Exploring Historical Links between Scotland and India using Geoparsing

Chandramauli Tyagi[*1] and Bruce M. Gittings [†2]

[1]Digitalisation, Ramboll UK
[2]School of Geosciences, The University of Edinburgh

GISRUK 2023

Summary

A significant amount of spatial information can be derived from unstructured datasets available in web pages, e-books, and digital archives. Geoparsing is one such concept that is very useful in extracting spatial data from any unstructured text source. Geoparsing complemented with Natural Language Processing algorithms can effectively automate this process of identifying and geo-tagging the extracted spatial data. The research illustrates the power of geoparsing by extracting place names from a corpus of biographies of famous Scots who travelled to India from the 18th to the early 20th Century to give an impression of the spread of the Scottish diaspora at that time.

KEYWORDS: Geoparsing, Natural Language Processing (NLP), Geo-tagging, Historical GIS (HGIS), Named Entity Recognition

1. Introduction

In this data-driven age, there has been a continuous expansion of digital resources via optically scanning historical texts, generating a pool of valuable information for the researchers. Applying Geoparsing and NLP algorithms in the historical context can often give new insights into the geographies embedded in various historical texts (Paterson et al., 2018; Gregory et al., 2015). This 'Mutualism' between GIS and humanities can offer a whole range of novel analysis techniques, which is uncommon in traditional humanities research. This research paper explores the quantitative and qualitative capabilities of Geographical Information Systems (GIS) and related algorithms/technologies to analyse and study a spatiotemporally significant link.

The main challenge to geoparsing in a historical context is the associated 'spatiotemporal ambiguity' of the resolved placenames. For this study a global gazetteer that could cover all the regions across Scotland and India was required. Though a global gazetteer gives greater coverage but is restricted by its feature density (Acheson, 2017). Thus, for an identical study area, a global gazetteer would be limited in performance compared to a regional gazetteer (higher feature density). Table 1 and 2 gives a comparative statistic on feature density (features per unit area) between GeoNames (Global Gazetteer) and Gazetteer for Scotland (Regional Gazetteer).

Table 1 Coverage statistics of the Gazetteer for Scotland (Gittings, 2021)

Entity	Number of features[‡]	Area in Sq. Km.	Feature Density (Feature/km^2)
Scotland	97565	78,789[§]	1.238

[*] chandramauli.tyagi@ramboll.co.uk | chandramauli.tyagi@gmail.com
[†] bruce@geo.ed.ac.uk
[‡] Include features like settlement/farm, geographical, tourist attraction, counties, and parishes.
[§] https://www.scotlandinfo.eu/scotland-facts-and-figures/

Table 2 Coverage statistics of GeoNames Gazetteer with respect to India and United Kingdom (GeoNames, 2021)

Entity	**Number of features**[**]	**Area in Sq. Km.**	**Feature Density (Feature/km^2)**
India	648,766	32,87,263[††]	0.197
United Kingdom	62,977	244,820	0.257

The study shall provide a case study of automatically extracting place names (via geoparsing) from the historical biographies of the famous Scottish personalities who travelled to India from the 18th Century to the mid-20th Century. The research aims of the study are therefore to:

- Investigate the utility of existing geoparsing algorithms in extracting place names from historical biographies.
- Qualitative analysis of the geoparsed places and investigate the occurrence of any regional Spatio-temporal patterns in Scotland and India.
- Enhancing the Gazetteer for Scotland by enriching its database of Scots associated with India from 1707-1947.

2. Methodology

The study is based on the archives of historical biographies available digitally on the web. An Open-source biographical corpus fulfilling was preferred for the research.

2.1. Data Source

'The biographical dictionary of eminent Scotsmen' originally compiled by Robert Chambers, is the principal source for this study. However, it does not cover the entire span of 240 years (1707-1947) needed for the study. The Dictionary of National Biography (DNB), published between 1885-1900, is available in 63 volumes, with additional supplements published later (DNB, 2021).

The Gazetteer for Scotland (G*f*S) was used to identify famous Scottish personalities within the DNB biographies. Since G*f*S is a fully georeferenced linked database, all the Scottish places in G*f*S's biographies are already place-resolved. Although a work-in-progress, the GfS was able to signpost the names of individuals whose biographies could be extracted from the DNB.

Table 3 The list of data repositories used for geoparsing

Data Source	**Description**	**Temporal Scale of Data**
Primary Data Source		
The Biographical dictionary of eminent Scotsmen, published in 1875, originally compiled by Robert Chambers.	Available at: https://digital.nls.uk/biographical-dictionary-of-eminent-scotsmen/archive/ under Creative Commons Attribution 4.0 International Licence	Biographical information available till 1875

[**] Includes features like administrative boundary, hydrographic, area, populated place, road/rail, spot, hypsographic, undersea, vegetation.

[††] https://www.india.gov.in/india-glance/profile

The Dictionary of National Biography (DNB), published between 1885-1900 (including supplements). The 1912 supplement volume being the last version of the DNB.	Available at: https://en.wikisource.org/wiki/Dictionary_of_National_Biography under [Creative Commons Attribution-ShareAlike License](#).	Biographical information till available 1912
The Biographies text of Famous Scots available in the Gazetteer for Scotland (G*f*S)	Available at: https://www.scottish-places.info/people.html[‡‡] or https://www.scottish-places.info/anyword.html	Biographical information for the entire study period (1707-1947)
Miscellaneous Data Source		
The Indian Biographical Dictionary, published in 1915, originally compiled by C. Hayavadana Rao	Available at: https://en.wikisource.org/wiki/The_Indian_Biographical_Dictionary_(1915) under [Creative Commons Attribution-ShareAlike License](#)	Biographical information available till 1915

2.2. Data Processing

Figure 1 gives an overview of the steps followed for processing the data, from finding the biographical text to building relationship tables.

