

# Changes in Tweet Geolocation over Time: A Study with Carmen 2.0

Jingyu Zhang and Alexandra DeLucia and Mark Dredze


Department of Computer Science

Johns Hopkins University

{jzhan237, aadelucia, mdredze}@jhu.edu

## Abstract

Researchers across disciplines use Twitter geolocation tools to filter data for desired locations. These tools have largely been trained and tested on English tweets, often originating in the United States from almost a decade ago. Despite the importance of these tools for data curation, the impact of tweet language, country of origin, and creation date on tool performance remains largely unknown. We explore these issues with Carmen, a popular tool for Twitter geolocation. To support this study we introduce Carmen 2.0, a major update which includes the incorporation of GeoNames, a gazetteer that provides much broader coverage of locations. We evaluate using two new Twitter datasets, one for multilingual, multiyear geolocation evaluation, and another for usage trends over time. We found that language, country origin, and time does impact geolocation tool performance.

 <https://github.com/AADeLucia/carmen-wnut22-submission>

## 1 Introduction

Demographic studies leverage location-specific social media posts to track impactful events such as civil unrest (Sech et al., 2020; Chinta et al., 2021; Alsaedi et al., 2017; Littman, 2018), natural disasters (Wang et al., 2015), and disease spread (Xu et al., 2020). For social media posts from Twitter, researchers either collect posts from locations of interest in real-time with the Twitter API, or use third-party *Twitter geolocation* tools to identify tweet locations on an existing dataset. In Han et al. (2016), the authors distinguish between *user* and *tweet* geolocation. We focus on tweet geolocation in this work. These tools identify the location of a user or tweet based on tweet metadata (Dredze et al., 2013), tweet content (Alsaedi et al., 2017; Rahimi et al., 2016; Han et al., 2014; Wu and Gerber, 2018; Izbicki et al., 2019), and social networks (Rout et al., 2013; Jurgens, 2013).

While widely used, geolocation tools tend to be English-centric and are often not evaluated for global coverage or performance across time and language. These factors are important to study, since available user metadata, Twitter policies, and content patterns on which the tools depend on can change significantly over time.

In this work, we assess how the following factors impact geolocation tools:

1. **Language:** Is there a performance difference between languages, specifically between English and non-English tweets?
2. **Country:** Is there a performance difference between countries, specifically inside and outside the US?
3. **Time:** How does geolocation performance change over a large span of time? What differences in the data contribute to this performance change?

We measure performance in geolocation by coverage, i.e. the number of tweets that can be mapped to a location, and accuracy, the correctness of the assigned locations. When evaluated together, these metrics provide analogues to recall and precision, respectively.

To answer the above research questions, we analyze the performance of Twitter geolocation tool Carmen (Dredze et al., 2013), across time, language, and country of origin. In order to study performance across these factors, we introduce TWITTER-GLOBAL, a new geocoded multilingual and multiyear dataset (2013–2021) of 15.3M tweets. We created this dataset to fill a gap in other geolocation evaluation datasets that are either English-only (Han et al., 2012), or multilingual but restricted to short periods of time (Izbicki et al., 2019). We focus on Carmen since it is a rule-based tool that can be run quickly on large collections of tweets.

Since Carmen was built for English tweet

datasets from 2013, we update the tool and introduce Carmen 2.0. This updated version relies on GeoNames,<sup>1</sup> an open-source geographical dictionary, or gazetteer. In contrast to Carmen’s US and English-centric database, GeoNames provides global coverage in many languages. Through comparisons of GeoNames-augmented Carmen 2.0 with the original Carmen location database, we can study the effects of incorporating more non-US and non-English locations on geolocation performance.

In addition to studying Carmen 2.0’s performance with regard to different factors, we also include a longitudinal study of Twitter demographics over time from 2013–2021, with respect popularity across different countries, languages, and geolocation metadata. This study is on a collection of 5.7M tweets sampled from the 1% Twitter stream, which we refer to as TWITTER-RANDOM. The demographic and metadata analysis provides statistics to support design decisions for researchers developing their own geolocation tools.

