Plotting film toponyms: A study in cultural geo-analytics

Andrea Ballatore*[1], Stefano De Sabbata[2] and Daniel Chavez Heras[1]

[1]Department of Digital Humanities, King's College London
[2]Department of Geography, University of Leicester

Spatial Humanities 2022

Abstract

Films are deeply geographical. Externally, they are produced in places, across increasingly complex and shifting global networks that connect organisations, cities, professionals, and equipment. Internally, their imagined geographies are set in either real or fictional places, and refer to their social, political, and cultural facets. In this study, we adopt a cultural analytics approach to commence an investigation of the spatial dimension of films, focussing on toponyms in film plots. Using geoparsing, we extract toponyms from about 42,000 film plots from Wikipedia and we analyse their spatial distribution by country. We then consider the relationship between a film's country of origin and the plot toponyms, charting the flows from places where films are produced to the geographies evoked in their stories.

KEYWORDS: film geography, cultural geo-analytics, geoparsing, imagined geographies, digital humanities

1 Introduction

As cultural objects, films possess distinct geographies (Hallam and Roberts, 2014). They are conceived, set, produced, distributed, and consumed in places (Anton, 2006). They depict places shaping their imaginary and, in turn, are influenced by the spatial context of their production (Reijnders, 2016). Geographies can be traced about the film industry and its transnational spatial networks (Shaw, 2013), and about the diegetic worlds in films. The imagined geographies in which cinematic stories unfold can be real, fictional, or hybrid, and are conjured up through a combination of real locations, sets, and computer-generated imagery. The choice of locations at the level of storytelling is influenced by many interlocking factors, such as production constraints and countries' tax regimes.

Attracted to its sheer size and global reach, data scientists have been exploring many facets of the film industry, for example investigating gender balance in films (Yang et al., 2020) and the visual patterns that connect genres and film directors (May and Shamir, 2019). Although marginally, the spatial humanities have also intersected with film studies: geographic information systems (GIS)

*andrea.ballatore@kcl.ac.uk

have been used to study the spatialities of film history (Klenotic, 2011) and of narrative structures (Caquard and Naud, 2019). In this short study, we tackle a fundamental research question: *Where are films set?* This deceptively simple question guides our first step to charting novel film geographies, using cultural analytics (Manovich, 2016) on these ubiquitous and influential cultural objects, looking at the interplay between the spatial ontology of film (i.e., where films are produced) and the film ontology of space (i.e., where films are set).

2 Finding place references in film

Film plot dataset. The CMU Movie Summary Corpus provides a collection of annotated film plot summaries and film metadata extracted from the English Wikipedia, from 1888 to 2013 (Bamman et al., 2013). These plot summaries are co-authored by non-expert editors, without following a specific protocol. An important limitation lies therefore in the heterogeneity of plot length and level of detail, as some Wikipedia editors may pay more attention to place names than others. The country of origin of the 42,306 films included in the dataset shows a strong presence of films produced in the US (42%), India (10%), United Kingdom (7%), and France (%5), with other countries representing each less than 3% of the total. Besides the linguistic preference for films in English, this distribution reflects the inclusion criteria of Wikipedia editors, who might include foreign films that were distributed in the English-speaking world and that reached some visibility.

Geoparsing film plots. To identify the toponyms mentioned in the plot summaries, we deployed geoparsing (Gregory et al., 2015). First, we extracted the 17,461 unique n-grams identified as locations by NER in the CMU Movie Summary Corpus. We then harnessed the Edinburgh Geoparser[1] (Grover et al., 2010) to geocode the n-grams using GeoNames[2] as a gazetteer, identifying 7,628 places. To improve the results, we complemented the geoparser's matches with a simple lexical approach using the same gazetteer, identifying 5,913 places, including 1,514 not identified by the first method. After a comparison of the two result sets, 42 n-grams geocoded as different places by the two approaches were manually resolved, 3 n-grams geocoded as different places by the two approaches were discarded as not relevant. The two most commonly occurring of the 7,175 unresolved n-grams ("LA" and "States") were manually resolved and the rest was discarded. The quality of the results was considered sufficient for this exploratory study.

