AGILE: GIScience Series, 3, 9, 2022. https://doi.org/10.5194/agile-giss-3-9-2022
Proceedings of the 25th AGILE Conference on Geographic Information Science, 2022.
Editors: E. Parseliunas, A. Mansourian, P. Partsinevelos, and J. Suziedelyte-Visockiene.
This contribution underwent peer review based on a full paper submission.

Geoparsing: Solved or Biased? An Evaluation of Geographic Biases in Geoparsing

Zilong Liu [1,2], Krzysztof Janowicz[1,2,3], Ling Cai [1,2], Rui Zhu [1,2], Gengchen Mai [1,2,4], and Meilin Shi [1,2]

[1]STKO Lab, Department of Geography, University of California, Santa Barbara, USA
[2]Center for Spatial Studies, University of California, Santa Barbara, USA
[3]Department of Geography and Regional Research, University of Vienna, AT
[4]Department of Computer Science, Stanford University, USA

Correspondence: Zilong Liu (zilongliu@ucsb.edu)

Abstract.

Geoparsing, the task of extracting toponyms from texts and associating them with geographic locations, has witnessed remarkable progress over the past years. However, despite its intrinsically *geospatial* nature, existing evaluations tend to focus on overall performance while paying little attention to its variation across geographic space. In this work, we attempt to answer the question whether *geoparsing is solved or biased* by conducting a spatially-explicit evaluation, namely an evaluation of the regional variability in geoparsing performance. Particularly, we will analyze the spatial autocorrelation underlying this regional variability. By performing hot and cold spot detection over results of several open-source geoparsers, we observe that none of them performs equally well across geographic space, and some are geographically biased towards some regions but against others. We also carry out a comparative experiment showing that state-of-the-art geoparsers developed with neural networks do not necessarily outperform the off-the-shelf tools across geographic space. To understand the implications behind this observed regional variability, we evaluate geographic biases involved in geoparsing research centered around data contribution and usage, algorithm design, and performance evaluations. Particularly, our spatially-explicit performance evaluation serves as an approach to evaluation bias mitigation in geoparsing. We conclude that previous performance evaluations published in the literature are overly optimistic, thus hiding the fact that geoparsing is far from solved, and geoparsers require debiasing in addition to further considerations when being applied to (geospatial) downstream tasks.

Keywords. geoparsing, spatially-explicit evaluation, regional variability, geographic bias, evaluation bias mitigation

1 Introduction

Geoparsing is a key part of geographic information retrieval (Jones and Purves, 2008). Two consecutive steps constitute the pipeline of geoparsing systems (i.e., geoparsers), namely toponym recognition and toponym resolution. Toponym recognition is often treated as a subtask of Named Entity Recognition (NER), or more precisely speaking, Named Entity Recogntion and Classification (NERC) (Nadeau and Sekine, 2007). It refers to the process of identifying toponyms from texts in various forms of literature, such as news and Wikipedia articles. Each recognized toponym is then fed into a toponym resolution model that selects the correct place reference along with its geo-location information from all candidates with the same place name. Toponym resolution is also referred to as place name disambiguation (Overell and Rüger, 2008; Ju et al., 2016) in the related literature.

In recent years, neural network architectures have contributed to the improvement in the quality of geoparsers. Examples in toponym recognition include Wang et al. (2020) that used an improved version of the bidirectional Long Short-Term Memory (LSTM) neural network with a Conditional Random Field (CRF) layer (BiLSTM-CRF) (Lample et al., 2016), and Hu et al. (2021) that used C-LSTM (Zhou et al., 2015) combining a Convolutional Neural Network (CNN) with LSTM. Similarly, Gritta et al. (2018a) and Kulkarni et al. (2021) applied CNN while Fize et al. (2021) used LSTM in toponym resolution. As the use of deep learning techniques has greatly improved geoparsing performance, Wang and Hu (2019b) recently asked whether the geoparsing performance obtained with state-of-the-art geoparsers is good enough to essentially call the

problem *solved*. The authors argued that geoparsing can be claimed to be solved when it comes to prominent place names in well-formatted texts because their evaluation shows that these geoparsers, particularly deep-learning-based ones, can reach an outstanding level of toponym recognition performance and also relatively low errors in toponym resolution on multiple benchmark datasets.

However, there is a loophole in how the performance evaluation is carried out. Current evaluations only measure the overall performance of a geoparser on a corpus, thereby ignoring the fact that, as a *geospatial* task geoparsing also needs to be evaluated from a geospatial perspective. Put differently, how geoparsers perform across geographic space is not evaluated, thus making such evaluations susceptible to biases, e.g., uneven geographic coverage. Recent work has shown spatial heterogeneity in the performance of a deep-learning-based toponym resolution model developed by Fize et al. (2021), which fails to work in most of the southern and western regions in the US except several major cities, and similar poor performance is also obtained in small regions in the southwest of France and the north of Japan. Our work tries to expand the scale of geoparsing performance evaluation to both toponym recognition and resolution.

In addition, we aim at looking beyond spatial heterogeneity and study spatial autocorrelation (Griffith, 1987; Legendre, 1993; Getis, 2008) underlying regional variability in geoparsing performance. There are two reasons why we choose spatial autocorrelation as the focus of our performance evaluation. First, spatial autocorrelation is inherent in data with a spatial structure (Sokal and Oden, 1978; Koenig, 1999; Getis, 2007), and geoparsing is affected by such autocorrelation effects in place names that also exhibit distance decay patterns in their collective similarity (Hu and Janowicz, 2018). Second, we consider that an analysis of second-order spatial variations in geoparsing performance will shed light on where a geoparser is biased towards or against, which we argue is necessary for a spatially-explicit evaluation on geoparsing performance. The term *spatially-explicit* has been frequently used in ecological studies (Dunning Jr et al., 1995; Irwin and Geoghegan, 2001; DeAngelis and Yurek, 2017), and has been recently regarded as an vital characteristic that a GeoAI (Janowicz et al., 2020) should demonstrate by satisfying at least one of four tests, including the invariance test, representation test, formulation test, and outcome test (Goodchild, 2001). Examples of such spatially-explicit machine learning models include geo-aware image classification (Yan et al., 2018) and multi-scale spatial representation learning (Mai et al., 2020). Similarly, we consider an evaluation to be spatially-explicit if it fulfills any of the four tests above.

If we observe strong spatial effects in geoparsing performance, this would be an important indicator of geographic biases in existing systems and their evaluations. Geographic biases cause disparities in the geographic distributions between sampled and ground-truth data (Reddy and Dávalos, 2003; Yang et al., 2013; Syfert et al., 2013), and they can also lead to quality issues in data contribution to volunteered geographic information through crowdsourcing (Basiri et al., 2019; Janowicz et al., 2016). Such biases have inspired an interest in sampling bias mitigation (Syfert et al., 2013; Beck et al., 2014) and representativeness assessment (Zhang and Zhu, 2018). Meanwhile, they have also drawn attention from the machine learning community as machine learning researchers have faced with the same lack of geographic diversity in open datasets, such as ImageNet (Russakovsky et al., 2015) and Open Images (Krasin et al., 2017), which results in biasing their image classifiers towards Europe and North America (Shankar et al., 2017). In addition, geographic biases also affect spatial data aggregation, which can result in reduced reliability of multivariate statistical analysis (Fotheringham and Wong, 1991) and perturbations in feature embeddings that destabilize neural networks used in scenarios such as deep-learning-based traffic predictive modeling (Zeng et al., 2020). These potential consequences of geographic biases motivate us to investigate geographic bias issues that have not been discussed in geoparsing by studying the datasets and beyond, i.e., algorithms and performance evaluations. Put differently, we examine whether the claim that geoparsing is essentially solved is true, or whether the data, models, and the usage of evaluation metrics are simply biased.