We contribute the following:

1. Analysis of the effects of time, language, and country origin on Twitter geolocation tool performance.
2. Longitudinal study of user geolocation metadata availability and changes in frequency of tweets from different countries and languages.
3. Carmen 2.0, an improved version of the popular geolocation tool.
4. TWITTER-RANDOM, a randomly (1% based) sampled 5.7M Twitter dataset to support analysis of metadata and user trends over time (2013–2021).
5. TWITTER-GLOBAL, a geocoded multilingual, multiyear 15.3M Twitter dataset to support temporal and global geolocation evaluation.

All experiment code and data (tweet IDs) are released on in the GitHub code repository.

## 2 Related Work

Most work in Twitter geolocation focuses solely on tool development and performance, usually on English-centric datasets published years ago. In this paper we question how those tools would perform on Twitter datasets today, but focus on a single tool, Carmen.

**Geolocation Analysis** Kruspe et al. (2021) analyze the impact of Twitter policy changes on research. The authors study tweet metadata availability over time, such as exact coordinate availability and granularity of place objects. Most importantly, the authors discuss the impact of the 2019 Twitter policy change to remove precise locations from tweets (starting from 2019), and how that affects geolocation tools and researchers who depend on the coordinates. Their work limits their study to tweets from 2020–2021, and in our work we study these metadata patterns over a larger span of time, 2013–2021, in addition to the impact of other factors, such as language and country of origin, on geolocation tool performance. We compare our multi-year trend analysis to theirs in Section 6. This multi-year analysis is useful for researchers geolocating tweets in older Twitter datasets.

**Geolocation Tools** Most approaches for social media geolocation use tweet/user-level metadata (Dredze et al., 2013), tweet content, including hashtags, (Alsaedi et al., 2017; Rahimi et al., 2016; Han et al., 2014; Wu and Gerber, 2018; Halterman, 2017; Izbicki et al., 2019), and social networks (Rout et al., 2013; Jurgens, 2013). The UnicodeCNN geolocation tool (Izbicki et al., 2019) is notable because it is not English- or US-centric, and can infer location from multilingual tweet content.<sup>2</sup> Izbicki et al. (2019) also introduced a large, global geotagged dataset of 900M tweets across 100 languages, but this dataset is not appropriate for our temporal evaluation since it only includes tweets from 2017 to 2018. The authors did not provide a trend analysis on the dataset for comparison to our analysis in Section 6. Huang and Carley (2019) use a combination of all these features, and Ribeiro and Pappa (2017) create an ensemble classifier to combine existing methods, improving accuracy and coverage. Geolocation approaches for other social media platforms, such as Reddit, use similar methods (Harrigan, 2018).

There are a few ways to ascertain the location of a user or tweet: (1) use the coordinates embedded in one or more of the user’s tweets, (2) use the embedded place metadata, (3) use the user’s location string in their profile, (4) infer a location from the tweet content, and (5) leverage social network information. Methods (1) and (2) are most accurate, but less than 2% of tweets contain location

<sup>2</sup>The UnicodeCNN model is unavailable for comparison at time of writing.

<sup>1</sup><https://www.geonames.org/>

metadata (Kruspe et al., 2021). Method (4) is common (see tools above that use tweet content), but requires building more sophisticated language models as opposed to examining the metadata. Method (5) also performs well, but requires access to significantly more tweets in order to build the social network structure (Jurgens, 2013).

### 3 Carmen 2.0

In this paper we present Carmen 2.0, an updated version of geolocation tool Carmen. We aim to increase the coverage and robustness of Carmen to language and countries by using an open-source database, GeoNames.

In addition to a new location database, we include other performance improvements, such as compatibility with Twitter API v2 (see Appendix §B). Since Carmen 2.0 does not change the core functionality, we focus on the construction and use of the internal location database, and we direct the reader to Appendix §A for a review of the location resolvers or Dredze et al. (2013) for more details.

