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 considered to benefit from this practice, uniting 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. Here we discuss the challenges of combining data across disciplines, focusing in particular on associating geographic location data with genetic data in the context of a project involving Crop Science and Geospatial Information Science disciplines. This project aims to improve 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

Research challenges in the 21st century require an interdisciplinary approach, to advance fundamental understanding or to tackle problems whose solutions are beyond the scope of a single discipline or body of knowledge (Academies, 2004). Interdisciplinarity is encouraged, with funding organisations recently increasing support for research that integrates multiple disciplines. Interdisciplinary science represents the current reality for increasing numbers of scientists, through the composition of research teams and the nature of the hypotheses being examined (Dyer, 2015). In this context, sharing and reusing datasets from different disciplines is 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 varying concepts of quality and abundance.

Geographic Information Science (GIScience) explores the location property of entities such as objects, events and processes, associating them with co-ordinates, 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 lives. However, it is still a challenge because of the nature and contexts of the facts examined. In museums and herbaria, artefacts and specimens have usually been collected over decades or centuries 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 names of locations and vague addresses. The lack of historical address databases 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 these seeds is necessary to characterise the environment around their collection location, and to go beyond basic measurements of diversity. However, again, location information about their origin may be vague or associated with markets where the seeds were obtained instead of where they grew or were originally sourced from (Richards, 2011).

A MODEL FOR INTEGRATION OF MULTIPLE DISCIPLINE DATA

In the cases just mentioned, integration of data from other disciplines with GIScience is not straightforward. In fact, the integration often involves transformation and filtering operations using arbitrary criteria. Given two datasets from different disciplines, these operations typically discard or reduce records or items, often greatly, resulting in a relative 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) fits the criteria of both disciplines (D1 and D2 or D1 and D3). As the number of disciplines involved increases, so there is further diminution of the available data.

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

Another issue with integrating data from distinct disciplines concerns variation of concepts such as abundance and diversity. A dataset may fit criteria of relatively high abundance and diversity in discipline D1, but be classed as scarce and uniform in discipline D2. Figure 2A shows one seedling being planted in the glasshouse. In the context of Crop Science, this individual has the 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. However, genetic information among individuals of the same species can be very similar and in order to get data that represents the differences it is important to choose highly polymorphic molecular markers in the genotyping process. In the context of GIScience, this individual seedling, whose origin information is available at the population level, has only one associated location, and would be considered scarce if the objective were to 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 population 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 different populations would provide a better representation of the geographic space. However, it would be necessary to investigate the characteristics of interest around this location in order to guarantee a reasonable representation of environmental variables. Experiment design and sampling strategy are important, and should be discussed taking into account the characteristics of the data of the disciplines involved.

THE BAMBARA GROUNDNUT STUDY CASE

Bambara groundnut is classed as a neglected and underutilised species of legume, mainly cultivated in Sub-Saharan Africa. It is believed that the process of its domestication and further cultivation started thousands of years ago. However, 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 spatial patterns of genetic variation. We included environmental datasets in order to analyse distinct measures of distance (geographic, genetic and environmental) of Bambara groundnut landraces. The genetic dataset was composed of the genotyping information about the presence or absence of twenty Single Sequence Repeat (SSR) molecular markers of 33 distinct landraces of Bambara groundnut, with 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). Although the environmental and genetic dataset were large, putting the two together led to a small dataset that was only just big enough to analyse.

So far, we have conducted exploratory analyses using the datasets and process presented here. We identified some imprecision in the location data that did not affect the initial analyses (mostly PCA and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2248v2 | CC BY 4.0 Open Access | rec: 6 Jul 2016, publ: 6 Jul 2016

k-means cluster analysis of the genetic data and mapping of the first and second axes); however, future analyses of distance matrices may require the exclusion of the most compromised samples. Although the environmental and genetic datasets were large, putting the two together led to a small dataset that was only just big enough to analyse.

CONCLUSIONS

If, as a scientific community, we are serious about interdisciplinarity then we need a lot more work in co-ordinating data-collection activities, to guarantee the data acquired are useful for all disciplines involved. We propose that existing research on interoperability, an established concept in GIScience, be extended to other areas of science, and particularly the co-ordination of data collection. It holds potential for helping to address the challenges presented in the integration of multidisciplinary data.