Our research contributions are as follows:

- Rather than focusing on overall geoparsing performance, we conduct a spatially-explicit evaluation on how geoparsers perform across geographic space. We unveil spatial autocorrelation underlying regional variability in geoparsing performance, and analyze its comparison between deep-learning-based models and off-the-shelf tools in terms of toponym recognition and toponym resolution, respectively.
- We analyze and summarize representation biases, aggregation biases, algorithmic biases, and evaluation biases in geoparsing, along with recent work that attempts to mitigate them. Particularly, our spatially-explicit performance evaluation serves as an approach to evaluation bias mitigation. To the best of our knowledge, our work is the first to provide such geographic bias evaluation in the field of geoparsing (evaluation).

The remainder of this paper is organized as follows. Section 2 provides an overview of related work on spatially-explicit performance evaluations and different kinds of geoparsing evaluation studies. Section 3 introduces our spatially-explicit evaluation of geoparsing performance. Section 4 describes an exploratory analysis on normalized frequency distributions of geoparsing performance indicators, and the results about our performance evaluation. Section 5 discusses geographic biases involved in geoparsing research, and how recent work attempts to mitigate

these issues. Finally, we summarize our work and propose future directions in Section 6. Details about the reproducibility of our study can be found in Section 7.

2 Related Work

2.1 Spatially-Explicit Performance Evaluations

There are two potential avenues for a performance evaluation to become spatially-explicit. First, a spatially-explicit performance evaluation might use evaluation metrics where spatial information (e.g., distance) is incorporated. For instance, Xu and Zhang (2013) conducted a sensitivity analysis on land suitability evaluation (LSE), in which the Earth Mover's Distance is applied to identify spatial variations between the original map and the simulated LSE map. Second, there is usually a geospatial perspective from which a spatially-explicit performance evaluation is carried out. Examples include a "policyscape analysis" for biodiversity conservation (Barton et al., 2013), an evaluation framework for integrated carbon sequestration and biodiversity conservation (Forsius et al., 2021), and a land use conflict evaluation approach (Cui et al., 2021).

According to a recent review of evaluation metrics used in geoparsing (Wang and Hu, 2019a), toponym recognition performance can be evaluated with metrics such as *Precision*, *Recall*, and *F-Score*. *Precision* measures the percentage of correctly-recognized toponyms among all recognized toponyms, and *Recall* measures the percentage of correctly-recognized toponyms among all annotated toponyms. *F-Score* is the harmonic mean of *Precision* and *Recall*. No spatial information is involved in the calculation of these metrics, thus hindering the spatial explicitness of performance evaluations of toponym recognition. On the other hand, evaluation metrics are mostly distance-based for toponym resolution. Commonly-used metrics include *Mean Error Distance (MED)*, *Median Error Distance (MdnED)*, *Accuracy@161*, and *Area Under the Curve (AUC)*. These metrics evaluate toponym resolution performance by comparing the error distance determined by how far a resolved location is from its corresponding annotated location. *MED*, *MdnED*, and *AUC* calculate the mean, median, and overall deviation of error distances, respectively. *Accuracy@161* is used to calculate the percentage of correctly-resolved locations among all annotated locations. It considers a distance threshold of 161 kilometers, within which a resolved location is regarded as correct. However, when using these metrics to evaluate toponym resolution, the performance across geographic space was often not considered.

2.2 Geoparsing Evaluation

Geoparsing evaluation is an important part of geoparsing research, which is concerned with concrete metrics, progress reviews, reproducibility issues, and benchmark dataset construction in geoparsing.

Gritta et al. (2020) discussed standard metrics used to evaluate geoparsing performance, and provided an evaluation framework. For example, they divided toponym resolution evaluation metrics into coordinate-based, set-based, and ranking-based metrics. Also, the authors highlighted that spatial scopes of geoparsing, i.e., whether a geoparser is applied to a local or global coverage, should be taken into considerations in performance evaluations.

Wang and Hu (2019a) moved one step forward by not only conducting a more comprehensive review of evaluation corpora, state-of-the-art models, and evaluation metrics. The authors built an Extensible and Unified Platform for Evaluating Geoparsers (EUPEG)[1], which is a benchmark platform aimed at improving reproducibility in geoparsing research for comparative experiments. Both geoparsers and evaluation corpora hosted on this platform are open-source. These geoparsers use different toponym recognition techniques, such as general NER tools (e.g., Stanford NER[2]) and in-house NER tools (e.g., LT-TTT2 (Grover, 2008)). In the meantime, various toponym resolution models were adopted, such as heuristic rule-based models (e.g., CLAVIN[3]), geostatistical models (e.g., TopoCluster (DeLozier et al., 2015)), and deep-learning-based models (e.g., CamCoder (Gritta et al., 2018a)). In addition, there are multiple categories of evaluation corpora available, including news articles (e.g., TR-News (Kamalloo and Rafiei, 2018)), Wikipedia articles (e.g., WikToR (Gritta et al., 2018b)), social media posts (e.g., GeoCorpora (Wallgrün et al., 2018)), and web pages (e.g., Hu2014 (Hu et al., 2014)).

Along with eight annotated benchmark datasets and eight evaluation metrics on top of EUPEG, Wang and Hu (2019b) carried out a performance evaluation on nine state-of-the-art geoparsers, including those already hosted on the platform and others developed by top-ranked teams in a geoparsing competition called SemEval-2019 Task 12. They discussed the circumstance under which geoparsing can be considered as solved, and introduced three future directions in geoparsing. These directions include population-free toponym resolution, fine-grained geoparsing, and the usage of additional gazetteers in toponym resolution. In this paper, we will showcase that their performance evaluation would benefit from being more spatially-explicit, because such spatially-explicit evaluations will allow us to compare geoparsing performance among different locations.

[1] https://geoai.geog.buffalo.edu/EUPEG/
[2] https://nlp.stanford.edu/software/CRF-NER.html
[3] https://github.com/Novetta/CLAVIN

More recently, Laparra and Bethard (2020) combined Wikipedia[4] and OpenStreetMap[5] to construct a new kind of benchmark dataset for geoparsing compositions of place mentions into geographic regions. The authors also proposed accompanying evaluation metrics that can be used to compare predicted geometries with ground-truth geometries in either a strict or relaxed way. As our intention is to reproduce studies on geoparsing individual toponyms rather than complex geographic descriptions, we do not use their evaluation framework in our study.

3 A Spatially-Explicit Geoparsing Performance Evaluation

3.1 Geoparser Selection

Previous work has studied toponym recognition and resolution individually, and therefore, we evaluate each of them separately. The criterion of our model selection is that their relevant resources, including evaluation corpora (and training corpora, if applied), are available online so that they can be easily reproduced as baseline models for geoparsing research. For both toponym recognition and resolution, we use a deep-learning-based model that has achieved state-of-the-art results and an off-the-shelf tool to analyze whether the former necessarily outperforms the latter across geographic space.