Distribution of place references. Film plots contain a highly variable number of place references. 51% of films (20,821) exhibit at least one toponyms, leaving 49% of films without any explicit platial reference. It is important to note that this does not imply that these films are not explicitly set in a named place, but that the place of setting was not deemed salient enough by the Wikipedia editors. Among the geo-located films, about 30% have one toponym and 60% fall between 2 and 6, with a tail of high values (5% between 7 and 87 references). The distribution appears to follow a power law, with few films with many toponyms and vice-versa, observed in many natural and human systems. The most cited places are shown in Table 2, highlighting the disproportionate

[1] `https://www.ltg.ed.ac.uk/software/geoparser`, accessed in February 2022.
[2] `https://www.geonames.org`, accessed in February 2022.

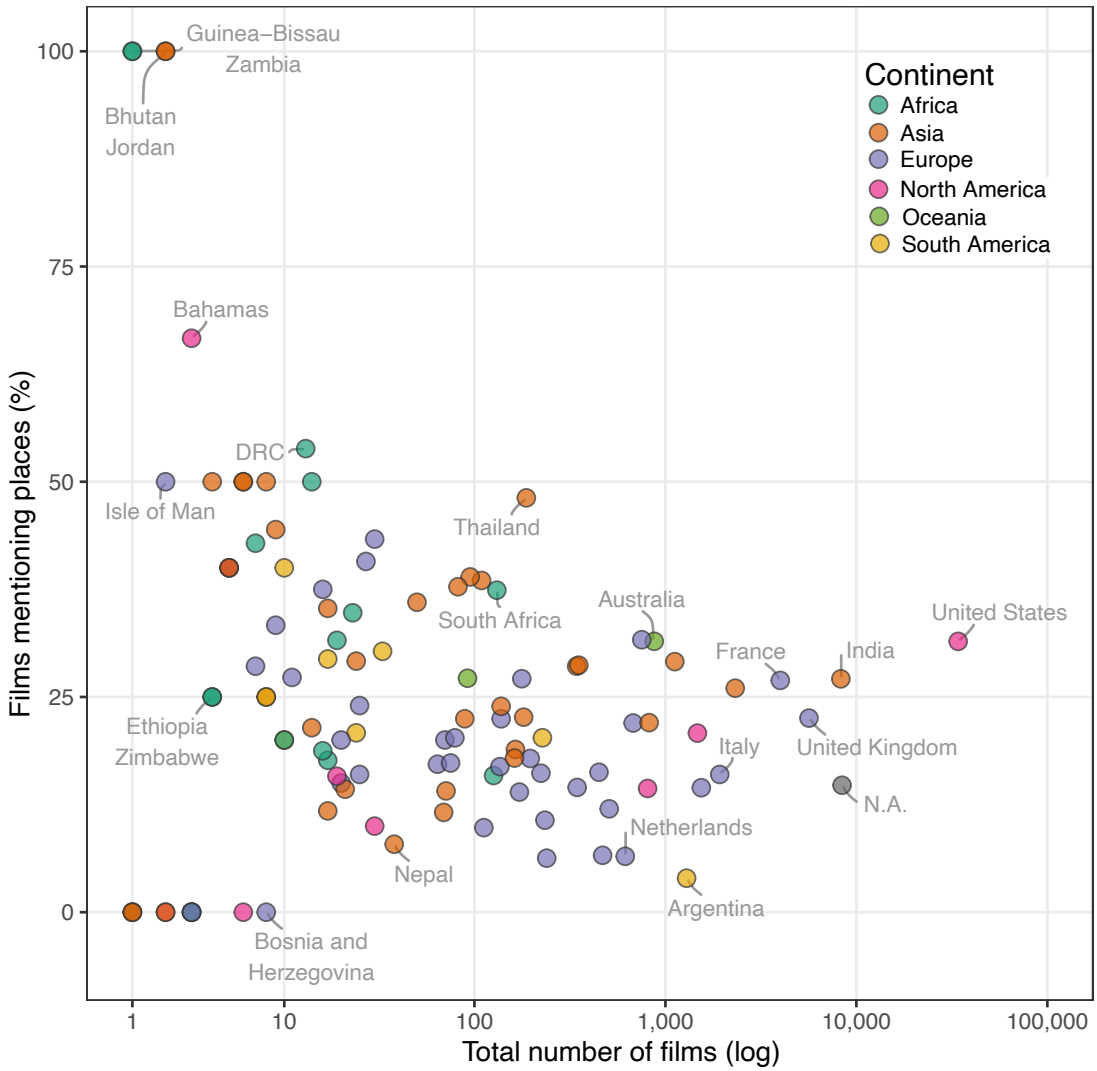

Figure 1: Percentage of films mentioning places in their plots grouped by country. The colours represent continents. Data source: 20,821 film plots from Wikipedia (1888–2013).

visibility of dominant countries and global cities such as London, New York and Paris.

