Integrating GIScience and Crop Science datasets: a study involving genetic, geographic and environmental data

Roberto Santos[1,2], **Adam Algar**[1], **Richard Field**[1], **and Sean Mayes**[3]

[1]**School of Geography, University of Nottingham, UK**
[2]**Nottingham Geospatial Institute, University of Nottingham, UK**
[3]**Plant and Crop Sciences, University of Nottingham, UK**

ABSTRACT

Sharing and reusing data in research is a welcome and encouraged practice since it maximises the scientific outcomes given limited financial, material and human resources. Interdisciplinary research is usually benefitted from this practice, reuniting researchers and data from two or more disciplines to advance fundamental understanding or tackle problems whose solution is beyond the limit of an individual body of knowledge. In this work, we discuss the challenges of associating localisation with other data types, particularly genetic data, in a project involving Crop Science and Geospatial Information Science disciplines that aim to improve the understanding of how geographical, environmental and anthropocentric factors affect the genetic variation in a neglected and underutilised crop called Bambara groundnut.

Keywords: Data integration, GIScience, Crop Science, Landscape Genetics, Bambara groundnut

INTRODUCTION

The research challenges in the 21st century require an interdisciplinary approach to advance fundamental understanding or to tackle problems which solution are beyond the scope of a single discipline or body of knowledge (Academies, 2004). Interdisciplinarity is the current reality in the life of many scientists, through the composition of research teams and the nature of the hypotheses being examined (Dyer, 2015), and funding organisations have been increasing the support for research integrating multiple disciplines. In this context, sharing and reusing datasets from different disciplines is a common practice and aims to strength the research by adding new dimensions to the data available or verifying the results obtained from a different perspective. However, integration of discipline associated datasets is not a smooth process and is subject to associated concepts of quality and abundance.

Geographic Information Science (GIScience) explores the location property of facts such as objects, events and processes, associating them with data, such as longitude and latitude (Goodchild, 2010; Stevens and Pfeiffer, 2015). This process may seem straightforward given the increasing presence of location based sensors and mapping technologies in our daily life. However, it is still a challenge given the nature and conditions of the facts examined. In natural museums, artefacts and specimens have usually been collected over the course of many years and their finding location is often ambiguous or very imprecise (van Erp et al., 2015). In Health Science, investigation of historical records of disease occurrence in individuals relies on the name of locations and vague addresses. The lack of a historical address database undermines the potential use of those health data (Lash et al., 2012). In Crop Science, breeding programs rely on the collections of seeds available to germplasm banks. Knowledge about the origin of theses seeds is necessary to characterise the environment around their collect location and go beyond the basic measurements of diversity. However, as natural museums, location information about their origin may be vague or associated with markets where the seeds were collected instead of their grown site (Richards, 2011).

A MODEL FOR INTEGRATION OF MULTIPLE DISCIPLINE DATA

In the previous examples, we showed cases where integration of data from distinct disciplines with GIScience was not straightforward. In fact, the integration often involves transformation and filtering operations to attend some criteria. Given two datasets from different disciplines, these operations usually discard or reduce records or items, defining a subset of the original dataset that can be used in both disciplines. In Figure 1, a Venn diagram shows this concept using three distinct disciplines. Given the initial amount of data in a discipline (D1), only a subset (D1 $\cap$ D2 or D1 $\cap$ D3) that attends the criteria of both disciplines (D1 and D2 or D1 and D3) can be used. As the number of disciplines involved increases, also increases the criteria to be attended.

These criteria involve requirements common and unique to the disciplines involved. These criteria might include the following cases. Exclusion of records with missing data (i.e. records must have longitude and latitude or postcode). Temporal scales (i.e. growing season must match respective weather data to investigate potential plant stress). Spatial scales (i.e. association of soil and disease resistance among individuals in a small to medium farm demands detailed soil maps). Sample units (i.e. given the location is available at the population level; the genetic information of individuals must be grouped by the same definition of the population), among others.