Toponym Recognition Models For toponym recognition, we choose a pre-trained version of NeuroTPR[6] and the named entity recognition module of spaCy[7] (version 2.1) coupled with a large-sized trained English pipeline[8]. NeuroTPR is a BiLSTM-conditional random field toponym recognition model that deals with social media messages (Wang et al., 2020), and spaCy is an open-source Python library for natural language processing.

Toponym Resolution Models For toponym resolution, we compare a pre-trained version of CamCoder[9], a CNN-based toponym resolution model that integrates both lexical and geographic knowledge (Gritta et al., 2018a), with the rule-based Edinburgh Geoparser[10] (Grover et al., 2010) (version 1.2).

[4] https://www.wikipedia.org/
[5] https://www.openstreetmap.org/
[6] https://github.com/geoai-lab/NeuroTPR
[7] https://spacy.io
[8] https://spacy.io/models/en#en_core_web_lg
[9] https://github.com/milangritta/Geocoding-with-Map-Vector
[10] https://www.ltg.ed.ac.uk/software/geoparser/

3.2 Evaluation Corpus Selection

Toponym Recognition Corpus To evaluate toponym recognition, we select GeoCorpora (Wallgrün et al., 2018), a social media corpus containing geo-annotated tweets along with place mentions. This dataset was also used as one of the two evaluation corpora in Wang et al. (2020) and is available on Github[11]. Among all the 2,122 tweets, only those toponyms that have annotated coordinates are covered in our evaluation.

Toponym Resolution Corpora For toponym resolution evaluation, we select LGL (Lieberman et al., 2010), GeoVirus (Gritta et al., 2018a), and WikToR (Gritta et al., 2018b), which were also used in the evaluation of Gritta et al. (2018a) and shared along with CamCoder[9]. Both LGL and GeoVirus are news corpora. LGL is also the most frequently-used geoparsing evaluation dataset, consisting of 588 news articles from 78 local newspapers. GeoVirus contains 229 news articles from WikiNews, and they are centered around global disease outbreaks and epidemics. WikToR is a Wikipedia dataset containing 5,000 articles within which annotated toponyms are widely distributed across the world. The reason why we choose the three corpora here is that we want to analyze how CamCoder performs across geographic space on benchmark datasets with different levels of ambiguity. According to Gritta et al. (2018a), WikToR has higher place name ambiguity than GeoVirus, and GeoVirus has higher ambiguity than LGL.

Kulkarni et al. (2021) found that WikToR contains wrong coordinates for some places because the sign of either their latitude or longitude is flipped, and in all three corpora the same toponym will have slightly different coordinates because they were created differently. To correct these inconsistencies, we follow their method to unify three corpora by using the shared data patches[12]. Annotated toponyms without location information in the data patches are not included in our study.

3.3 Evaluation Metrics

In our performance evaluation, we select *Recall* and the coordinated-based *Median Error Distance (MdnED)* as they are frequently used to evaluate the performance of toponym recognition and toponym resolution, respectively. We apply them to measuring geoparsing performance at each annotated location individually, so that we can disclose how geoparsers perform across geographic space.

Toponym Recognition Evaluation Metric For toponym recognition, we use *Recall* to measure the proportion of the number of times an annotated location being

[11] https://github.com/geovista/GeoCorpora
[12] https://github.com/google-research-datasets/mlg_evaldata

correctly identified by a model among the number of times it is annotated (in a corpus). In Equation 1, $Recall_i$ is the $Recall$ of the i^{th} annotated location in a corpus; tp_i is the number of times it is recognized; and fn_i is the number of times it fails to be recognized. The range of $Recall$ is $[0, 1]$. A higher $Recall_i$ indicates a better toponym recognition performance with respect to the i^{th} annotated location. We adopt exact matching (Gritta et al., 2018b), meaning that only toponyms that match exactly with their ground-truth annotations are considered valid.

$$Recall_i = \frac{tp_i}{tp_i + fn_i} \tag{1}$$

Toponym Resolution Evaluation Metric For toponym resolution, *MdnED* is calculated as the median of error distances from the location of an annotated toponym to its resolved location (in a corpus). Compared with the commonly-used *Mean Error Distance (MED)* that calculates the mean, *MdnED* is better at dealing with outliers that exist in computed error distances, and therefore, *MdnED* can minimize the distortion of evaluation results. In Equation 2, $MdnED_i$ is the *MdnED* of the i^{th} annotated location in a corpus; ed_{ij} is its j^{th} error distance computed; $\mathbf{x}_i = (x_i, y_i)$ is its annotated location; n_i is the number of its resolved locations; $\mathbf{x}_{ij} = (x_{ij}, y_{ij})$ is its j^{th} resolved location; and $Dist(\cdot, \cdot)$ is the error distance between a pair of geographic coordinates, computed as the great circle distance in our experiment. A larger *MdnED* indicates worse toponym resolution performance with respect to an annotated location. Only the annotated locations that are recognized by a geoparser and can be found in the GeoNames gazetteer[13] are considered in the computation of *MdnED*. A shared version of GeoNames by Gritta (2018) is used to reproduce CamCoder, while the online version of GeoNames is directly accessed by Edinburgh Geoparser during place name disambiguation.

$$MdnED_i = Median(\{ed_{ij} | ed_{ij} = Dist(\mathbf{x}_i, \mathbf{x}_{ij}),\ \forall j \in [1, n_i]\}) \tag{2}$$

3.4 Spatial Autocorrelation Detection

To evaluate spatial autocorrelation effects in geoparsing performance, we apply the Getis-Ord Gi* statistic (Ord and Getis, 1995; Getis and Ord, 2010) that can help visually reveal hot spots and cold spots in geoparsing performance. In Equation 3, G_i^* is the Getis-Ord Gi* statistic of the i^{th} annotated location (in a corpus); v_i and v_j denote the geoparsing performance indicators (i.e., *Recall* or *MdnED*) of the i^{th} and the j^{th} annotated locations, respectively; $\bar{V}$ is the average geoparsing performance indicator of all annotated toponyms and N is the number of all annotated locations; w_{ij} is the spatial weight between the i^{th} and the j^{th} locations. We assign 1 to w_{ij} if the j^{th} annotated location is in the neighborhood of the i^{th} annotated location, and 0 otherwise. The Getis-Ord Gi* statistic is a z-score. A larger positive (or negative) z-score indicates a stronger clustering effect of high (or low) values, which represents hot (or cold) spots. We select the K-nearest neighbors of an annotated location as its neighborhood when calculating the Getis-Ord Gi* statistics. K is defined as 8 in the experiment.

$$G_i^* = \frac{\sum_{j=1}^{N} w_{ij} v_j - \bar{V} \sum_{j=1}^{N} w_{ij}}{\sqrt{\frac{\sum_{j=1}^{N} v_j^2}{N} - (\bar{V})^2} \sqrt{\frac{N \sum_{j=1}^{N} w_{ij}^2 - (\sum_{j=1}^{N} w_{ij})^2}{N-1}}} \tag{3}$$

While focusing on how geoparsers perform across space in general, we are also interested in whether regional variability in toponym resolution performance exists for highly ambiguous toponyms as toponym resolution is sensitive to ambiguity determined by the frequency of places with the same name. Therefore, we calculate the standard deviations of $MdnED$ produced by CamCoder and Edinburgh Geoparser, respectively, for highly ambiguous place names in WikToR since it has the highest ambiguity among all three evaluation corpora for toponym resolution in our experiment. The standard deviations are reported in Section 4.3.

