Spatial analysis of historical Indian cranial measurements

GEOG 897 Spatial Analysis with R

Fall 2021

Andrew Laws

Introduction

In the field of forensic anthropology, the measurements of human bone sections have been used in the identification of human remains, to track the spatial and temporal migrations of humans, and to understand when genetic intermixing between cultural or ethnic groups occurred (Spradley, 2013, 2014). This is accomplished with the help of databases containing measurements of bones from populations with known spatial and temporal locations. But not all measurements are created equal for each of the above tasks or even for each gender. Therefore, forensic anthropologist need not only accurate and uniform data collection, but the establishment of each measurements use case.

The area of study for this pilot project is currently found in modern-day India and Sri Lanka. When the data was collected in 1909, the area was under the governance of Great Britain. To further the understanding of human spatial patterns in and around our study area, this project seeks to understand the clustering associated with cranial measurements.

Background

There are three initial components when a forensic anthropologist is building the biological profile of human remains, which are the estimation of sex, height, and ancestry (Spradley, 2013; Spradley et al., 2008). To perform these estimations, scientists rely on known reference data that has been shown to differentiate these criteria (Spradley, 2013; Spradley et al., 2008). However, when reference data is incomplete or lacking, the result can be poor estimations due to ancestral and population differences (Spradley et al., 2008). This further highlights the need for quality reference data from diverse groups of people (Spradley, 2013; Spradley et al., 2008).

The most common method for creating reference data is to take morphological measurements (cranial measurements are a subset of this) and use a discriminant function analysis (DFA) to determine the primary identifiers (Ross & Williams, 2021, 2021; Spradley, 2013; Spradley et al., 2008). In fact, the DFA has been seen as the traditional craniometric methodology (Spradley, 2013) alongside other statistical parametric methodologies in the forensic sciences (Dunn et al., 2020). There has generally been a lack of spatial analysis methodologies applied in the field. When spatial analysis does occur, it is normally done after other non-spatial statistical methods have identified significant identifiers of origin, with the main computed statistic being the Moran's I, a measure of spatial autocorrelation (Harding, 1990; Maestri et al., 2016).

Methods

Based on the background and previous research, the following methodology was devised to answer the research question.

Technology

R-Studio (version 1.4.1717) was the project IDE and R (version 4.1.1) was the base programming language for this project. The following primary, non-base packages (versions in parentheses) were used: sf (1.0-2), tmap (3.3-2), tidyverse (1.3.1), readxl (1.3.1), rgeoda (0.0.8-6), spdep (1.1-11), stringr (1.4.0), gt (0.3.1), and webshot (0.5.2). ArcGIS Online was used for limited geoprocessing. Microsoft Excel was used to prepare the tabular data for ingestion.

Tabular and Spatial Data Procurement/Pre-processing

The tabular data for this project is from the data collection conducted and recorded in Indian Museum, 1909. This was an electronic scan of a paper media and was found in the University of Toronto Library online archives. This was procured by the principal investigators, Dr. William Belcher (University of Nebraska-Lincoln) and Dr. Joe Hefner (Michigan State University). Dr. Belcher transcribed the information found in Gupta 1909 into an Excel spreadsheet. The cranial measurements included the glabello-occipital length, max cranial breadth, basi-bregmatic height, minimum frontal diameter, stephanic asterionic, frontal longitudinal chord, parietal longitudinal chord, occipital longitudinal chord, length of foramen magnum, basi-nasal length, basi-alveolar length, interzygomatic breadth, mid-orbital width, nasio-alveolar length, nasal height, nasal width, orbital width, orbital height, palato-maxillary length, and palato-maxillary breadth.

A review of the spatial extent of each observation showed that the states and regions[1] in India and Sri Lanka had changed since 1909 due to geopolitical differences i.e., independence from Great Britain. While the Sri Lankan border was relatively stable, a simple crosswalk was performed using the names of Indian states and not a spatial overlay. This was performed using 2020 Indian state and territory data from a Wikipedia article detailing the transformation of Indian states and territories. While Wikipedia is not an ideal or overall trustworthy source of data, the article provided the clearest description of these processes with regards to language considerations. After performing the crosswalk, an identifier column was added with the state's corresponding, non-hyphenated International Organization for Standardization (ISO) 3166 code. A final review and correction of column headers was completed to ensure uptake into the data pipeline.