#### 3.1 Carmen: A Review

Carmen, introduced by Dredze et al. (2013), uses tweet and user profile metadata for geolocation.<sup>3</sup> Carmen has three “resolvers” which use different information from the tweet: (1) embedded coordinates in the `geo` object,<sup>4</sup> (2) matching the Place object to the internal locations database, and (3) mapping the user profile location string to the internal location database.

##### 3.1.1 Original Location Database

The tweet location is resolved to an entry in an internal database of 7041 places.<sup>5</sup> Locations are stored in JSON form, where each location object has city, county, state/province, country, coordinates, and “aliases.”

The original database was developed from tweets available at the time that Carmen was released in 2013, specifically 10K tweets sampled from the Bergsma et al. (2013) dataset. This dataset consists of roughly 4 billion tweets between May 2009 and August 2012, in addition to 80 million tweets from users who follow specific feeds for locations and

<sup>3</sup>The Python version of the tool is available at <https://github.com/mdredze/carmen-python>.

<sup>4</sup>According to Kruspe et al. (2021), Twitter stopped including coordinates in 2019.

<sup>5</sup>The original paper says 4K locations, but the database was expanded by the authors between 2013 and 2022.

	Original		GeoNames	
	Count	Percent	Count	Percent
City	4401	62.51%	24568	33.24%
County	1995	28.33%	45154	61.08%
State	461	6.55%	3947	5.34%
Country	184	2.61%	252	0.34%
Total	7041		73921	

Table 1: The statistics of city, county, state, and country-level locations in the original Carmen location database and the new GeoNames database versions developed for Carmen 2.0. The GeoNames-augmented databases have more than 10 times the number of location entries than `Original`. Percentage refers to portion of the database dedicated to each granularity.

languages. The internal location database was constructed from the geotagged places in the development set, and then augmented through manual and automatic collection of **aliases**, or alternate names. The motivation for including aliases stemmed from inconsistent names for Twitter places, mostly due to location references in different languages, e.g., “polnia” and “poland.” Aliases also include colloquial names for a place, such as “the big apple” for New York City, which could be found in a user profile location string. Place information was included as much as possible (i.e., province) by obtaining full location information from Yahoo’s PlaceFinder API.<sup>6</sup>

Thus, because of the origin of its location database, which was built from tweets between 2009–2012, Carmen’s database is biased towards common locations and languages in tweets before 2012, primarily English tweets from the US. Further, the database does not align with an external knowledge base, so Carmen locations cannot be directly matched against other place information. These limitations prompted our updates, without which we would be unable to answer the questions of this paper.

#### 3.2 Expanded Database: GeoNames

The original Carmen location database was crafted from a Twitter sample (see §3.1.1). This decision biased Carmen towards locations popular with Twitter users from 2009–2012, which is not representative of today’s users. Further, the location identifiers were unique to Carmen, and thus could not be meaningfully shared for external analysis or easily augmented with other place information.

<sup>6</sup>This API is no longer available.

To remedy both issues, we augmented the internal Carmen database with GeoNames, an open source geographical database that covers all countries and millions of place names.

**GeoNames Structure** GeoNames has a hierarchical structure, where every entry has a link to its parent. For example, Austin (`city`) has a link to Texas (`admin1`), which in turn has a link to United States (`US`, `country`). Sometimes `admin2` is also present, which refers to a county in the US. All counties, administrative regions (i.e., state or province), and countries were also added to the database.

Similar to Carmen’s aliases, GeoNames contains a list of “alternate names” of each entry. While Carmen’s methods were geared towards colloquial names, GeoNames contains the name of each entry in many languages, in addition to a few colloquial names.