4 Evaluation Results

4.1 The Normalized Frequency Distributions of Geoparsing Performance Indicators

First, we provide an exploratory analysis showing that the performance indicator distributions of both toponym recognition and toponym resolution are highly skewed. Figure 1(a) and Figure 1(b) describe the normalized frequency distributions of $Recall$ for spaCy and NeuroTPR, respectively. We can see there is a peak indicating that more than 50% annotated locations have a $Recall$ ranging from 0 to 0.1 for both toponym recognition models. Figure 2(a), Figure 2(b), and Figure 2(c) show the normalized frequency distributions of *MdnED* for Edinburgh Geoparser with respect to LGL, GeoVirus, and WikToR, respectively, while Figure 2(d), Figure 2(e), and Figure 2(f) show those distributions for CamCoder. On the contrary to $Recall$, $MdnED$ is expected to be as small as possible, but in most cases it can reach up to 16,000 kilometers.

4.2 Spatial Autocorrelation in Toponym Recognition Performance

Figure 3(a) shows that both hot spots and cold spots are detected in toponym recognition performance of spaCy

[13] https://www.geonames.org/

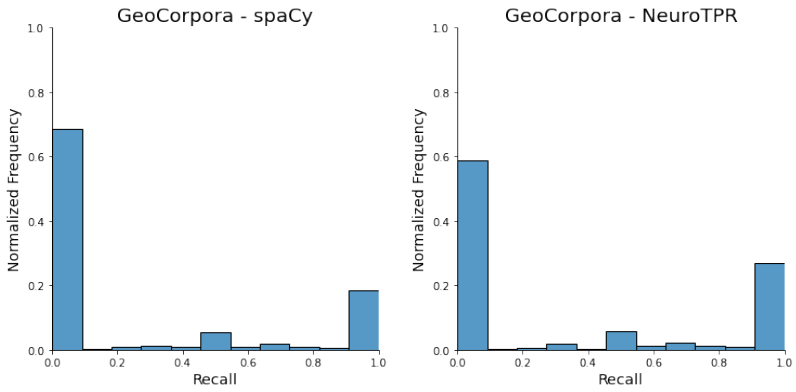

Figure 1. The normalized frequency distributions of *Recall* with respect to all annotated locations in GeoCorpora for spaCy and NeuroTPR, respectively

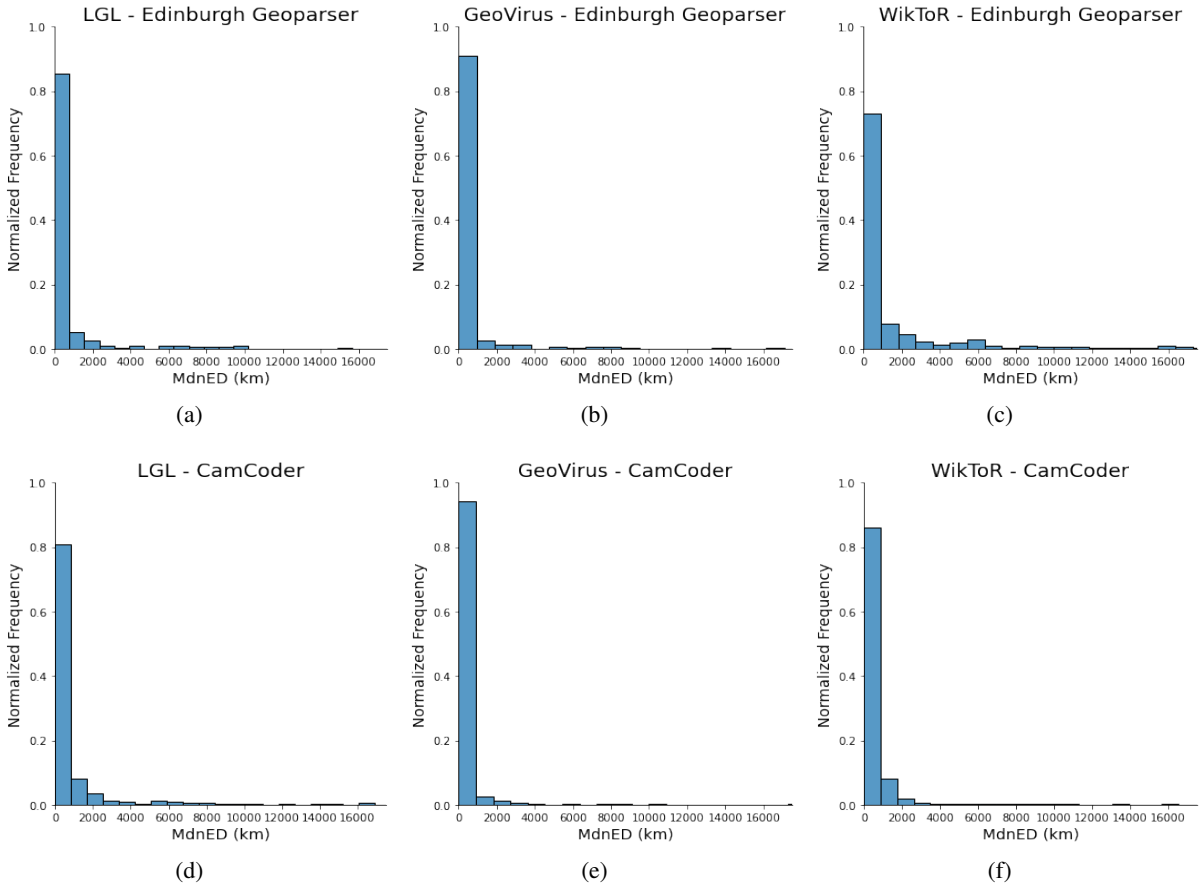

Figure 2. The normalized frequency distributions of *MdnED (km)* with respect to all annotated locations in LGL, GeoVirus, and WikToR for Edinburgh Geoparser and CamCoder, respectively

on GeoCorpora. For spaCy, most of all the hot spots are found in the United States, the United Kingdoms, and Malaysia on a global scale. Cold spots can be seen not only in several South American countries near the equator, Middle East countries (e.g., Bahrain and Qatar), South Asia countries (e.g., India), and so forth, but also in the United States and the United Kingdoms, where hot spots have been found as well. Figure 3(b) shows hot spots and cold spots in toponym recognition performance of NeuroTPR on GeoCorpora, where there is a larger proportion of more cold spots with 95% confidence compared with Figure 3(a). Again, we can see coexistence of both hot spots and cold spots in the United States and the United Kingdoms.

In general, the observed regional variability in toponym recognition performance shows that toponym recognition is geographically biased towards some regions and against others on both global and local scales.