Procurement of spatial data for Indian states came from ArcGIS Online. This was accomplished by performing a union between the Indian State Boundaries 2020 (ESRI, 2021) and a polygon that overlayed the entire county then downloading the resulting feature layer as a shapefile. The Sri Lankan data was acquired from the GADM website, an open-source data repository (GADM, 2021), in the shapefile format.

[1] *Collectively referred to as state or states except when originating country is relevant*

Data Ingestion

Spatial data was read in as a layer of class sf and it's usability ensured using st_make_valid from the *sf* package. Spatial data were all transformed to the projected coordinate system Asia South Albers Equal Area Conic (ESRI 102028). This was chosen as the CRS due to its limiting of distortion in the area of interest. The Sri Lankan data was mutated to include the ISO 3166 code. The spatial extents were then combined and unnecessary columns dropped (assigned to *aoi*). Finally, rows with ISO code INAN were dropped, with the reasoning for this highlighted in the **Results** section. The tabular data was read in, unnecessary columns dropped, and rows with ISO code INAN were dropped (assigned to *cran*).

Exploratory Spatial Data Analysis (ESDA)

Both the spatial and tabular data were visualized. The spatial data was visualized as a map to see extent. The tabular data was plotted in a histogram of each cranial measurement. Additionally, the dispersal of observations into sex and state bins was tallied.

Two cranial measurements, glabello-occipital length (GOL) and max cranial breadth (XCB) were chosen to test neighbor formalization and subset from *cran*. With multiple observations normally distributed per geometry, the records were grouped by ISO code and aggregated using two mathematical functions, mean and median (i.e. GOL.mean and GOL.median). As the geometries for the polygons in *aoi* were non-contiguous, two nearest-neighbor methods were tested: distance and k-nearest neighbor (KNN). For the distance neighbor, a minimum threshold was calculated using the *rgeoda* library and multiplied by 1.25 to ensure each geometry had at least one neighbor before a row-standardized list of weights was calculated. For the KNN, the *k* was tested at values of 2 and 3 before a row-standardized list of weights was calculated. The result of this testing led to the selection of the distance nearest neighbor method of formalization due to lower p values.

Spatial Autocorrelation

The *cran* dataset was grouped by ISO code and each cranial measurement aggregated using the *mean* and *median* functions inside the *summarize_all* function (assigned to *cran.stats*) then joined with *aoi* to create the sf data.frame *cran.sf.* The list of weights was calculated using the distance nearest neighbor described above.

The Moran's I was calculated using a custom function, *all.moran*, that was developed with help and testing from a colleague (Sebastiano de Bona, personal communication, November 12, 2021). The function has one parameter, which is an aggregated cranial measurement column, runs the function *moran.test* from the *spdep* library, and returns a dataframe containing the Moran's I score and p-value if the p-value is significant (<0.05). To map *all.moran* over the columns, the geometry of *cran.sf* is dropped (assigned to *cran.df*) and the *map_dfr* function from the *purr* library handles the mapping. The output is a dataframe with the statistically significant Moran's I values of the aggregated cranial measurements (assigned to *moran.stats*).

To validate the findings in *moran.stats*, a second function was created. *all.moranmc* replaces *moran.test* with *moran.mc*, which runs a Monte Carlo simulation. *moran.mc* was run with 10,000 iterations due to examples from the course textbook using that amount (Brunsdon, 2018). The function returns statistically significant (<0.05) Moran's I and is mapped over *cran.df*, similar to *moran.stats* (assigned to *moranmc.stats*). Values that were in the congruous to both *moran.stats* and *moranmc.stats* were retained for future work.

The Local Indicators of Spatial Autocorrelation (LISAs) were calculated for the two aggregated cranial measurements with the lowest p-values in *moranmc.stats*. A list of weights was created using distance nearest neighbor and the 1.25 times the minimum distance threshold. The resulting LISAs were mapping.