**Database Merging** While GeoNames contains cities from all over the world, we only include cities with a population over 15K. This is to ensure a more efficient location resolution process, since tweets are more likely to originate from highly populated places. Also, only including more populated locations is important for user privacy, since a user can remain more anonymous when aggregated in a large group. We discuss more ethical concerns surrounding geolocation, such as privacy, in Section 7. We use GeoNames to create two versions of a new Carmen location database: (1) GeoNames only and (2) a merged GeoNames and Carmen database. The **GeoNames-only** database, (1), converts the GeoNames format for cities, states/provinces, and countries to Carmen-formatted JSON objects.

The **GeoNames-combined** database, (2), required matching Carmen database entries to GeoNames entries. We matched locations based on string similarity of location name, the distance between coordinates, and country name. To maintain accuracy of mapped locations, our mapping criteria was strict and 4,467 out of 7,041 (63.44%) Carmen locations were successfully mapped to GeoNames. We then added the alternate names of each location in Carmen to the new GeoNames backed location entries. The merged version also contains all county, state/province, and country entries as (1). The remaining entries in Carmen were that were unable to be matched were disregarded. A spot-check on these unmatched locations confirmed they were

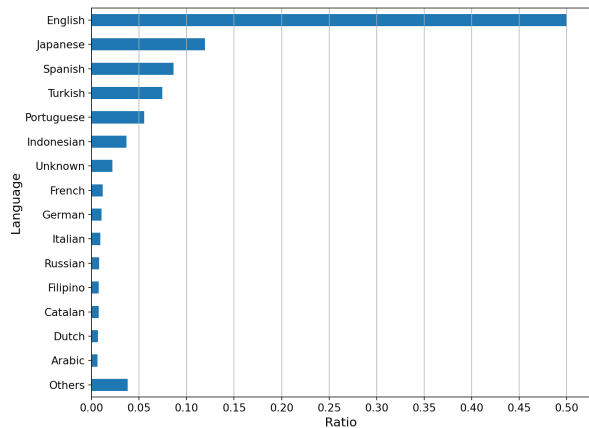


Figure 1: Language distribution for tweets in TWITTER-GLOBAL. Only the top 15 languages are shown. Languages are identified by tweet metadata

cities with population less than 15K or contained errors, such as incorrect county or province information.

Both **GeoNames-only** and **GeoNames-combined** contain 73921 entries. The number of entries is the same since they differ only in alias lists. Database details are in Table 1.

## 4 Geolocation Evaluation

Through comparing Carmen’s original database (see §3.1.1) with the new GeoNames based database (see §3.2), we can answer our research questions from Section 1 to see how geolocation tool coverage and accuracy change with respect to language, country of origin, and time. The performance is evaluated with similar metrics of other geolocation tools mentioned in §2. In addition to updating Carmen, we also created two datasets to support our analysis.

### 4.1 Ground Truth Data

There are a handful of “standardized” datasets for Twitter tweet geolocation evaluation (Han et al., 2012, 2016), but they often are not global, multilingual, or recent. In this work we introduce two datasets, TWITTER-GLOBAL and TWITTER-RANDOM. Despite the temporal (2011–2012) and language (English only) limitations of the popular TWITTER-WORLD (Han et al., 2012) geolocation evaluation dataset, we include Carmen’s performance in Appendix §D so others can refer to it for comparison.

**Twitter-Global** This new geolocation evaluation dataset is collected from multiple geolocation filter-

ing Twitter streams that are designed to cover the world.<sup>7</sup> The data from these streams was collected from 2013 to 2021 for a total of 15.3M tweets, balanced over the years. Due to the nature of the stream, all tweets are “geotagged” with Twitter Place objects. The ground truth for tweets are the place names and coordinates in the Place metadata. We follow previous work in using geotagged tweets as ground truth, although we note the bias introduced by only evaluating on geotagged data (Pavalanathan and Eisenstein, 2015).