4.3 Spatial Autocorrelation in Toponym Resolution Performance

In terms of toponym resolution, however, only cold spots can be found in the performance of both Edinburgh Geoparser and CamCoder on LGL, GeoVirus, and WikToR. Figure 4(a), Figure 4(c), and Figure 4(e) show that the geographic distribution of cold spots covers more regions on a global scale as the place name ambiguity of the corpus increases from LGL, GeoVirus to WikToR. Cold spots are mostly observed in the United States for LGL, in the United States, Mexico, the United Kingdoms, Australia, and Fiji for GeoVirus, and in the United States, Canada, the Philippines, Australia, and New Zealand for WikToR. Similar observations can be found for CamCoder's toponym resolution performance on LGL, GeoVirus, and WikToR, which are shown in Figure 4(b), Figure 4(d), and Figure 4(f), respectively.

In addition, CamCoder produces more cold spots (with higher confidence) and these cold spots are widely distributed across the world. By comparing the toponym resolution performance between Edinburgh Geoparser and CamCoder, we can see while the overall performance of CamCoder is higher than Edinburgh Geoparser as reported in previous research (Gritta et al., 2018a; Wang and Hu, 2019b), there are more regions (e.g., South America and Africa) where it fails to perform well. Note that the total of cold spots produced by CamCoder is greater than that of Edinburgh Geoparser because these two geoparsers use different toponym recognition models. Put aside this difference, the wider geographic distribution of cold spots in toponym resolution performance of CamCoder indicates a stronger geographic bias in this deep-learning-based model in comparison with the rule-based Edinburgh Geoparser.

In general, the observed regional variability in toponym resolution performance shows that toponym resolution is even worse than toponym recognition because there is no hot spots found. Put differently, toponym resolution is simply biased against many regions, both globally and locally.

The Standard Deviations of *MdnED* for Highly Ambiguous Toponyms in WikToR Then we examine the top ten most ambiguous place names in WikToR, and analyze regional variability in their corresponding toponym resolution performance according to the standard deviation of $MdnED$. We find that Edinburgh Geoparser fails to resolve half of the selected toponyms, which is because of a failure either to identify them during toponym recognition or to retrieve their coordinates from the gazetteer. For toponyms resolved only by CamCoder, we observe very large standard deviations of $MdnED$ for Springfield and Georgetown, which are 2,770.63 and 2,742.79 kilometers, respectively. For toponyms (including Washington County, Greenville, Kingston, Hamilton, and Newport) resolved by both models, Edinburgh Geoparser results in greater standard deviations of $MdnED$ than CamCoder. Particularly, this difference is extremely large for Hamilton and Newport, respectively. As all these standard deviations (except the one for Newport resolved by CamCoder) are greater than 161 kilometers, which is regarded as a threshold within which a resolved location can be considered correct relative to its annotated location (Wang and Hu, 2019a), both models exhibit very strong regional variability in toponym resolution performance for place names with high ambiguity.

Table 1. The standard deviations (SD) of $MdnED$ for highly ambiguous toponyms in WikToR, respectively

Toponym	Model	MdnED SD (km)
Washington County	Edinburgh Geoparser	807.70
	CamCoder	692.18
Clinton	Edinburgh Geoparser	/
	CamCoder	539.43
Greenville	Edinburgh Geoparser	407.52
	CamCoder	514.68
Springfield	Edinburgh Geoparser	/
	CamCoder	2770.63
Georgetown	Edinburgh Geoparser	/
	CamCoder	2742.79
Kingston	Edinburgh Geoparser	680.48
	CamCoder	226.45
Franklin County	Edinburgh Geoparser	/
	CamCoder	414.99
Hamilton	Edinburgh Geoparser	4105.57
	CamCoder	622.24
Jefferson County	Edinburgh Geoparser	/
	CamCoder	384.38
Newport	Edinburgh Geoparser	5705.19
	CamCoder	145.05

(a) (b)

Figure 3. Hot spots and cold spots in toponym recognition performance of spaCy and NeuroTPR, respectively, on GeoCorpora

5 Geographic Biases in Datasets, Algorithms and Performance Evaluations

While we have shown that geoparsing is geographically *biased* instead of being *solved*, we are also interested in the implications behind the observed autocorrelation effects in its performance, which is likely to be attributed to geographic biases in datasets, algorithms, and beyond that are involved in geoparsing. In this section, we analyze these biases, and discuss their potential influence on geoparsing. While following bias categorization of the machine learning community (Mehrabi et al., 2021), we also hope to highlight their geospatial characteristics and to point out that they have been (un)intentionally introduced in geoparsing.

5.1 Representation Bias

The first kind of geographic bias is the representation bias in training/evaluation corpora. Here we provide quantitative measurements of how representation biases in datasets commonly used in recent geoparsing research are across geographic space. Two metrics *Spatial Misalignment (SM)* and *Spatial Diversity Misalignment (SDM)* are used to measure the difference in place coverage and the difference in the geographic diversity of place coverage, respectively, between a corpus and the GeoNames Gazetteer. These metrics are introduced in Quattrone et al. (2015) to compare the content mapped by power users, i.e., users that represent only a small portion of the entire OpenStreetMap community but produce most of the content, and the content mapped by the crowd. In our experiment, each country/region is divided into $10km \times 10km$ grids, and each grid cell is assigned to the number of places within it (in a corpus or in the gazetteer). The country/region data used is a 1:10m shapefile from Natural Earth 5.0.0[14].

Equation 4 describes the calculation of *SM*, in which the i^{th} element in the vector $\overrightarrow{g_c}$ (or $\overrightarrow{g_d}$) is the number of places in the i^{th} grid of a country/region mapped by a corpus (or a gazetteer).

$$SM = 1 - \frac{\overrightarrow{g_c} \cdot \overrightarrow{g_d}}{||\overrightarrow{g_c}|| \cdot ||\overrightarrow{g_d}||} \tag{4}$$

Equation 5 describes the calculation of *Spatial Diversity Misalignment (SDM)*, which is an extension of the Shannon's Diversity Index (Shannon, 1948). The variable $g_{c,i}$ (or $g_{d,i}$) is the i^{th} element in $\overrightarrow{g_c}$ (or $\overrightarrow{g_d}$). As many regions of the world are sparsely populated with annotated locations in LGL and GeoVirus, we only provide measurements for global-scale evaluation corpora used in our spatially-explicit performance evaluation, i.e., GeoCorpora and WikToR. We also provide the measurements for a training corpus named GeoWiki shared by Gritta (2018). This is a version of the English Wikipedia dump that contains 1.4M Wikipedia articles, and was used in the training process of CamCoder without overapping with WikToR. The English Wikipedia dump has commonly served as an easily-accessible and frequently-updated large corpus for the training purpose of many other toponym resolution models and toponym recognition models as well.

$$SDM = \frac{(-\sum_{g_{c,i} \in \overrightarrow{g_c}} g_{c,i} \ln g_{c,i}) - (-\sum_{g_{d,i} \in \overrightarrow{g_d}} g_{d,i} \ln g_{d,i})}{max(-\sum_{g_{c,i} \in \overrightarrow{g_c}} g_{c,i} \ln g_{c,i}, -\sum_{g_{d,i} \in \overrightarrow{g_d}} g_{d,i} \ln g_{d,i})} \tag{5}$$