To explain the spatial patterns in how films are grounded geographically, the country of origin emerges as particularly important. To provide a cross-sectional outlook, countries of origin not in existence are merged with the most similar current ones (e.g., Soviet Union and Russia). When grouping the films by this dimension, the proportion of plots that mention places varies widely (see Figure 1).

Toponym	No. films	Toponym	No. films
New York/N.Y.C.	1,767	France	447
United States	1,712	Japan	319
London	787	Germany	302
Earth	623	Europe	301
Paris	589	Chicago	279
Los Angeles	521	China	278
England	500	Mexico	271
America	481	Texas	258
California	481	Italy	257
India	474	Hollywood	251

Table 1: Most cited toponyms in 20,821 film plots from Wikipedia (1888–2013). The results include abbreviations and alternative spellings.

From country of origin to plot toponyms. At a large scale, it is reasonable to expect that films produced in a country tend to be set in and refer to proximal places. Figure 2 illustrates the flow from countries of origin to place references in film plots. For example, Italy has produced 134 films that reference toponyms in Italy, 352 in the U.S., and 108 in France. This does not imply that those films are set in the target country, as these counts include any place reference, including, for example, the birthplace of a character. These country-level intersections can be interpreted in light of clusters of countries that tend to co-produce films.[3] The data highlights the interconnected film geographies in Western Europe and in the Anglosphere. These flows can also be highly asymmetrical. Notably, 108 Italian films mention places in France, while only 12 French films mention places in Italy, perhaps reflecting the higher productive power of the latter.

3 Towards cultural geo-analytics

This study illustrated the possibilities enabled by the application of geoparsing to a large corpus of films. Because of the limitations and biases in Wikipedia, more work is needed to trace this film geography at a higher accuracy, harnessing data sources that reflect the variety and richness of global film production. While we observed countries, we acknowledge the need for a more nuanced approach that would account for the transnational nature of film production. Several research directions can be taken from this initial investigation, with a focus either on nomothetic trends or on idiographic accounts of specific geographic regions, periods, genres, and artists. We believe that a fruitful interplay between geographic data science and cultural analytics await to be explored for films and other cultural objects.

References

Anton, E. (2006). The Geography of Cinema — A Cinematic World. *Erdkunde*, 60(4):307–314.

Bamman, D., O'Connor, B., and Smith, N. A. (2013). Learning latent personas of film characters. In

[3] `https://stephenfollows.com/most-frequent-co-producing-nations`, accessed in February 2022.