Results

Spatial Analysis

Tabular data preprocessing led to a final dataset that included 138 observations with 18 variables. The observations were split based on sex as 18 female, 89 male, and 21 unknowns. Spatial data preprocessing resulted in 37 total geometries. The final aggregated and joined sf dataframe *cran.sf* contained 12 geometries with 34 aggregated cranial measurements.

Results from the ESDA showed that the Andaman and Nicobar Islands (ISO code INAN) were approximately 1243 km from their nearest neighbor and resulted in no statistically significant Moran's I values from *moran.test*. Therefore, the related records were dropped from all data (Figure 1). With INAN out of the data, the minimum distance threshold became approximately 449 km. The histogram of each cranial measurement showed that each was normally distributed (Figure 2). The results of the neighbor formalization were described in the methodology section to ensure clarity for the rationale.

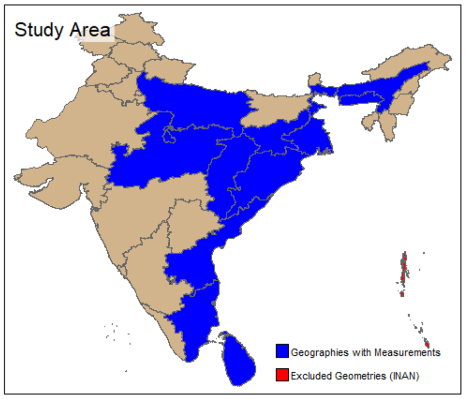


Figure 1. Map of study area

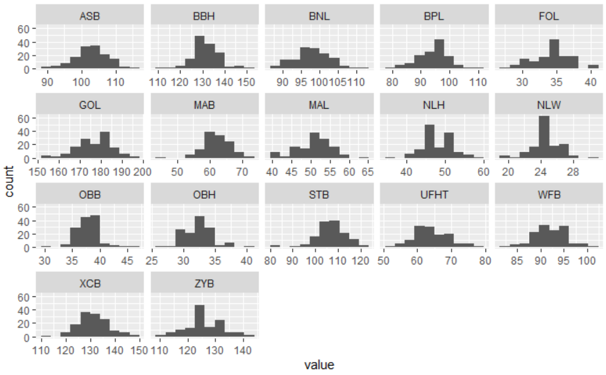
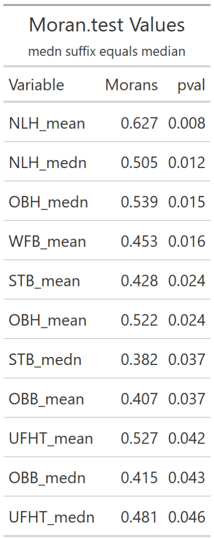


Figure 2. Histograms of cranial measurements

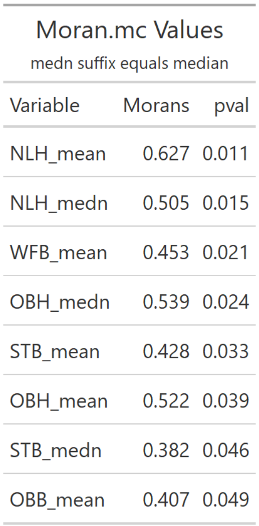
The results from the Moran test showed 11 aggregated cranial measurements (6 cranial measurements) with Morans I values ranging from 0.382 to 0.627 and p-values ranging from 0.008 to 0.046 (Figure 3). The results from most of the Monte Carlo Moran tests showed 8 aggregated cranial measurements (5 cranial measurements) with Morans I values ranging from 0.382 to 0.627 and p-values ranging from 0.011 to 0.049 (Figure 4). Of note, the Monte Carlo Moran test would occasionally not return the aggregated cranial measurement OBB_mean (p-value = 0.049) due to the iterative nature of the function and borderline p-value. In both test results, mean and median values for nasal height (NLH) had the highest Moran's I values and lowest p-values. These results were likely influenced due to 2 geometries having no NLH values. The calculated LISA's for these (NLH_mean, NLH_medn) are shown in Figure 5.