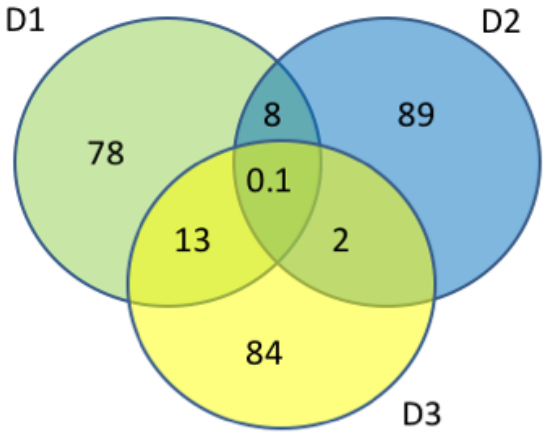

Figure 1. This figure shows a model for integration of data from distinct disciplines. Circles in the Venn diagram represent various disciplines (D1, D2 and D3) and their respective data criteria. Numbers represent a hypothetical amount (%) of the data that only fit the criteria of each discipline (D1 = 78, D2 = 89 and D3 = 84), or the combined criteria of two disciplines (D1 ∩ D2 = 8, D1 ∩ D3 = 13, D2 ∩ D3 = 2), or the combined criteria of all disciplines (D1 ∩ D2 ∩ 3 = 0.1). Given two or more datasets from different disciplines, the combined criteria typically discard or reduce records or items, often greatly, resulting in a relatively small subset of the original datasets that can be used by the combined bodies of knowledge.

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.

Nei, M. (1972). Genetic Distance between Populations. *The American Naturalist*, 106(949).

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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2248v2 | CC BY 4.0 Open Access | rec: 6 Jul 2016, publ: 6 Jul 2016

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.2248v2 | CC BY 4.0 Open Access | rec: 6 Jul 2016, publ: 6 Jul 2016