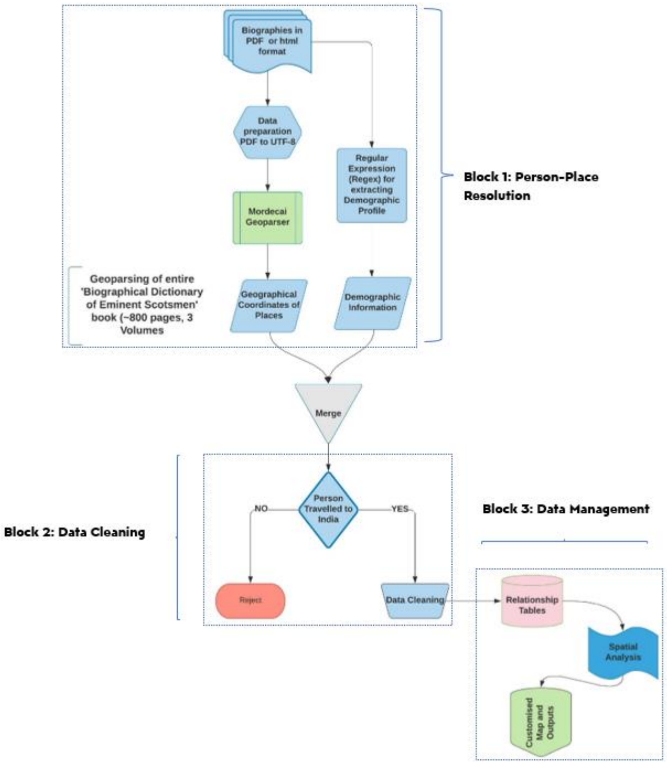

Biographies in PDF or html format
Data preparation PDF to UTF-8
Regular Expression (Regex) for extracting Demographic Profile
Mordecai Geoparser
Block 1: Person-Place Resolution
Geoparsing of entire 'Biographical Dictionary of Eminent Scotsmen' book (~800 pages, 3 Volumes
Geographical Coordinates of Places
Demographic Information
Merge
Block 2: Data Cleaning
NO
Person Travelled to India
YES
Block 3: Data Management
Reject
Data Cleaning
Relationship Tables
Spatial Analysis
Customised Map and Outputs

Figure 1 Data Processing Workflow

[‡‡] (c) The Gazetteer for Scotland, 2020. Used with permission from The Gazetteer for Scotland at https://www.scottish-places.info/

2.2.1. Block 1: Person-Place Resolution

The first step after collecting the biographies involves geoparsing the text for placename resolution. For this purpose, Mordecai [§§] (Halterman, 2017) geoparser was used. The biographical data is available either as PDF or TXT/HTML, while the input requirement for the geoparser is the string data type. Consequently, before geoparsing, all the biographies are converted into TXT (UTF-8). The Place resolution is carried out by extracting and geo-tagging the place names via geoparsing. The output of this step gives a list of place names attached with latitude and longitude. As a result of these two steps, places associated with every person are obtained.

2.2.2. Block 2: Data Cleaning

The geoparsing does not always resolve a placename correctly. There can be several factors responsible for this, but the primary being diversity of natural languages, placename ambiguity, and metonymic language (Gritta et al., 2020). The process of data cleaning is preceded by selecting the persons associated with India between 1707 to 1947.After compiling the list of Scottish personalities related to India, the places associated with every person are validated for data integrity and consistency. The table provides the list of errors commonly encountered during the process of data cleaning.

Table 4 Type of errors encountered during data processing

Type	Associated Risk	Confidence Level (provided by geoparser)	Description
Typo/Garbage words	Nil	Not Applicable	Created while converting OCR scanned PDFs to text. These can be easily removed on manual inspection.
Recognised but not Resolved	Low	Low to Medium	The place is recognised, but geoparser could not geo-resolve the place. The places with a low confidence level (less than 0.5) can be removed on inspection, while for medium-level confidence (0.5 to 0.75), the original biographical text is rechecked.
Neither Recognised nor Resolved	Medium	Not Applicable	It was noticed that the NLP algorithm sometimes entirely missed a few places. Such entries can only be picked up on physical scrutiny of the biographies. Relevant places/locations are then added manually to the table of geoparsed places.
Recognised but wrongly Resolved	High	High	Places are wrongly resolved (in context to the text). They pose a high risk to the overall performance, as they are usually resolved with a high confidence level and cannot be removed just by physical inspection.

[§§] https://github.com/openeventdata/mordecai

2.2.3. Block 3: Data management

The conceptualisation of ER Model is the next step after the data cleaning process (Figure 2).