Another aspect of the integration of data from distinct disciplines is the concepts of abundance and diversity. A dataset may attend the criteria of abundance and diversity in the discipline D1 and scarce and uniform in the discipline D2. Figure 2A shows one seedling being planted in the glasshouse. In the context of Crop Science discipline, this individual has potential to generate a significant amount of genetic data through genotyping or sequencing processes (see figure 2C), and the resulting dataset could be considered abundant. The interesting point is that genetic information among individuals of the same species is very similar and in order get data that represent the differences it is important to choose high polymorphic molecular markers in the genotyping process. In the context of GIScience, this individual which origin information is available at the population level, has one location data associated and could be considered scares if the objective would analyse the genetic data over a broad geographic area. Even if more seedlings from the same population were cultivated (see figure 2B), the number of locations is still one, and all the genetic data generated from these samples are still associated with one point in space (see figure 2D). Cultivating seeds from distinct population could provide a better representation of the geographic space. However, it would be necessary investigate the characteristics of interest around this location in order the guarantee a reasonable representation of the varies categories. Experiment design and sampling strategy are important and should be discussed taking into account the characteristics of data of the disciplines involved.

THE BAMBARA GROUNDNUT STUDY CASE

Bambara groundnut is a neglected and underutilised species (NUS) of legume, mainly cultivated in Sub-Saharan Africa. It is believed that the process of domestication and further cultivation started thousands of years ago. And, despite its long history, this crop is still cultivated from landraces (locally developed mixtures of genotypes) (Molosiwa et al., 2015).

We used genetic and geographic datasets to explore potential spatial patterns of genetic variation. Further, we included environmental datasets in order to analyse distinct measures of distance (geographic, genetic and environmental) of Bambara groundnut.

The genetic dataset was composed of the genotyping information about the presence of absence of twenty Single Sequence Repeat (SSR) molecular markers of 33 distinct landraces of Bambara groundnut, having a total of 128 samples. We calculated the allele frequency for each group of landraces and based on the allele frequency we produced a matrix of genetic distance among each pair of landraces using Nei's genetic distance method (Nei, 1972). The geographic information about the origin of these seeds was provided by the International Institute of Tropical Agriculture (IITA) at landrace level. The seeds were cultivated in glasshouses in England and Botswana (Molosiwa et al., 2015). We used the WorldClim database (Hijmans et al., 2005) to characterise temperature, rainfall and altitude around the origin locations (see Figure 3).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2248v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

CONCLUSIONS

So far exploratory analyses have been conducted using the datasets and process presented. We identified some level of imprecision on location data that did not affect the initial analyses (mostly PCA and k-means of the genetic data and mapping of the first and second axis); however, future comparisons among matrices of distance may impose the exclusion of most compromised samples. The process was challenging and a valuable experience. Integrating datasets from different disciplines involved a better understanding of the concepts and criteria of the disciplines involved. We plan to advance in the analysis including different molecular markers and proxy variables to assess the anthropocentric factors that might have affected the genetic variation as well as least cost path analysis instead of the great circle distance.

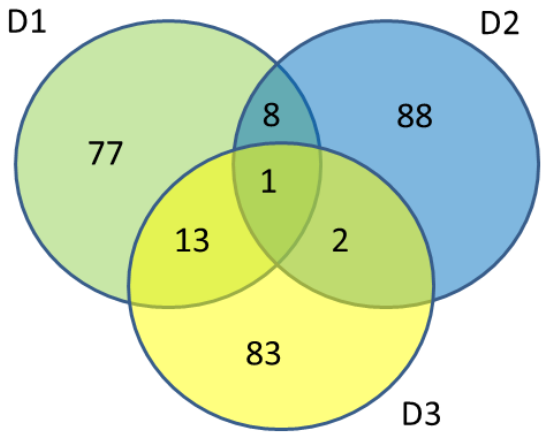

Figure 1. A model for integration of multiple discipline data.

REFERENCES

Academies, N. (2004). *Facilitating Interdisciplinary Research*. National Academies Press, Washington, District of Columbia.