Unlike popular geolocation evaluation dataset TWITTER-WORLD, our dataset is multilingual. While Han et al. (2012) removed non-English tweets in order to not make it “easy” on the tool, we want to ensure the geolocation tools work in a multilingual setting. Language distribution is in Figure 1. Since TWITTER-GLOBAL includes samples from North America, we omit evaluation on another popular evaluation dataset, TWITTER-US (Han et al., 2012). Izbicki et al. (2019) introduced a larger, global geotagged dataset of 900M tweets across 100 languages, but is not appropriate for our temporal evaluation since it only includes tweets from 2017 to 2018.

**Twitter-Random** In addition to the new geolocation evaluation dataset, we introduce a multiyear random sample. This sample is useful for analyzing shifts in usage patterns across the world with respect to metadata inclusion, language, etc.

We created this dataset by sampling 100K tweets per month from the Twitter Streaming API between 2013 and 2021, resulting in 5.7M tweets.

## 4.2 Evaluation Metrics

Evaluating geotagging performance is grouped into two categories: (1) *coverage*, or percentage of the data our method successfully found a location, and (2) *accuracy*, or how well the proposed locations compare to the ground truth. We use metrics similar to other work on Twitter geolocation. Formulas are provided in Appendix §C.

### 4.2.1 Coverage

Given a tweet, Carmen resolves it to an entry in the internal database if such mapping can be found. Since Carmen only uses information from tweet and user profile metadata, we define **coverage** as the fraction of resolved tweets among all tweets

<sup>7</sup>While streams are meant to cover the entire world, there are gaps due to Twitter API restrictions.

that have location information (i.e. has a Twitter Place object). Coverage is similar to *recall* and *sensitivity*, but does not incorporate whether the prediction is correct.

### 4.2.2 Accuracy

Coverage gives us a good metric of Carmen’s sensitivity to locations contained in tweets. However, it does not evaluate the *correctness* of the mapped results. We measure the location mapping accuracy by string comparison (country, state/province, city) and by geographical distance. These metrics are referred to as **match ratio** and **distance**.

**Match Ratio** A predicted location can be accurate on different levels of granularity, such as a correct state or province prediction, but incorrect city prediction. The **match ratio** metric awards partial credit for correct identification of a country or state even if another portion of the prediction is incorrect, such as the city. Match ratio on level  $L$ , denoted  $mr_L$ , where  $L \in \{\text{country, admin, city}\}$ , is the ratio of the number of resolved tweets where the prediction is correct on level  $L$  over the total number of tweets where  $L$  is available in the ground truth. We restrict the denominator to tweets where the level is available, since it is unfair to penalize the model for an “incorrect” city prediction when the city is not available in the ground truth.

**Distance** We also use geographical distance to measure accuracy. This metric is inspired by Eisenstein et al. (2010) and Cheng et al. (2010) and their calculation of regression performance, or mean and median distance between proposed location coordinates and ground truth.

Distance,  $d$ , is measured as the geodesic distance, calculated with `geopy`, between the resolved location and the ground-truth tweet coordinates. We calculate distance at the dataset level, which is the average distance over all tweets, where 0 is best. In addition to the average distance, we also consider “accuracy at  $K$ ”, or  $\text{Acc}@K$ , the ratio of resolved tweets such that the distance error does not exceed  $K$  miles (Ribeiro and Pappa, 2017; Han et al., 2014). This metric is less influenced by outliers than  $d$ .

## 5 Experiments

As enumerated in §1, we are interested in how the following factors impact geolocation tool performance: **language**, **country**, and **time**.

To answer these questions we perform an ablation study over Carmen location databases (see §3.2) and different subsets of TWITTER-GLOBAL (see §4.1).

### 5.1 Performance across Language

Many geolocation tools (and even evaluation datasets (Han et al., 2012)) focus on English tweets. We can analyze the performance difference of English-biased tools by comparing the performance of Carmen’s original English-centric database with the GeoNames-augmented ones on multilingual data. Since TWITTER-GLOBAL is multilingual, we create two subsets of English and Non-English data, as identified by the tweet language metadata. Tweets with “unknown” language tag are omitted. Since the GeoNames-only and GeoNames-combined location databases contain translations for location names, we expect Carmen to perform better with these over the original database, as corroborated in Table 2.