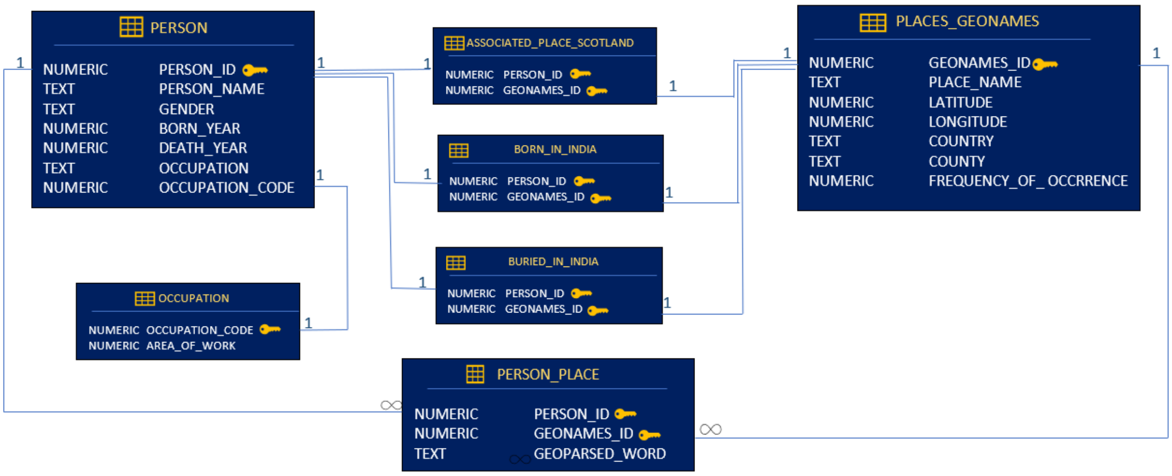


Figure 2 ER model for relationship tables

3. Results and Discussion

Figure 3 gives a visual overview of the locations geoparsed from the historical biographies of the famous Scots from 18th Century to early 20th Century. It is critical to mention that the world map shows places after filtering only those Scots (sample size of 160) that visited India, which explains aggregation of points over Scotland and India. The places visited in Scotland and India are shown separately in Figure 4.

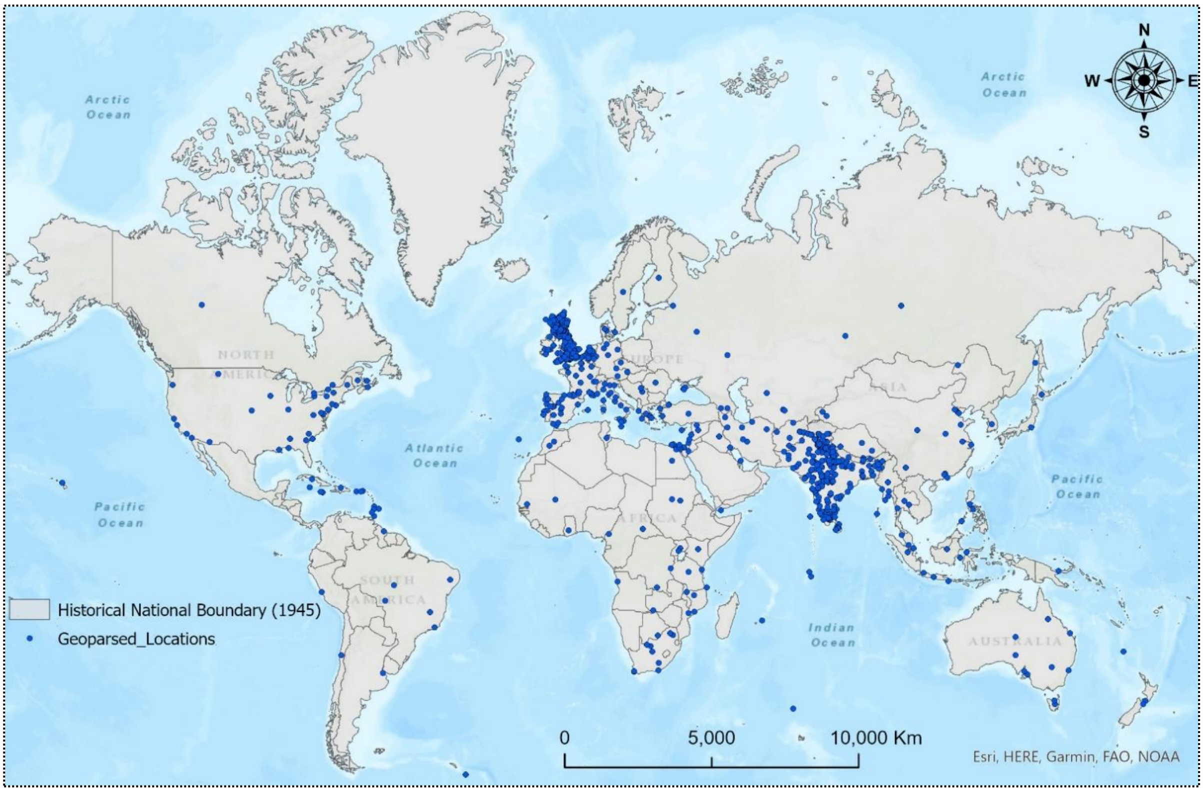
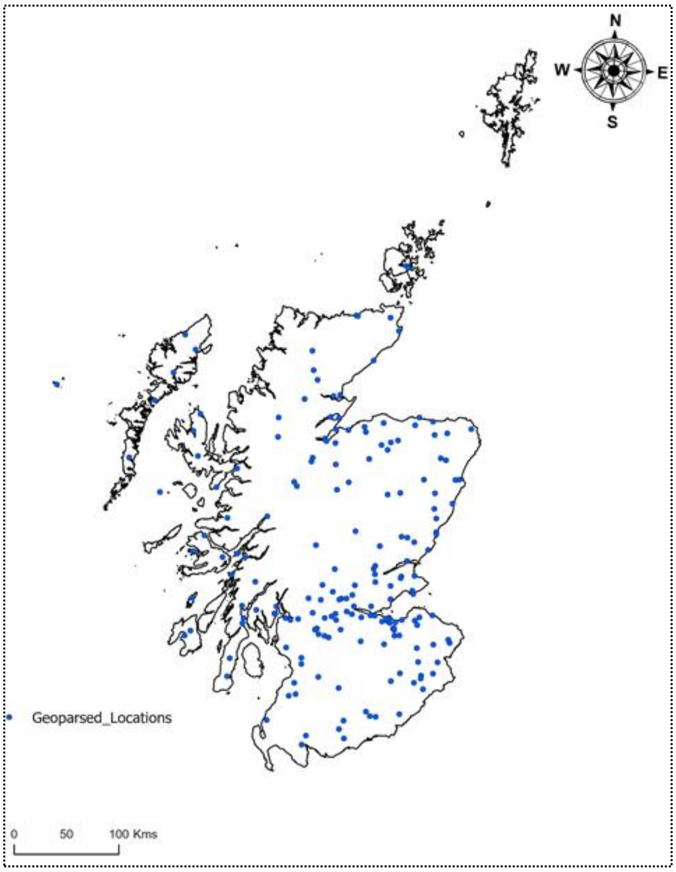