Dyer, R. J. (2015). Is there such a thing as landscape genetics? *Molecular Ecology*, 24(14):3518–3528.

Goodchild, M. F. (2010). Twenty years of progress: GIScience in 2010. *Journal of Spatial Information Science*, (1).

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15):1965–1978.

Lash, R., Carroll, D. S., Hughes, C. M., Nakazawa, Y., Karem, K., Damon, I. K., and Peterson, A. (2012). Effects of georeferencing effort on mapping monkeypox case distributions and transmission risk. *International Journal of Health Geographics*, 11(1):23.

Molosiwa, O., Aliyu, S., Stadler, F., Mayes, K., Massawe, F., Kilian, A., and Mayes, S. (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.

Richards, G. M. V. C. M. (2011). Integration of Georeferencing, Habitat, Sampling, and Genetic Data for Documentation of Wild Plant Genetic Resources. *HortScience*, 46(11):1446–1449.

Stevens, K. B. and Pfeiffer, D. U. (2015). Sources of spatial animal and human health data: Casting the net wide to deal more effectively with increasingly complex disease problems. *Spatial and Spatio-temporal Epidemiology*, 13:15–29.

van Erp, M., Hensel, R., Ceolin, D., and van der Meij, M. (2015). Georeferencing Animal Specimen Datasets. *Transactions in GIS*, 19(4):563–581.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2248v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