Overall, Carmen has better coverage for English tweets than on non-English with all location databases (roughly 49% compared to 32-41%). While the coverage on the English data is the same for the three databases (less than 2% difference), there is a large difference in coverage for non-English tweets. Both GeoNames-based databases were able to provide predictions for 42% of tweets and the Original database only provided matching 32% of tweets. Accuracy also differs between databases and language splits, but only at the higher granularities of admin (state/province) and city level, where the match ratio drops from 95% on English data to 66-75% for non-English at the admin level and from 48% to 14-20% at the city level across databases. Country-level accuracy remains stable at a 99% match ratio. The decrease in accuracy at the admin and city levels is also apparent through the distance metrics, where average distance is higher for non-English than English tweets. The high distance error for the GeoNames-only database can be attributed to different coordinates between the GeoNames and Twitter places gazetteer entries, and prediction error within large countries, such as the US and India, which can be detrimental.

In summary, using a multilingual geolocation tool can increase geolocated data for studies, with highest accuracy at the country level.

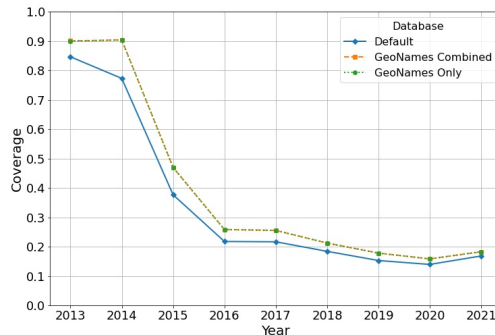


Figure 2: Ablation over Carmen location database and performance over data from different years of Twitter data, 2013-2021. Evaluated on TWITTER-GLOBAL. Metric is coverage, where higher score is better.

### 5.2 Performance across Countries

Similar to concerns with language bias, geolocation tools can also be US-centric. In order to analyze difference in performance across countries, we simplify the study to inside and outside of the US. We split TWITTER-GLOBAL into “US” and “non-US” categories for the evaluation. Similar to the language performance, we expect the GeoNames-augmented databases to provide an advantage over the original location database, due to the alternate names list. The results are shown in Table 3.

There is a similar trend between US/non-US split and English/non-English split. Overall, all databases have higher coverage of locations inside the US (50%) than outside of the US (32-42%), possibly confounded by differences in language. However, using a multilingual non-US based database helps with coverage significantly, as shown in the difference between GeoNames-augmented databases (42%) and the Original database (32%).

Accuracy is also better inside the US, as seen with the match ratio at the state/province (99% vs 60-66%) and city levels (54% vs 11-19%). Average distance is also higher for non-US locations, except for GeoNames-only which is most likely due to difference in coordinates for large countries.

### 5.3 Performance over Time

Carmen’s performance over time degrades significantly between 2015 and 2021, as shown in Figure 2. In 2013–2014, Carmen has 80-90% coverage with all databases, but this coverage drops to 40-50% in 2015 and below 20% after 2018. This drop is most likely due to Carmen’s heavy reliance on

Language	Database	Coverage	$mr_{country}$	$mr_{admin}$	$mr_{city}$	$d$	Acc@10	Acc@100	Acc@1000
English	GeoNames-Only	49.58%	99.42%	95.63%	47.49%	853.9	0.81	0.85	0.86
	GeoNames-combined	49.63%	99.43%	94.36%	47.69%	58.7	0.81	0.91	0.99
	Original	48.14%	99.35%	94.94%	48.90%	46.4	0.78	0.91	1.00
Non-English	GeoNames-Only	41.77%	99.36%	66.50%	20.13%	482.3	0.84	0.88	0.88
	GeoNames-combined	41.78%	99.35%	66.83%	20.27%	105.3	0.84	0.90	0.99
	Original	32.27%	98.95%	75.61%	14.22%	106.2	0.67	0.87	0.99

Table 2: Ablation over Carmen location database and performance on English and non-English tweets. Evaluated on TWITTER-GLOBAL. “Acc@ $K$ ” represents the ratio of tweets predicted within  $K$  miles of the ground truth. Higher values are best for all metrics except distance ( $d$ ).