Figure 3 Geoparsed locations visited by the famous Scots from 1707 to 1947



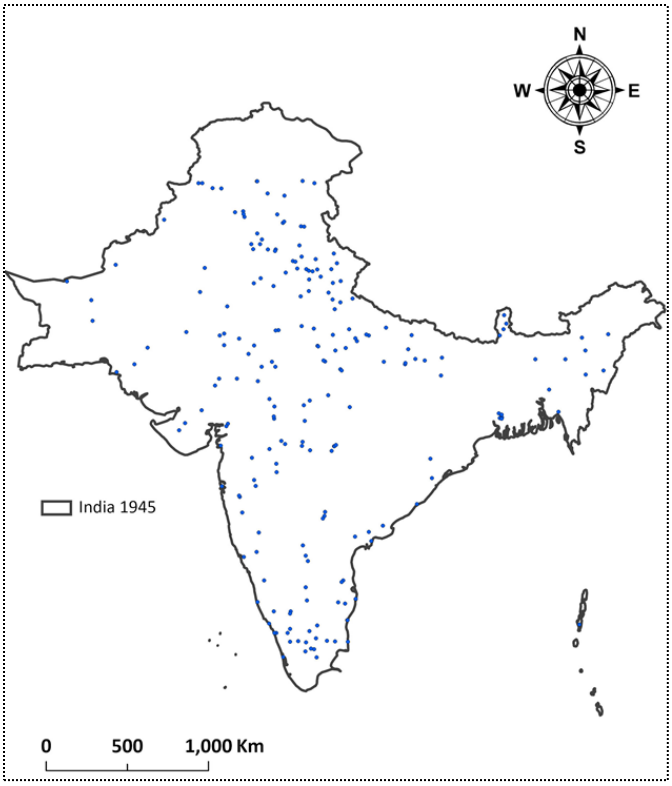


Figure 4 Geoparsed locations in Scotland*** and India††† visited by the famous Scots from 1707 to 1947

*** Scotland Country Boundary (Blackwood, 2017)

††† Historical National Boundary (University of Minnesota, 2021)

3.1. Scotland

For a better understanding of the socio-economic background of the Scots went to India between 1707 to 1947. Occupation profile for the sample size of 160 is plotted as a pie chart (Figure 5).

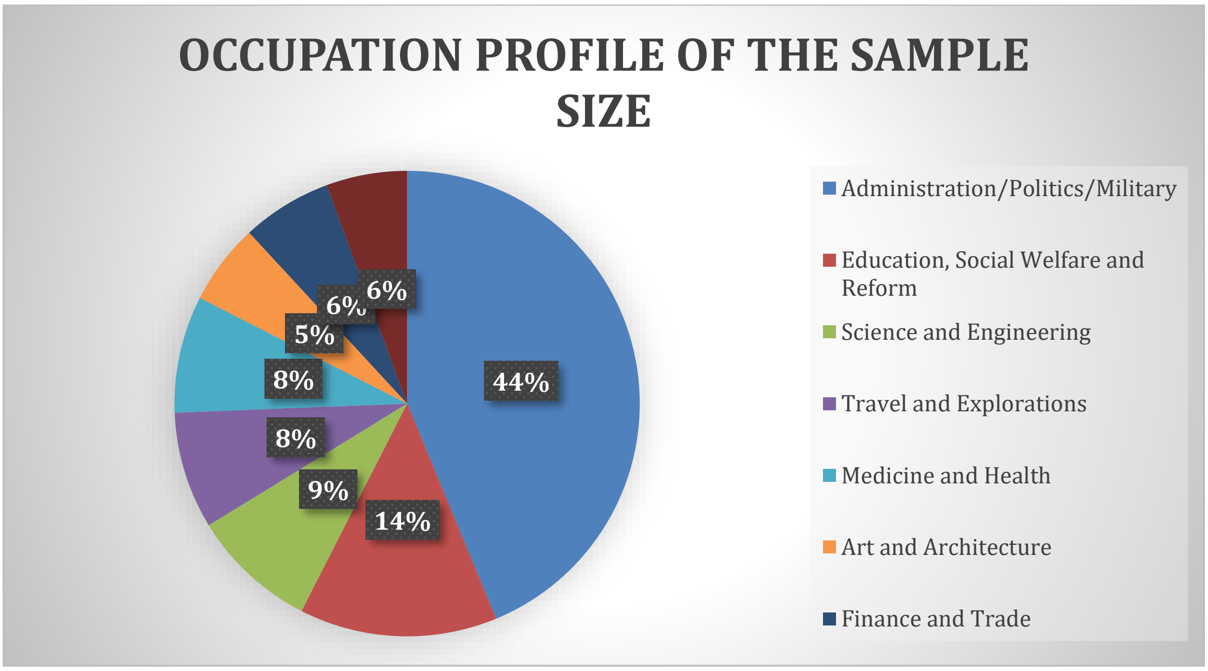


Figure 5 The Occupation Profile of the Scots travelled to India from 1707-1947

The regional variations across the counties with respect to people travelling to India is highlighted in Figure 6.

