Spatial analysis of historical Indian cranial measurements

GEOG 897 Spatial Analysis with R

Fall 2021

Andrew Laws

**Introduction**

In the field of forensic science, 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. 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 scientists need not only accurate and uniform data collection but the establishment each measurements use case.

Background

**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 regions1 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**

Tabular data preprocessing led to a final dataset that included 138 observations with 18 variables. Spatial data preprocessing resulted in 37 total geometries. The final aggregated and joined sf dataframe *cran.sf* contained 12 observations (ISO codes) with 34 aggregated cranial measurements.

Results from the ESDA showed that the Andaman and Nicobar Islands (ISO code INAN) resulted in no statistically significant Moran’s I values from *moran.test*. Therefore, the related records were dropped from all data (see Figure 1 for this visualization). The state dispersal for each cranial measurement was normally distributed.

Discussion

Asia South Albers Equal Area Conic (ESRI 102028) was chosen as the CRS due to its limiting of distortion in the area of interest.

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.

Conclusion

References