Origin	Database	Coverage	$mr_{country}$	$mr_{admin}$	$mr_{city}$	$d$	Acc@10	Acc@100	Acc@1000
US	GeoNames-only	50.56%	99.37%	99.87%	53.66%	994.2	0.79	0.84	0.84
	GeoNames-combined	50.60%	99.37%	99.87%	53.81%	23.6	0.79	0.91	1.00
	Original	51.03%	99.93%	99.96%	55.33%	23.7	0.79	0.91	1.00
non-US	GeoNames-only	42.63%	99.37%	61.51%	18.73%	439.3	0.84	0.89	0.89
	GeoNames-combined	42.65%	99.37%	60.81%	18.88%	121.2	0.84	0.90	0.98
	Original	32.89%	98.45%	66.11%	11.10%	118.0	0.67	0.87	0.99

Table 3: Ablation over Carmen location database and performance on tweets originating from and outside of the United States (US). Evaluated on TWITTER-GLOBAL. “Acc@ $K$ ” represents the ratio of tweets predicted within  $K$  miles of the ground truth. Higher values are best for all metrics except distance ( $d$ ).

tweet metadata as opposed to tweet content or other features, which has decreased over time. We discuss the impact of metadata availability further in Section 6.3.

## 6 Longitudinal Analysis of Twitter User Location

We have seen how using a less biased geolocation tool offers better performance with respect to coverage and accuracy. However, despite the overall better performance with the GeoNames-combined location database, coverage still varied greatly when evaluated over time, as in Section 5.3. To better understand this performance difference and to provide insights for other geolocation researchers, we present a longitudinal study of trends in location metadata availability and user demographics. All metadata and demographic statistics are gathered from TWITTER-RANDOM (see §4.1).

### 6.1 Location Metadata Availability

As discussed in §2, Twitter geolocation tools make use of tweet/user-level metadata, tweet text, and social network information. Tools that exclusively use tweet and/or user metadata are most at the mercy of changes to Twitter API or policy.

As shown in Figure 3, the rate of tweets in the random stream with tagged Places increased slightly from 2013 to 2014 and then decreased from 2% to 0.5% from 2014 to 2021. This 75% decrease

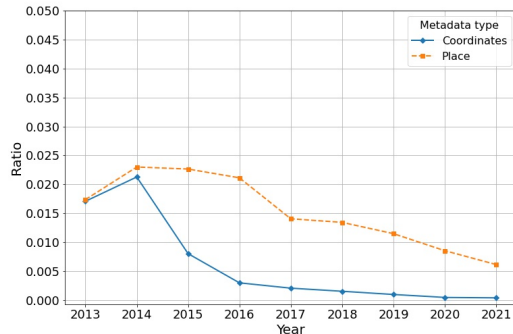


Figure 3: Prevalence of tweet metadata over time in TWITTER-RANDOM. We limit to metadata commonly used in geolocation of users or tweets. Note scale is from 0-5%.

represents millions of tweets. While inclusion of place information has declined, the rate of place types has remained the same. High-granularity types like points of interests (POIs) and neighborhoods are largely unused (less than 1% of Place objects), followed by country- and state/province-level tags (5% and 11%). The most common type by far is at the city level, comprising 83% of tagged place types.

The number of tweets with embedded coordinates has decreased even more than tagged places starting in 2015, even before the 2019 Twitter policy that removed coordinates. This decrease is most

likely due to another Twitter policy enacted April 2015 which changed the default “precise location” setting from enabled to disabled.<sup>8</sup> The only meta-data that has stayed consistent since 2013 is the user profile location field, which is available in 60% of tweets.