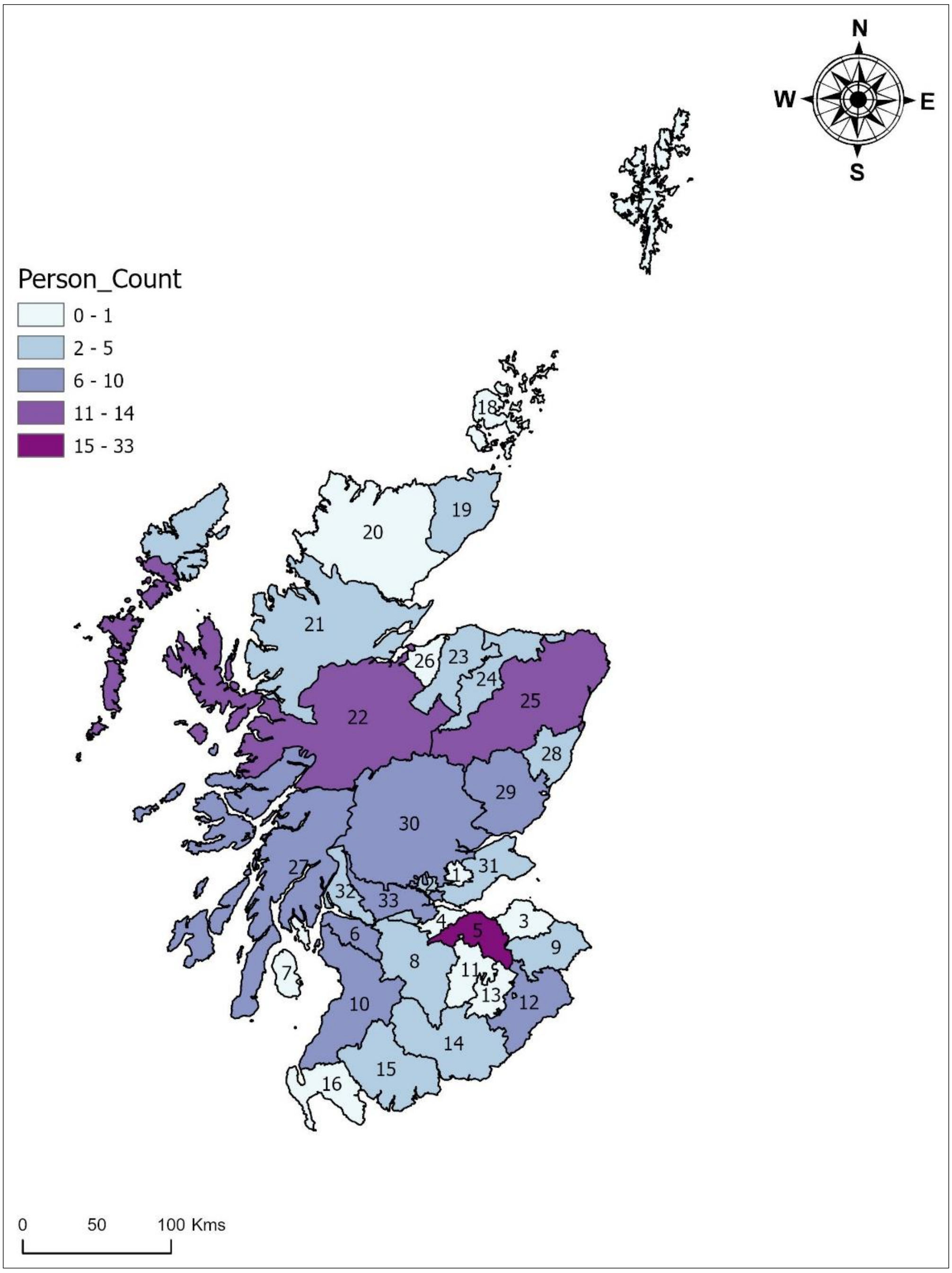


Figure 6. County wise map of the total number of people travelled to India. The Historical counties[‡‡‡] (pre-1890) are used for demarcating the boundary. The figure assumes an individual to be associated to a single county based on the place of birth or location of the estate/house. These assumptions are also necessary otherwise all the famous Scots may end up associated with capital city of Edinburgh.

[‡‡‡] Copyright (c) 2012 of the Great Britain Historical GIS Project and the University of Portsmouth available under Creative Commons Attribution-Non-commercial-ShareAlike 3.0

In contrast to Figure 6, Figure 7 highlights the regional variations across the counties as a function of person count density.