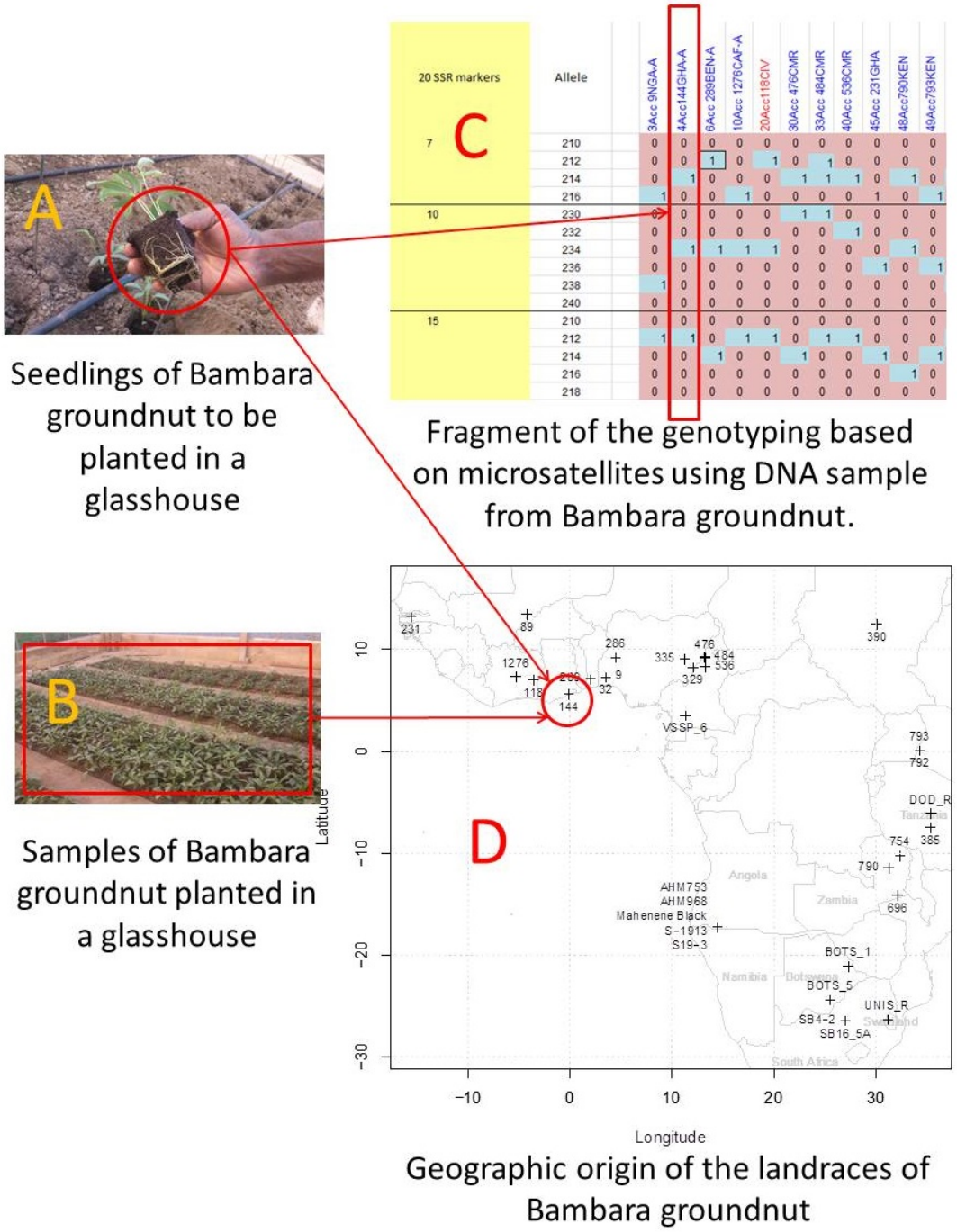

Figure 2. Integration of geographic and genetic datasets of Bambara groundnut. (A) shows a seedling to be planted in the glasshouse. (B) shows a trial of Bambara groundnut planted in the glasshouse. (C) genotyping results with highlighted data of a specific sample. (D) geographic localisation of the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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2248v2 | CC BY 4.0 Open Access | rec: 6 Jul 2016, publ: 6 Jul 2016

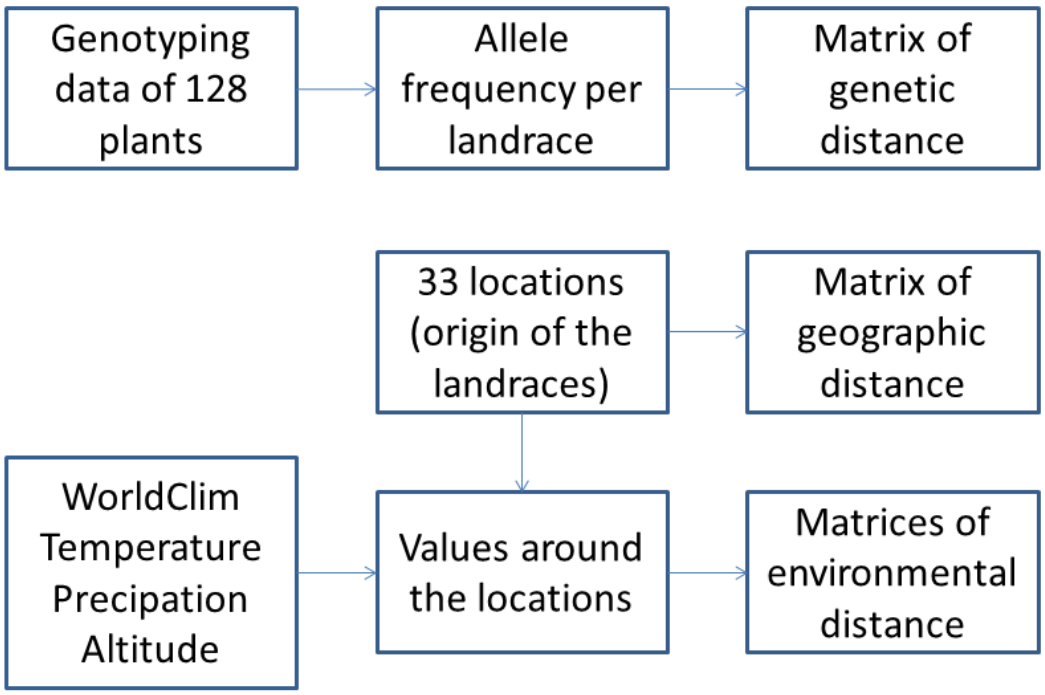

Figure 3. Flow of transformations and filter operations applied to the genetic, geographic and environmental data.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2248v2 | CC BY 4.0 Open Access | rec: 6 Jul 2016, publ: 6 Jul 2016