Table 2 and Table 3 show *SM* and *SDM* for GeoWiki, GeoCorpora, and WikToR, respectively. For all three corpora the median of *SM* is greater than or equal to 0.49, meaning there is a strong misalignment between their place coverage and those of GeoNames. This spatial misalignment is

[14] https://www.naturalearthdata.com/

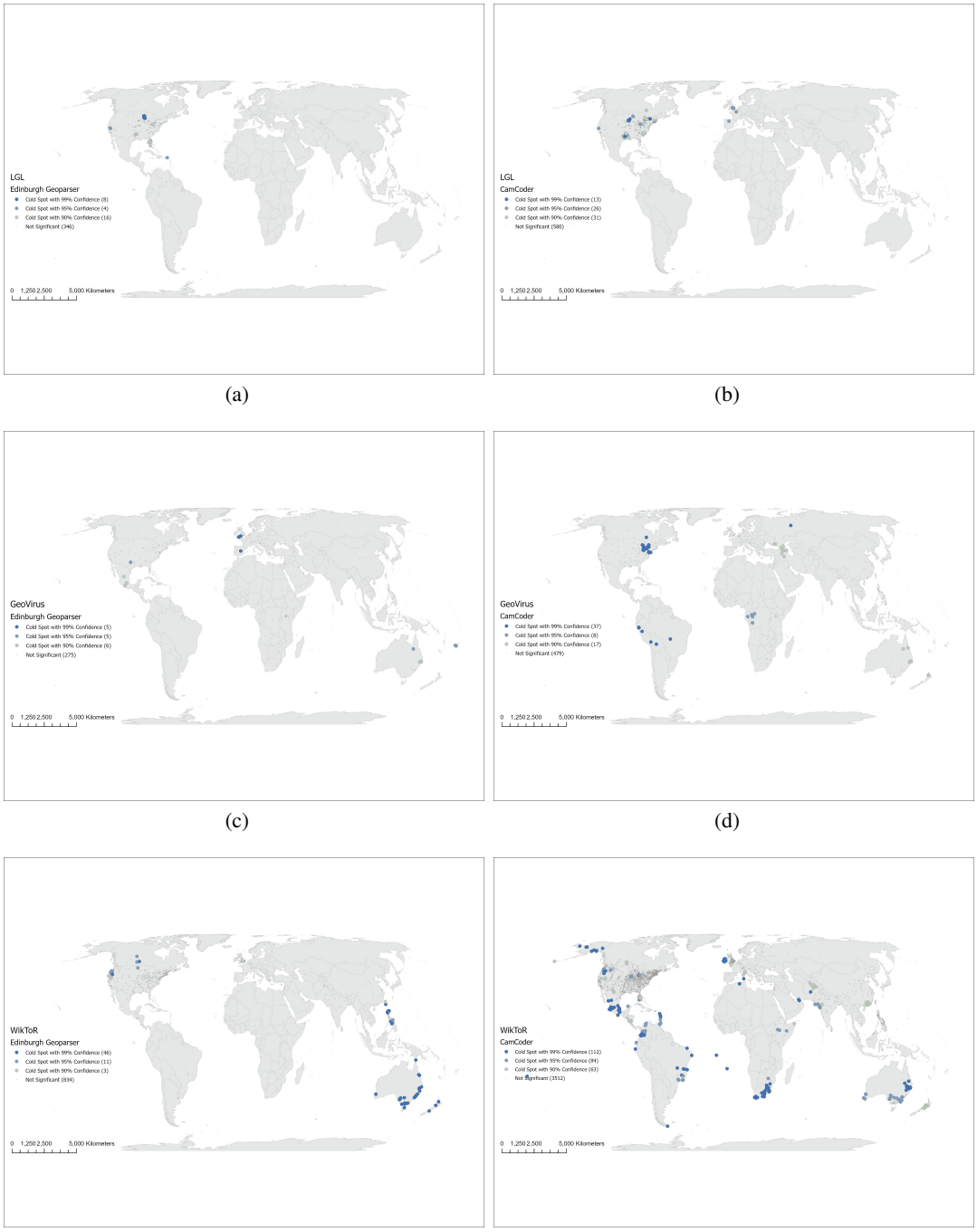

Figure 4. Cold spots in toponym resolution performance of Edinburgh Geoparser and CamCoder on LGL, GeoVirus, and WikToR, respectively

more prominent for evaluation corpora as both their first quantiles of *SM* are greater than 0.8. In addition, *SDM* for all three corpora is generally negative. This points out a strong misaligned geographic diversity in their place coverage, particularly for the social media corpus GeoCorpora with a *SDM* median of -1.00.

Table 2. *Spatial Misalignment* between training/evaluation corpora and the GeoNames gazetteer

Dataset	Genre	Min	1st Qu.	Median	3rd Qu.	Max
GeoWiki	Wikipedia	0.00	0.31	0.49	0.63	0.87
GeoCorpora	Social Media	0.35	0.90	0.96	0.99	1.00
WikToR	Wikipedia	0.49	0.87	0.92	0.96	1.00

Table 3. *Spatial Diversity Misalignment* between training/evaluation corpora and the GeoNames gazetteer

Dataset	Genre	Min	1st Qu.	Median	3rd Qu.	Max
GeoWiki	Wikipedia	-1.00	-0.30	-0.22	-0.14	0.43
GeoCorpora	Social Media	-1.00	-1.00	-1.00	-0.86	-0.43
WikToR	Wikipedia	-1.00	-1.00	-0.83	-0.70	-0.26

However, it is worth noting that there are also representation biases involved in gazetteer data that have often been used as references for geoparsers to search for candidate place information of a toponym. McDonough et al. (2019) questioned the colonist perspective from which databases such as GeoNames, the Alexandria Digital Library, and Wikipedia are built. They also pointed out the lack of temporal metadata for historial toponyms in gazetteers that have already been found to contain inadequate geographic knowledge about many parts of the world. This raises concerns that biases might accumulate as digital resources would replicate the content of their predecessors during their creation process. Therefore, we consider the study of historical aspects of representation biases involved in place name data as an intriguing future direction, which has now been made possible by tremendous efforts in applying knowledge graphs to building historical gazetteers such as Grossner and Mostern (2021).

5.2 Aggregation Bias

The second kind of geographic bias is the aggregation bias involved in toponym resolution. In many recent studies, the Earth's surface is divided into grid cells, and toponym resolution is approached as a classification task where the model predicts the most likely cell the current toponym should fall into based on loss function minimization. However, different discretization of the space will yield different prediction results. In the evaluation of Kulkarni et al. (2021), this Modifiable Areal Unit Problem (MAUP) in the prediction of toponym resolution is found to cause a trade-off between model generalization and prediction quality: finer granularity results in higher accuracy in denser regions with more toponyms in training data, while coarser granularity leads to better generalization over data on both global and local scales.

In addition to the prediction process, different patterns can also be learned by toponym resolution models when choosing different granularity during the training process. Coupled with joint minimization of losses at each level, the multi-level neural network architecture proposed by Kulkarni et al. (2021) can be one of the many possible solutions to dealing with the MAUP in the training of toponym resolution models.