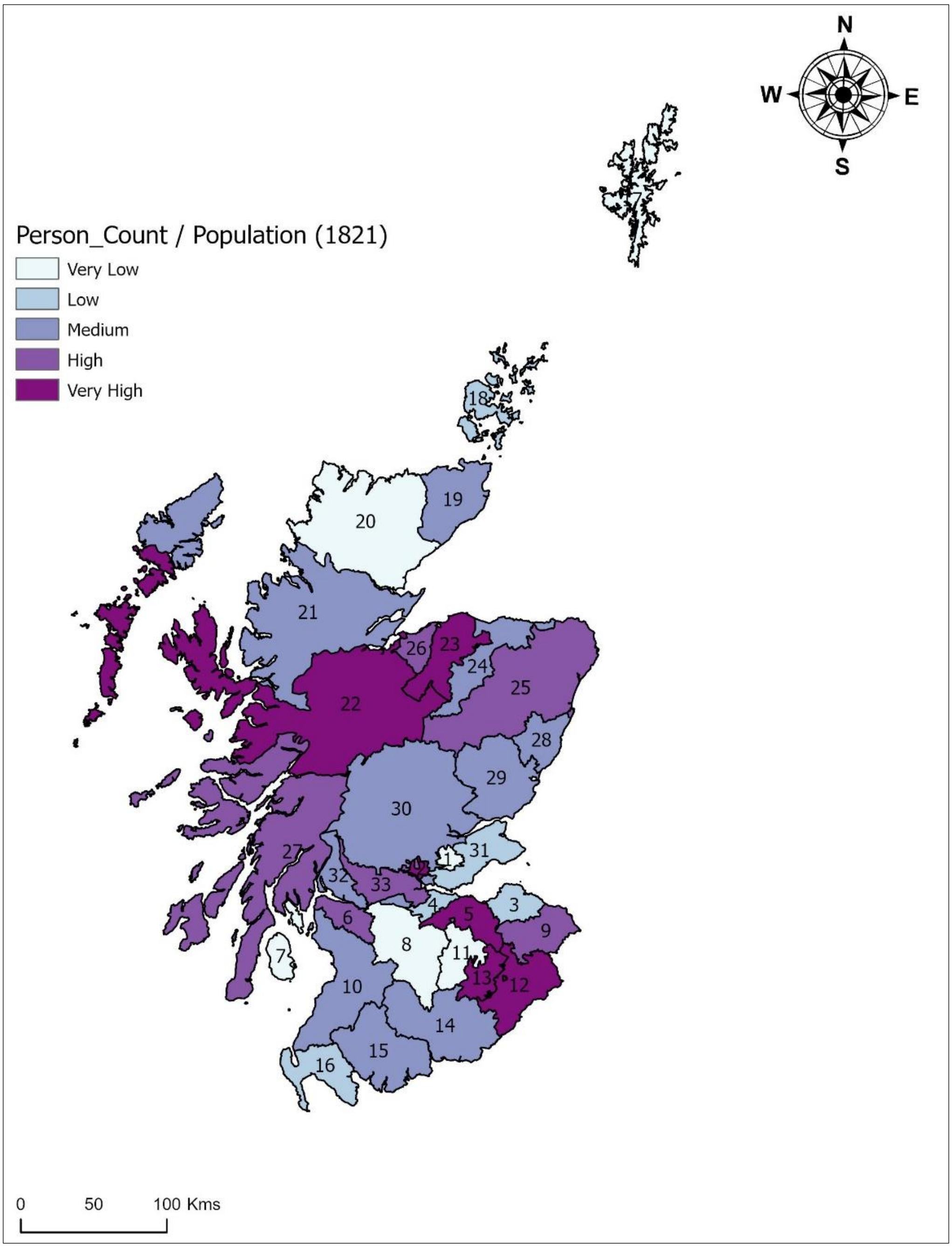


Figure 7. County wise map of the people travelled to India, corrected with population of each county. The Historical counties[§§§] (pre-1890) are used for demarcating the boundary. The 1821 Census is used as the reference as it lies on the mid-point of the timeline of this study.

[§§§] Copyright (c) 2012 of the Great Britain Historical GIS Project and the University of Portsmouth available under Creative Commons Attribution-Non-commercial-ShareAlike 3.0

3.2. India

To analyse the regional travelling patterns of Scots in India, the study period is divided into two phases: the early phase (1707-1864) and the later phase (1865-1947). The year of death is used as the reference, i.e., the early phase includes the individual who died on or before 1864, while the later phase includes people who died after 1864. For analysis, the historical boundary map (University of Minnesota, 2021) of the year 1800 (for early phase) and 1914 (for later phase) is used as the reference for representing the extent of India, where the country included present-day Pakistan and Bangladesh, together with other more minor boundary differences.

3.2.1. The Early Phase (1707-1864)

The heat map (Figure 8) identifies regions that were visited more compared to others. The hotspots around Mumbai, Kolkata, and North-Central Region (around Delhi and Lucknow) are noticeable. Further, there are two distinguishable hotspots in the southern part of India. The hotspot region over the northern part of India can be observed but is less prominent than other hotspots on the map. A few vacant areas in the map (South-East and West) are visible and may broadly correspond to forested regions in South-East India and the Thar Desert in Western India.