Moran.test Values

medn suffix equals median

Variable	Morans	pval
NLH_mean	0.627	0.008
NLH_medn	0.505	0.012
OBH_medn	0.539	0.015
WFB_mean	0.453	0.016
STB_mean	0.428	0.024
OBH_mean	0.522	0.024
STB_medn	0.382	0.037
OBB_mean	0.407	0.037
UFHT_mean	0.527	0.042
OBB_medn	0.415	0.043
UFHT_medn	0.481	0.046



Moran.mc Values

medn suffix equals median

Variable	Morans	pval
NLH_mean	0.627	0.011
NLH_medn	0.505	0.015
WFB_mean	0.453	0.021
OBH_medn	0.539	0.024
STB_mean	0.428	0.033
OBH_mean	0.522	0.039
STB_medn	0.382	0.046
OBB_mean	0.407	0.049

Figure 3: Result of Moran.test Figure 4: Result of Moran.mc

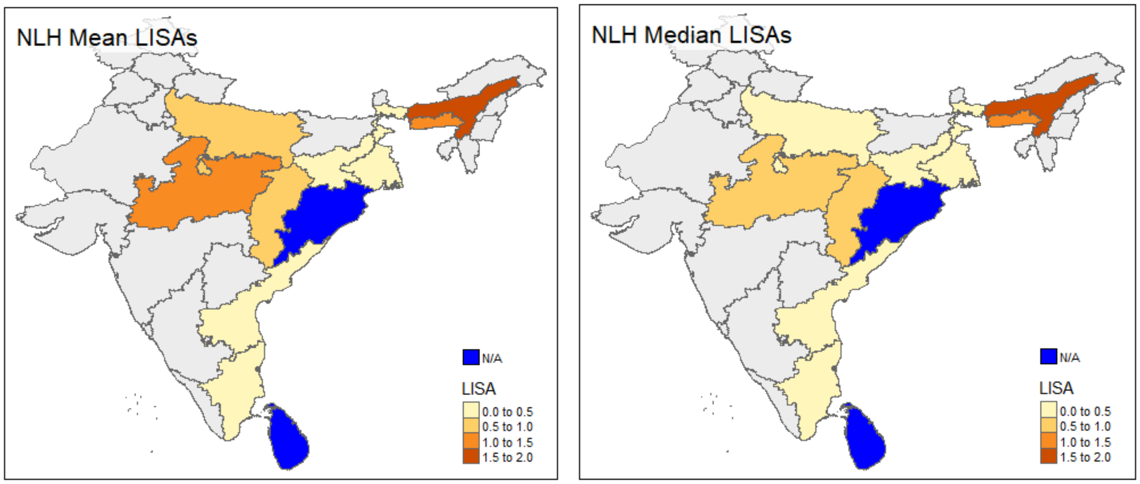


Figure 5: LISAs for NLH mean and median.

Project Success

Looking back on the project proposal, I met many of the goals I set out to, even if they required some rewriting along the way. My research question was refined to include spatial autocorrelation and the spatial scale altered to the state-level due to data constraints. I learned more about the statistical and spatial methodologies used in forensic anthropology as well the interesting ways they are used to track human movement. With a new tabular dataset containing more spatially granular data coming from Dr. Belcher in the near future, the use of R-Studio and R will allow me to readily replicate the entire process in a speedy and efficient manner. The greatest challenge was with the data. The tabular data was in its first draft, with a final dataset coming in Spring 2022. Spatial data from the Indian and Sri Lankan governments presented some unique challenges. The Indian government does not have an open data portal and the data is only accessible for free to residents of India. Otherwise, it carries a price tag of at least $1000 depending on which datasets are requested. The Sri Lankan government data can be easily accessed using an application programming interface (API) but the use policy for this data is not clearly spelled out.

Discussion

At the state-level, there is evidence of low to moderate spatial clustering with a subset of the aggregated cranial measurements, with NLH measurements showing the most clustering. However, there is a limited ability to draw definitive conclusions from these findings due to several limitations. The findings are the limited by the quantity of observations, the lack of spatial granularity, and the spatial and population extent these observations are meant to represent. The limited quantity of observations leads to being unable to group observations by sex, one of the building blocks of the biological profile. Combine few observations with a lack of spatial granularity and there rises the need to aggregate the data, reducing the significance of the spatial autocorrelation calculation. Finally, the limited number of observations are not representative of the spatial and population extents from which they are drawn (see Figure 1 for visualization).