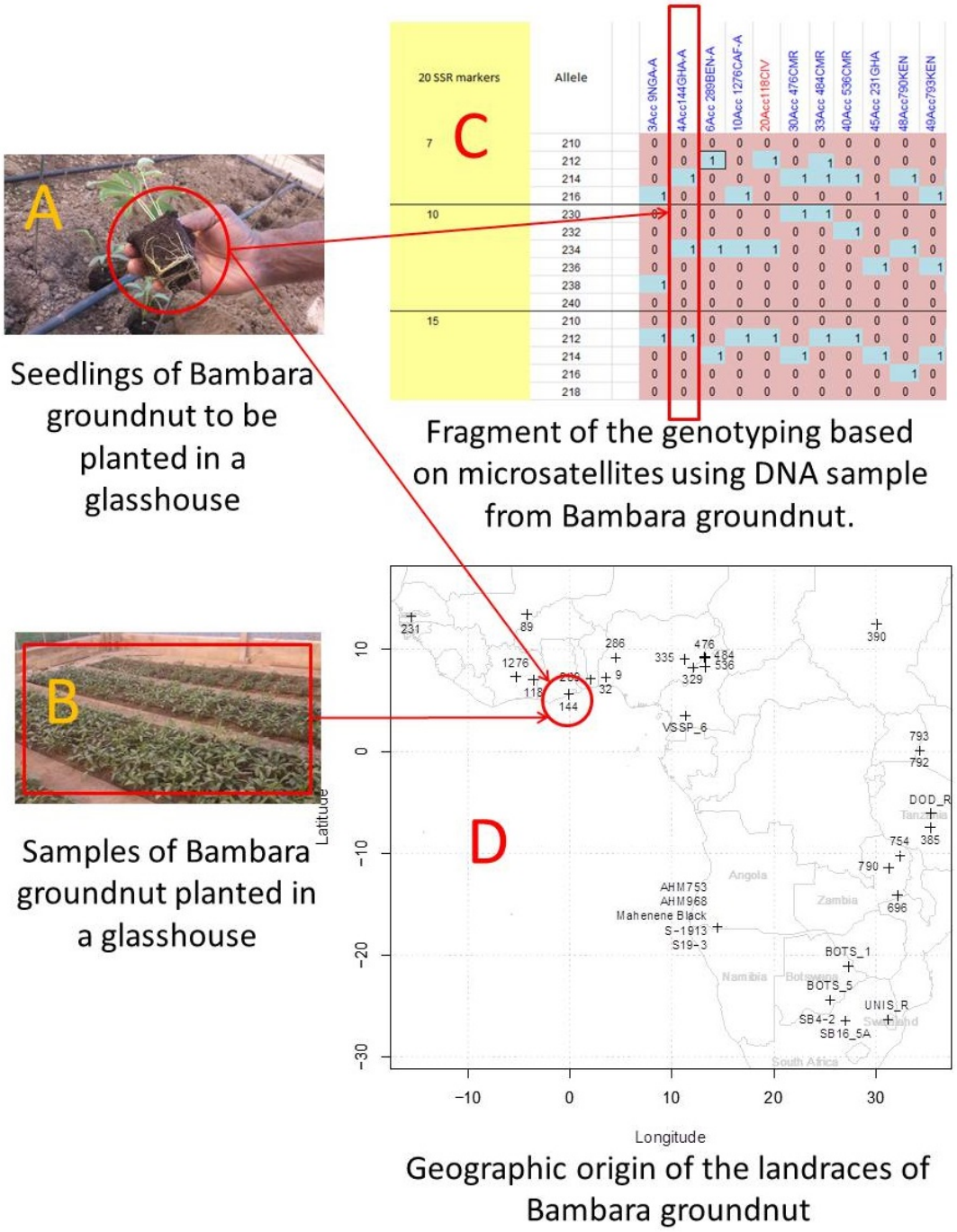

Figure 2. Integration of geographic and genetic datasets of Bambara groundnut. (A) shows a seedling to be planted in the glasshouse. (B) shows a trial of Bambara groundnut planted in the glasshouse. (C) genotyping results with highlighted data of a specific sample. (D) geographic localisation of the landraces (populations) of Bambara groundnut used in this study.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2248v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

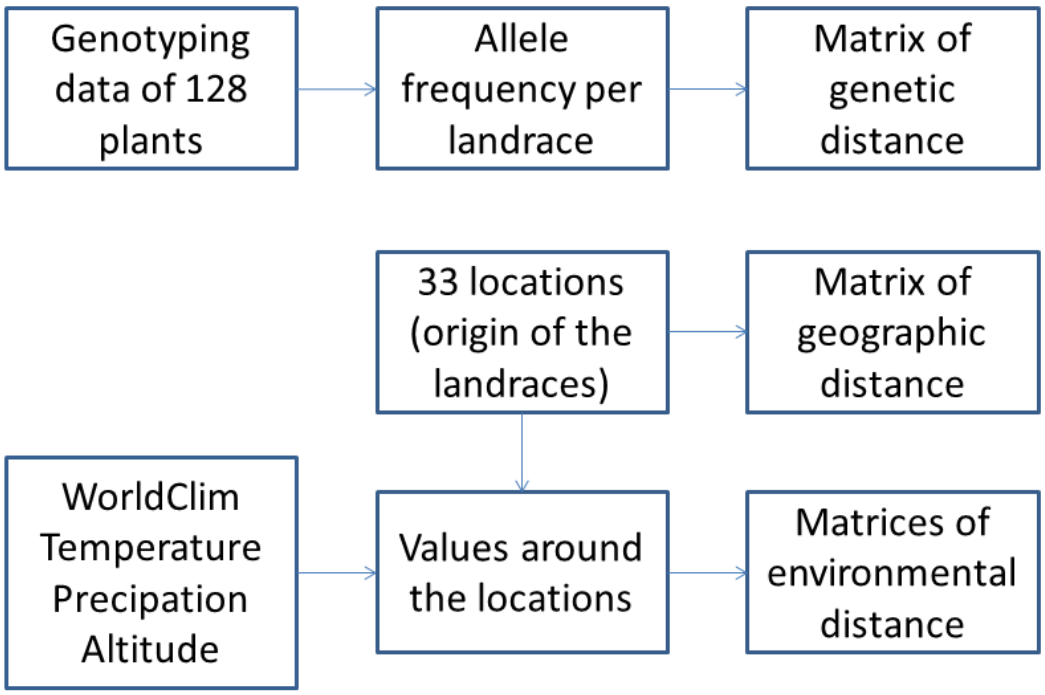

Figure 3. Flow of transformations and filter operations applied to the genetic, geographic and environmental data.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2248v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016