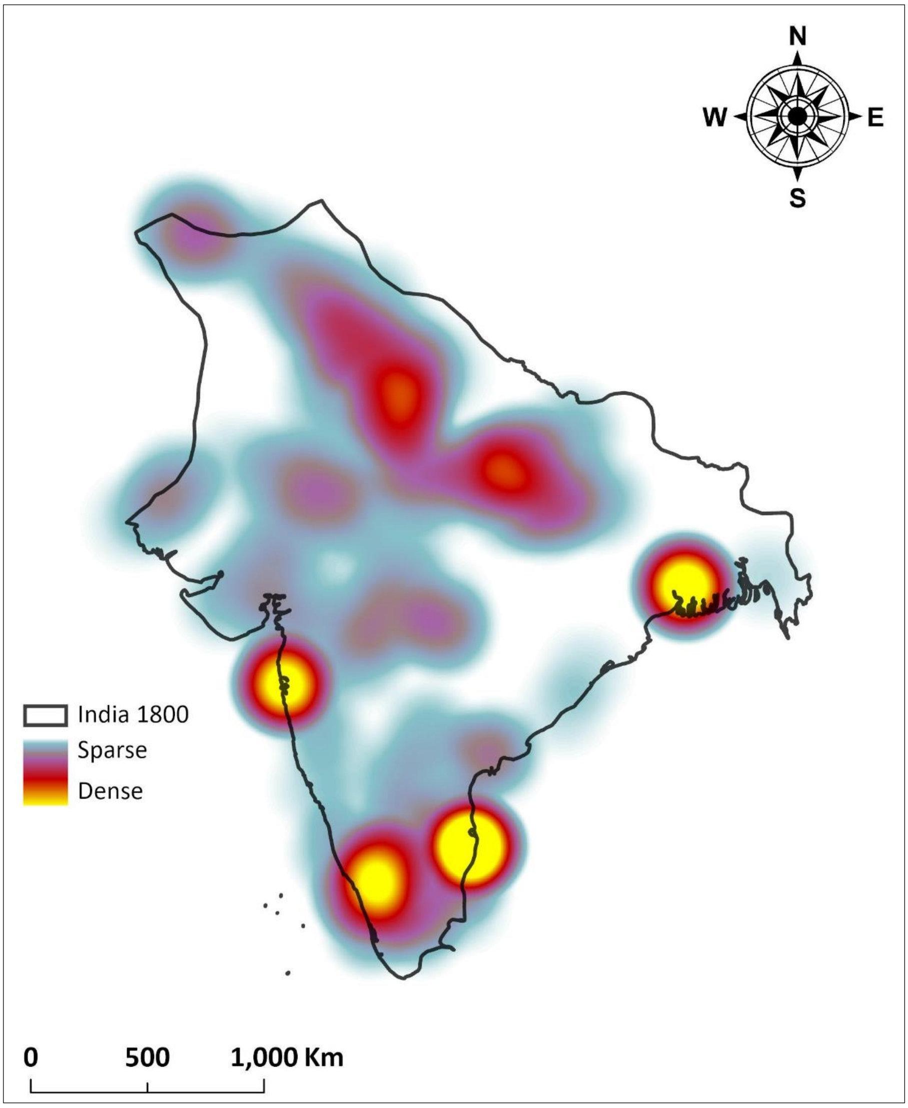


Figure 8 Regions most visited by Scots in India during 1707-1864

3.2.2. The Later Phase (1865-1947)

This phase represents the consolidation and expansion of the British Empire in the later years of British rule in India. From Figure 9, new emerging hotspots in the east and northwest part of India are noticeable. The hotspot in the northern region has become more prominent while the southern part has been reduced. There are fewer vacant areas compared to the early phase, implying the further expansion of British Empire in this phase.

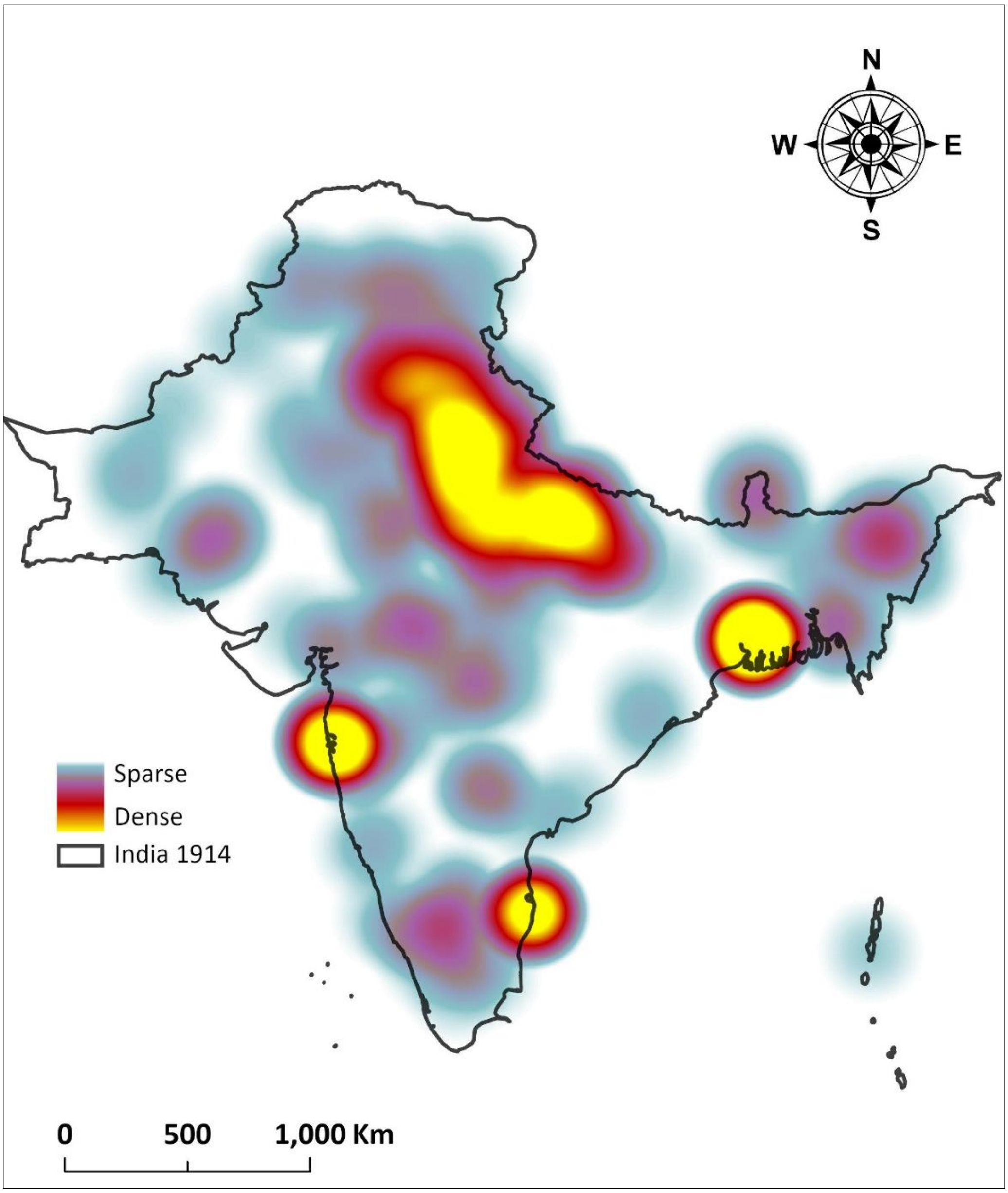


Figure 9 Regions most visited by Scots in India during 1865-1947

4. Findings and Limitations

This study assessed the utility of geoparsing algorithms for place resolution from historical biographies. The study demonstrates the limitations in extracting placenames due to the error associated with the geoparsing process (Table 4). The errors and associated ambiguity affect the ability to comprehensively extract and correctly geo-tag places. For this study, the GeoNames Gazetteer used by the geoparser provided fair coverage for places across India and Scotland (Figure 3 and 4), but a better gazetteer (for example the Gazetteer for Scotland) would undoubtedly produce better results. The Gazetteer for Scotland provides a feature density of 1.252, almost five times higher than that to the GeoNames Gazetteer (Table 1 and 2).

