: h ttp. *asjp. clld. org*.