5.3 Algorithmic Bias

Besides datasets, there is a common algorithmic bias intentionally introduced in toponym resolution. For instance, toponym resolution tends to prefer places with the largest population during place name disambiguation. While applying this simple population heuristic reduces computational complexity, and even brings about better performance than building more complex architecture in some circumstances, its limitations are evident. First, population information needs to be retrieved from gazetteers, and therefore, toponym resolution is made gazetteer-constrained. Second, population information serves as a geographic bias that hinders the fairness towards places with a smaller population in toponym resolution. Gazetteer-free toponym resolution has been studied as one of the many possible solutions recently and has shown promising results. Examples include topic modeling for place name disambiguation (Ju et al., 2016), modeling geographic profiles of words (DeLozier et al., 2015), spatial language representation learning at multiple levels (Kulkarni et al., 2021), and toponym co-occurrence representation learning (Fize et al., 2021). It is worth noting that gazetteer-free toponym resolution models can even output toponyms that are not inventoried in the gazetteer in the first place.

5.4 Evaluation Bias

Lastly, there is an evaluation bias (Suresh and Guttag, 2019) caused by how geoparsing performance was measured in previous research. How a geoparser performs across geographic space has not been taken into account, since the evaluation metrics are merely used to provide the overall performance instead of the performance at each annotated location (in a corpus). Our spatially-explicit performance evaluation has addressed this issue. However, the spatial (diversity) misalignment analysis in Section 5.1 raises another concern that geoparsing performance evaluation is more than geographically biased for lack of a geospatial perspective, because the evaluation bias in geoparsing can be exacerbated by the representation bias in these benchmark evaluation datasets.

6 Conclusions and Future Work

In this work, we presented a spatially-explicit evaluation of geoparsing performance across geographic space. We utilized $Recall$ and $MdnED$ to measure toponym recognition and resolution performance, respectively, and compared how deep-learning-based models that were claimed to achieve state-of-the-art results and their off-the-self counterparts perform across geographic space. By visualizing the normalized frequency distributions of the two geoparsing performance indicators, we discovered that all normalized frequency distributions are highly skewed. Then, we analyzed the spatial autocorrelation underlying regional variability in geoparsing performance by calculating the Getis-Ord Gi* statistic of $Recall$ and $MdnED$ at all annotated locations. We detected hot spots and cold spots in geoparsing performance, which reveals that geoparsing is geographically biased towards some regions and against others. Particularly, in toponym resolution that is sensitive to place name ambiguity, only cold spots were observed, and there was a stronger bias in the deep-learning-based CamCoder compared with the rule-based Edinburgh Geoparser. There was also strong regional variability in toponym resolution performance for highly ambiguous toponyms, as revealed by the standard deviations of $MdnED$ for annotated locations with the same name.

To probe the reason for the observed regional variability, we further evaluated geographic biases involved in geoparsing studies, ranging from spatial (diversity) misalignment of place coverage between (training and evaluation) corpora and the GeoNames gazetteer, the MAUP in toponym resolution, the preferences towards places with the largest population in place name disambiguation, to a biased perspective and usage of biased data in performance evaluation. While we have discussed how recent work, such as spatial representation learning and spatially-explicit geoparsing performance evaluations, can help remove some of the geographic biases embedded in geoparsing, there is still a long way to go towards directions such as developing geoparsers that can succeed in performing accurately in place name disambiguation across geographic space. We hope to highlight aforementioned biases for future geoparsing research, and we call for debiasing work on geoparsing. We also recommend further considerations about how much bias a task is able to bear when applying geoparsing in (geospatial) downstream tasks. Such considerations need to be taken into account in terms of dataset construction, algorithm design, and performance evaluations in geoparsing. Summing up, we rejected the claim that geoparsing is solved, and argued that it only appears so due to evaluation biases. Meanwhile, other geographic biases in geoparsing also need immediate attention.

7 Data and Software Availability

The toponym recognition and toponym resolution models we used are all publicly available as described in Section 3.1. The evaluation corpora and data patches used in our study are described in Section 3.2. The gazetteers used in our study are mentioned in Section 3.3. Section 5 contains information on how to access the training corpus and the country/region shapefile. The documentation about how to reproduce our study is available on GitHub[15].

Acknowledgement

This work was partially supported by the NSF award 2033521, "KnowWhereGraph: Enriching and Linking Cross-Domain Knowledge Graphs using Spatially-Explicit AI Technologies".

References

Barton, D. N., Blumentrath, S., and Rusch, G.: Policyscape—a spatially explicit evaluation of voluntary conservation in a policy mix for biodiversity conservation in Norway, Society & Natural Resources, 26, 1185–1201, 2013.

Basiri, A., Haklay, M., Foody, G., and Mooney, P.: Crowdsourced geospatial data quality: Challenges and future directions, 2019.

Beck, J., Böller, M., Erhardt, A., and Schwanghart, W.: Spatial bias in the GBIF database and its effect on modeling species' geographic distributions, Ecological Informatics, 19, 10–15, 2014.

Cui, J., Kong, X., Chen, J., Sun, J., and Zhu, Y.: Spatially explicit evaluation and driving factor identification of land use conflict in yangtze river economic belt, Land, 10, 43, 2021.

DeAngelis, D. L. and Yurek, S.: Spatially explicit modeling in ecology: a review, Ecosystems, 20, 284–300, 2017.

DeLozier, G., Baldridge, J., and London, L.: Gazetteer-independent toponym resolution using geographic word profiles, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, 2015.

Dunning Jr, J. B., Stewart, D. J., Danielson, B. J., Noon, B. R., Root, T. L., Lamberson, R. H., and Stevens, E. E.: Spatially explicit population models: current forms and future uses, Ecological Applications, 5, 3–11, 1995.

Fize, J., Moncla, L., and Martins, B.: Deep Learning for Toponym Resolution: Geocoding Based on Pairs of Toponyms, ISPRS International Journal of Geo-Information, 10, 818, 2021.

Forsius, M., Kujala, H., Minunno, F., Holmberg, M., Leikola, N., Mikkonen, N., Autio, I., Paunu, V.-V., Tanhuanpää, T., Hurskainen, P., et al.: Developing a spatially explicit modelling and evaluation framework for integrated carbon seques-

[15] https://github.com/zilongliu-geo/Geoparsing-Solved-Or-Biased

tration and biodiversity conservation: Application in southern Finland, Science of the Total Environment, 775, 145 847, 2021.

Fotheringham, A. S. and Wong, D. W.: The modifiable areal unit problem in multivariate statistical analysis, Environment and planning A, 23, 1025–1044, 1991.

Getis, A.: Reflections on spatial autocorrelation, Regional Science and Urban Economics, 37, 491–496, 2007.

Getis, A.: A history of the concept of spatial autocorrelation: A geographer's perspective, Geographical analysis, 40, 297–309, 2008.

Getis, A. and Ord, J. K.: The analysis of spatial association by use of distance statistics, in: Perspectives on spatial data analysis, pp. 127–145, Springer, 2010.

Goodchild, M.: Issues in spatially explicit modeling, Agent-based models of land-use and land-cover change, pp. 13–17, 2001.

Griffith, D. A.: Spatial autocorrelation, A Primer (Washington, DC, Association of American Geographers), 1987.

Gritta, M.: Research data supporting" Which Melbourne? Augmenting Geocoding with Maps", 2018.