The analysis of Indian places travelled to by Scots provided three persistent hotspots over the port cities of Mumbai, Chennai, and Kolkata throughout the time frame of this study (1707-1947). Regional variations were noticed in the early phase (1707-1864) and later phase (1865-1947). The Southern region was more prominent in the early phase (Figure 8), while the Northern region became more prominent in the later phase (Figure 9).

The future studies in this research domain should consider deploying customised geoparser specifically tailored for historical texts, rather than using an 'all-'purpose' geoparser. This will unquestionably improve the efficiency and effectiveness of placename extraction and resolution process. The quality of OCR scanned documents should also be considered as it creates a few challenges in correctly extracting the information (Bazzo et al., 2020).

Another important finding this study notes is that historical biographies are not truly balanced in terms of representation of women; only 3 % of those persons included are women – a feature of the time when women were not considered equal to men. Therefore, future work should attempt to identify more women to provide a more balanced study.

5. Conclusion

The study demonstrates the capabilities of GIS (spatial analysis and visualisation) to aid and enrich historical studies. It is a novel attempt to investigate historical linkages between India and Scotland from a spatial perspective. The research has shown the utility of geoparsing algorithms in a historical context. These algorithms provide a very effective workflow for automating the process of extracting and geo-resolving places. The study defines an effective methodology for cleaning, processing, and data management of the information derived from historical biographies within its confined scope. The study analysed the spatio-temporal patterns across India and Scotland through various geoprocessing and visualisation tools. The study also contributed the information gathered during the research to the "Editors of the Gazetteer for Scotland" to enrich its database of Scots associated with India.

The use of customised geoparsers and regional gazetteers, like the Gazetteer for Scotland, will further enhance the place-resolution capabilities of the geoparser. The research brings out the potential of merging demographic characteristics with georeferenced information. The study highlights the usefulness of innovative visualisation techniques for analysing historical links. Future research should continue geoparsing this vast number of historical archives to seize the opportunity of combining human creativity and machine intelligence.

References

Acheson, E., De Sabbata, S., & Purves, R. S. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. Computers, Environment and Urban Systems, 64, 309–320. https://doi.org/10.1016/j.compenvurbsys.2017.03.007

Barrow, I. J. (2017). The East India Company, 1600-1858: A Short History with Documents. Cambridge, MA: Hackett Publishing Company.

Biographical Dictionary of Eminent Scotsmen (BDEM) (2021). In National Library of Scotland. Available from: https://digital.nls.uk/biographical-dictionary-of-eminent-scotsmen/archive/74458002

Blackwood, Carol. (2017). Scotland Country Boundary, [Dataset]. EDINA. https://doi.org/10.7488/ds/1759.

Bazzo G.T., Lorentz G.A., Suarez Vargas D., Moreira V.P. (2020) Assessing the Impact of OCR Errors in Information Retrieval. In: Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12036. Springer, Cham. https://doi.org/10.1007/978-3-030-45442-5_13

Dictionary of National Biography (DNB). (2021). In Wikisource. Available from: https://en.wikisource.org/w/index.php?title=Dictionary_of_National_Biography&oldid=6744388

GeoNames (2021): Available at: https://www.GeoNames.org/. Last Accessed 05th August 2021.
Gregory, I. N, Bennett, C., Gilham, V. L. and Southall, H., R. (2002). The Great Britain Historical GIS Project: From Maps to Changing Human Geography. Cartographic journal, 39(1), pp.37–49, doi: 10.1179/caj.2002.39.1.37. Available at: https://doi.org/10.1179/caj.2002.39.1.37

Gregory, I., Donaldson, C., Murrieta-Flores, P., Rayson, P. (2015). Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. International journal of humanities and arts computing, 9(1), pp.1–14, doi: 10.3366/ijhac.2015.0135.

Gritta, M., Pilehvar, M.T. & Collier, N. (2020). A pragmatic guide to geoparsing evaluation: Toponyms, Named Entity Recognition, and pragmatics. Language Resources and Evaluation, 54(3), 683–712. https://doi.org/10.1007/s10579-019-09475-3

Gittings, B. M. (2021). The Gazetteer for Scotland. https://www.scottish-places.info/

Halterman, A. (2017). Mordecai: Full Text Geoparsing and Event Geocoding, Journal of Open Source Software, 2(9), 91, doi:10.21105/joss.00091

Paterson, L. L., & Gregory, I. N. (2018). Geographical Information Systems and Textual Sources. In Representations of Poverty and Place (pp. 41–60). Springer International Publishing. https://doi.org/10.1007/978-3-319-93503-4_3

University of Minnesota (2021). Historical National Boundaries, [Dataset]. Available at: https://www.arcgis.com/home/item.html?id=6b836b2859194fdfa156458f2d2842e9 . Last Accessed 18th July 2021.

University of Portsmouth (2012). Great Britain Historical GIS Project, [Dataset]. Available at: https://www.visionofbritain.org.uk/. Last Accessed 26thJuly 2021.

Biographies

Chandramauli Tyagi

Chandramauli Tyagi is currently a Graduate GIS Consultant in the Digitalisation Team of Environment and Health Division at Ramboll UK. He completed his MSc in Earth Observation and Geoinformation Management from the University of Edinburgh in September 2021. In addition to his interest in Remote Sensing and GIS, he has years of experience in Meteorological Services.

Bruce M. Gittings

Bruce Gittings is a Senior Lecturer at the University of Edinburgh, having been a member of the academic staff there since 1986. He is particularly responsible for the Masters programmes in Geographical Information Science and Earth Observation and Geoinformation Management. His research interests relate to database management, web integration and gazetteers, especially the Gazetteer for Scotland. He serves as the Chair of the Association for Geographic Information in Scotland, with particular interests in education, training and skills to bring new blood into the GI profession.