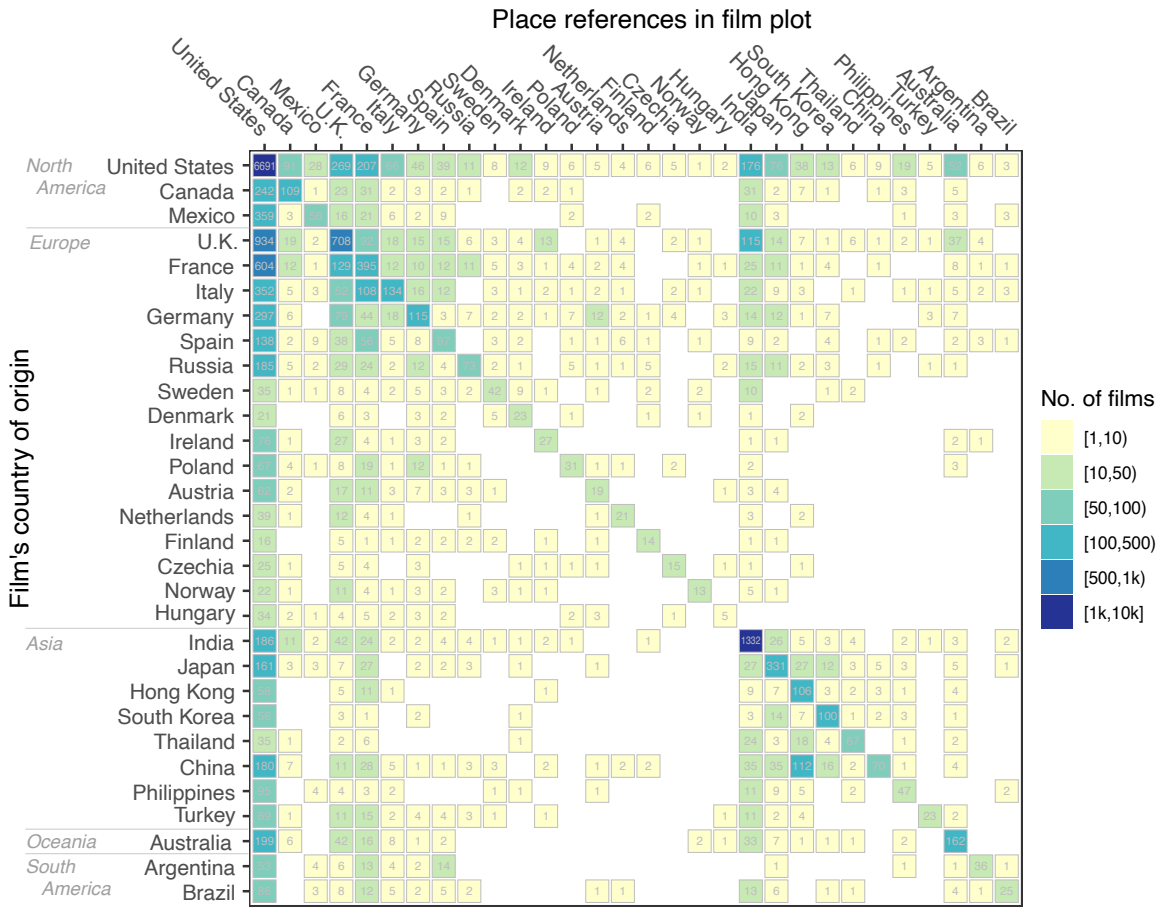

Figure 2: Flows at the country level from a film's country of origin to the country to which the plot refers to. The data was simplified by only considering only the main country of origin in the case of co-productions. If a film references toponyms in multiple countries, it is counted once in each cell. For example, 22 Italian films reference places in India. Data source: 20,821 film plots from Wikipedia (1888–2013).

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 352–361.

Caquard, S. and Naud, D. (2019). A spatial typology of cinematographic narratives. In Taylor, D. F., Anonby, E., and Murasugi, K., editors, *Further Developments in the Theory and Practice of Cybercartography*, volume 7 of *Modern Cartography Series*, pages 103–115. Elsevier, Oxford, UK.

Gregory, I., Donaldson, C., Murrieta-Flores, P., and Rayson, P. (2015). Geoparsing, GIS, and textual analysis: Current developments in spatial humanities research. *International Journal of Humanities and Arts Computing*, 9(1):1–14.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., and Ball, J. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1925):3875–3889.

Hallam, J. and Roberts, L., editors (2014). *Locating the moving image: New approaches to film and place*. Indiana University Press, Bloomington, IN.

Klenotic (2011). Putting Cinema History on the Map: Using GIS to Explore the Spatiality of Cinema. In Maltby, R., Biltereyst, D., and Meers, P., editors, *Explorations in new cinema history: Approaches and case studies*, pages 58–84. John Wiley & Sons, Oxford, UK.

Manovich, L. (2016). The science of culture? Social computing, digital humanities and cultural analytics. *Journal of Cultural Analytics*, 1(1):11060.

May, C. and Shamir, L. (2019). A data science approach to movies and film director analysis. *First Monday*, 24(6).

Reijnders, S. (2016). *Places of the imagination: Media, tourism, culture*. Routledge, London.

Shaw, D. (2013). Deconstructing and Reconstructing 'Transnational Cinema'. In Dennison, S., editor, *Contemporary Hispanic Cinema: Interrogating the Transnational in Spanish and Latin American Film*, pages 47–66. Tamesis, Woodbridge, UK.

Yang, L., Xu, Z., and Luo, J. (2020). Measuring Female Representation and Impact in Films over Time. *ACM Transactions on Data Science*, 1(4):1–14.