Gritta, M., Pilehvar, M., and Collier, N.: Which melbourne? augmenting geocoding with maps, 2018a.

Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N.: What's missing in geographical parsing?, Language Resources and Evaluation, 52, 603–623, 2018b.

Gritta, M., Pilehvar, M. T., and Collier, N.: A pragmatic guide to geoparsing evaluation, Language resources and evaluation, 54, 683–712, 2020.

Grossner, K. and Mostern, R.: Linked Places in World Historical Gazetteer, in: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities, pp. 40–43, 2021.

Grover, C.: LT-TTT2 example pipelines documentation, Edinburgh: Edinburgh Language Technology Group, July, 24, 2008.

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

Hu, X., Zhou, Z., Kersten, J., Wiegmann, M., and Klan, F.: GazPNE2: A general and annotation-free place name extractor for microblogs fusing gazetteers and transformer models, 2021.

Hu, Y. and Janowicz, K.: An empirical study on the names of points of interest and their changes with geographic distance, arXiv preprint arXiv:1806.08040, 2018.

Hu, Y., Janowicz, K., and Prasad, S.: Improving Wikipedia-based place name disambiguation in short texts using structured data from DBpedia, in: Proceedings of the 8th workshop on geographic information retrieval, pp. 1–8, 2014.

Irwin, E. G. and Geoghegan, J.: Theory, data, methods: developing spatially explicit economic models of land use change, Agriculture, Ecosystems & Environment, 85, 7–24, 2001.

Janowicz, K., Hu, Y., McKenzie, G., Gao, S., Regalia, B., Mai, G., Zhu, R., Adams, B., and Taylor, K.: Moon landing or safari? a study of systematic errors and their causes in geographic linked data, in: The Annual International Conference on Geographic Information Science, pp. 275–290, Springer, 2016.

Janowicz, K., Gao, S., McKenzie, G., Hu, Y., and Bhaduri, B.: GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, 2020.

Jones, C. B. and Purves, R. S.: Geographical information retrieval, International Journal of Geographical Information Science, 22, 219–228, 2008.

Ju, Y., Adams, B., Janowicz, K., Hu, Y., Yan, B., and McKenzie, G.: Things and strings: improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling, in: European Knowledge Acquisition Workshop, pp. 353–367, Springer, 2016.

Kamalloo, E. and Rafiei, D.: A coherent unsupervised model for toponym resolution, in: Proceedings of the 2018 World Wide Web Conference, pp. 1287–1296, 2018.

Koenig, W. D.: Spatial autocorrelation of ecological phenomena, Trends in Ecology & Evolution, 14, 22–26, 1999.

Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., et al.: Openimages: A public dataset for large-scale multi-label and multi-class image classification, Dataset available from https://github. com/openimages, 2, 18, 2017.

Kulkarni, S., Jain, S., Hosseini, M. J., Baldridge, J., Ie, E., and Zhang, L.: Multi-Level Gazetteer-Free Geocoding, in: Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics, pp. 79–88, 2021.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C.: Neural architectures for named entity recognition, arXiv preprint arXiv:1603.01360, 2016.

Laparra, E. and Bethard, S.: A Dataset and Evaluation Framework for Complex Geographical Description Parsing, in: Proceedings of the 28th International Conference on Computational Linguistics, pp. 936–948, 2020.

Legendre, P.: Spatial autocorrelation: trouble or new paradigm?, Ecology, 74, 1659–1673, 1993.

Lieberman, M. D., Samet, H., and Sankaranarayanan, J.: Geotagging with local lexicons to build indexes for textually-specified spatial data, in: 2010 IEEE 26th international conference on data engineering (ICDE 2010), pp. 201–212, IEEE, 2010.

Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L., and Lao, N.: Multi-Scale Representation Learning for Spatial Feature Distributions using Grid Cells, in: International Conference on Learning Representations, 2020.

McDonough, K., Moncla, L., and Van de Camp, M.: Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora, International Journal of Geographical Information Science, 33, 2498–2522, 2019.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A.: A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR), 54, 1–35, 2021.

Nadeau, D. and Sekine, S.: A survey of named entity recognition and classification, Lingvisticae Investigationes, 30, 3–26, 2007.

Ord, J. K. and Getis, A.: Local spatial autocorrelation statistics: distributional issues and an application, Geographical analysis, 27, 286–306, 1995.

Overell, S. and Rüger, S.: Using co-occurrence models for placename disambiguation, International Journal of Geographical Information Science, 22, 265–287, 2008.

Quattrone, G., Capra, L., and De Meo, P.: There's no such thing as the perfect map: Quantifying bias in spatial crowd-sourcing datasets, in: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 1021–1032, 2015.

Reddy, S. and Dávalos, L. M.: Geographical sampling bias and its implications for conservation priorities in Africa, Journal of Biogeography, 30, 1719–1727, 2003.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge, International journal of computer vision, 115, 211–252, 2015.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D.: No classification without representation: Assessing geodiversity issues in open data sets for the developing world, arXiv preprint arXiv:1711.08536, 2017.

Shannon, C. E.: A mathematical theory of communication, The Bell system technical journal, 27, 379–423, 1948.

Sokal, R. R. and Oden, N. L.: Spatial autocorrelation in biology: 1. Methodology, Biological journal of the Linnean Society, 10, 199–228, 1978.

Suresh, H. and Guttag, J. V.: A framework for understanding unintended consequences of machine learning, 2019.

Syfert, M. M., Smith, M. J., and Coomes, D. A.: The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models, PloS one, 8, e55 158, 2013.

Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., and Pezanowski, S.: GeoCorpora: building a corpus to test and train microblog geoparsers, International Journal of Geographical Information Science, 32, 1–29, 2018.

Wang, J. and Hu, Y.: Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers, Transactions in GIS, 23, 1393–1419, 2019a.

Wang, J. and Hu, Y.: Are we there yet? evaluating state-of-the-art neural network based geoparsers using eupeg as a benchmarking platform, in: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities, pp. 1–6, 2019b.

Wang, J., Hu, Y., and Joseph, K.: NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages, Transactions in GIS, 24, 719–735, 2020.

Xu, E. and Zhang, H.: Spatially-explicit sensitivity analysis for land suitability evaluation, Applied Geography, 45, 1–9, 2013.

Yan, B., Janowicz, K., Mai, G., and Zhu, R.: xnet+ sc: Classifying places based on images by incorporating spatial contexts, in: 10th International Conference on Geographic Information Science (GIScience 2018), Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Yang, W., Ma, K., and Kreft, H.: Geographical sampling bias in a large distributional database and its effects on species richness–environment models, Journal of Biogeography, 40, 1415–1426, 2013.

Zeng, W., Lin, C., Lin, J., Jiang, J., Xia, J., Turkay, C., and Chen, W.: Revisiting the modifiable areal unit problem in deep traffic prediction with visual analytics, IEEE Transactions on Visualization and Computer Graphics, 27, 839–848, 2020.

Zhang, G. and Zhu, A.-X.: The representativeness and spatial bias of volunteered geographic information: a review, Annals of GIS, 24, 151–162, 2018.

Zhou, C., Sun, C., Liu, Z., and Lau, F.: A C-LSTM neural network for text classification, arXiv preprint arXiv:1511.08630, 2015.