Despite these limitations, the findings are still useful as it involves the interpretation of a novel dataset and contributes to future research. With the addition of spatial granularity, additional insights may be gained about the clustering of cranial measurements in tribal and/or ethnic groups. These findings also must be reconciled with a principal component analysis being completed by Dr. Hefner. Overall, the results from this research are in-line with the expectations of a pilot study.

I have learned much from this project. My confidence in using R has increased overall but specially to tackle spatial analysis questions. However, I do think I will leave map making to a GUI-based GIS program such as QGIS. I gained a better understanding in the ways that humans at the individual and population level can be identified, which has many interesting implications. I also learned that not all fields use spatial statistics to answer spatial questions, which is hard to reconcile but useful, nonetheless. In the future, I will start researching spatial analysis methods as broadly applied when I find out a field doesn't normally use them in their methodologies.

Conclusion

Cranial measurements are used in the field of forensic anthropology for identifying the origin and movement among individuals and populations. In this pilot study, historical cranial measurements of skulls found in the collections of the Indian Museum in Calcutta were transcribed into spreadsheets and ingested into R-Studio. This tabular data was aggregated at the state-level and the spatial autocorrelation was calculated using a distance-band nearest neighbor formalization. Based on the results of a Monte Carlo Moran test, eight aggregated cranial measurements were found to have statistically significant, low to medium clustering. These findings provide novel insights into the clustering of cranial measurements in India in 1909 and form the basis for future research on the topic.

Data

The data and scripts for this final paper can be found at https://github.com/andrewroylaws/geog891/tree/main/Final_project.

References

Brunsdon, C. (2018). *An introduction to R for spatial analysis and mapping* (2nd edition). SAGE Publications.

Dunn, R. R., Spiros, M. C., Kamnikar, K. R., Plemons, A. M., & Hefner, J. T. (2020). Ancestry estimation in forensic anthropology: A review. *WIREs Forensic Science*, *2*(4). https://doi.org/10.1002/wfs2.1369

ESRI. (2021, March 21). *India State Bondaries 2020*. https://www.arcgis.com/home/item.html?id=6b15e6676364485a82f5cd36fa743c30

GADM. (2021). *GADM Maps and Data*. https://gadm.org/data.html

Harding, R. M. (1990). Modern European Cranial Variables and Blood Polymorphisms Show Comparable Spatial Patterns. *Human Biology*, *62*(6), 733–745.

Indian Museum. (1909). *Craniological Data from the Indian Museum, Calcutta*. Government Printing, Calcutta. http://www.archive.org/details/craniologicaldatOOindi

Maestri, R., Fornel, R., Gonçalves, G. L., Geise, L., Freitas, T. R. O., & Carnaval, A. C. (2016). Predictors of intraspecific morphological variability in a tropical hotspot: Comparing the influence of random and non-random factors. *Journal of Biogeography*, *43*(11), 2160–2172. https://doi.org/10.1111/jbi.12815

Ross, A. H., & Williams, S. E. (2021). Ancestry Studies in Forensic Anthropology: Back on the Frontier of Racism. *Biology*, *10*(7), 602. https://doi.org/10.3390/biology10070602

Sebastiano de Bona. (2021, November 12). *Function to create dataframe* [Personal communication].

Spradley, M. K. (2013). *Project IDENTIFICATION: Developing Accurate Identification Criteria for Hispanics* (Research No. 2008-DN-BX-K464; p. 69). US Department of Justice.

Spradley, M. K. (2014). TOWARD ESTIMATING GEOGRAPHIC ORIGIN OF MIGRANT REMAINS ALONG THE UNITED STATES-MEXICO BORDER: Origin of Migrant Remains Along the United States-Mexico Border. *Annals of Anthropological Practice*, *38*(1), 101–110. https://doi.org/10.1111/napa.12045

Spradley, M. K., Jantz, R. L., Robinson, A., & Peccerelli, F. (2008). Demographic Change and Forensic Identification: Problems in Metric Identification of Hispanic Skeletons. *Journal of Forensic Sciences*, *53*(1), 21–28. https://doi.org/10.1111/j.1556-4029.2007.00614.x