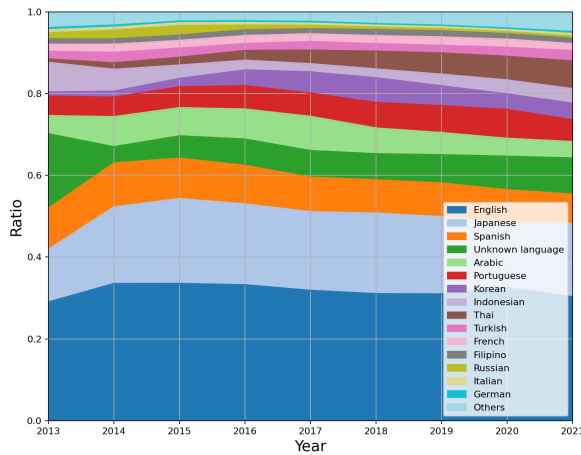


Figure 4: Language distribution for tweets in TWITTER-RANDOM over time. Languages are identified by tweet metadata

## 6.2 Twitter Demographics

In addition to changes in metadata availability, we also analyzed the change in countries and languages present in the random stream. No geolocation is needed for the language analysis, as it is included in tweet metadata, but we are limited by tweets that can be geotagged by Carmen 2.0 (GeoNames-combined database) for the country analysis. While this geotagging biases the sample to tweets with location information, this is representative of the distribution other researchers can expect from geotagged tweets in the random stream. Carmen was able to identify locations for 21% of the data, or 1.2M tweets.

**Country** The top 10 countries in the dataset are (in descending order): United States (US), United Kingdom (UK), Brazil, Indonesia, Japan, Argentina, India, France, Philippines, and Thailand. The US has a significantly larger share of tweets, roughly 30%. In comparison, tweets from the UK are only 6% of all tweets. The share of every country in the dataset is shown in Figure 5. The overall numbers can often hide year-specific trends. Within the top countries, Indonesia decreases from 11% in

<sup>8</sup><https://www.wired.com/story/twitter-location-data-gps-privacy/>

2013 to less than 5% after 2015. Inversely, India starts with very few tweets and steadily grows to roughly 9% of tweets. The other countries remain largely stable over the years.

**Language** The top languages follow the languages spoken in the top countries very closely, as shown in Figure 4. English comprises about 30% of all languages, followed by Japanese, Spanish, Arabic, Portuguese, Korean, Indonesian, Thai, and Turkish. Following the decrease in tweets from Indonesia, Indonesian tweets also decreased from 7% in 2013 to 4% in 2021. In the same time frame, tweets in Hindi follow the pattern of tweets from India, increasing from 0% to 1% of all languages. While not in the top languages or countries, there is also a decrease in tweets from Russia and in Russian from 2015 (2.5%) to 2021 (0.5%).

The Twitter language identification system likely changed between 2013 and 2014, as “unknown” languages dropped from 18% to 4%. This rate of unknown languages steadily increases to 8% in 2021, possibly due to increase of users tweeting in languages not officially supported by Twitter.

## 6.3 Impact on Geolocation Tools

Researchers applying existing tools to their own datasets should consider the locations and languages best represented by the tools, in addition to which metadata (if any) the tool relies on. Due to the large distribution of languages, it is important for geolocation tools to have multilingual support to increase coverage and accuracy. Further, the metadata trends in Section 6.1 suggest that geolocation tools should be frequently checked for API and policy compatibility.

Carmen’s performance analyzed over time in Section 5.3 is a prime example of how Twitter policy changes can affect geolocation tools. Carmen relies heavily on tweet metadata, specifically the presence of coordinates and place objects, but the prevalence of this information has decreased since 2015. A tool less reliant on metadata and more based on content or other signals, could be more temporally robust.

Ensuring a geolocation tool is temporally robust, i.e., has the same performance over time, is important for identifying tools that need to be periodically updated with new data (Dredze et al., 2016). This is especially important for tools that use features that can be subject to distribution shift, such as social networks, tweet content, and metadata availability.



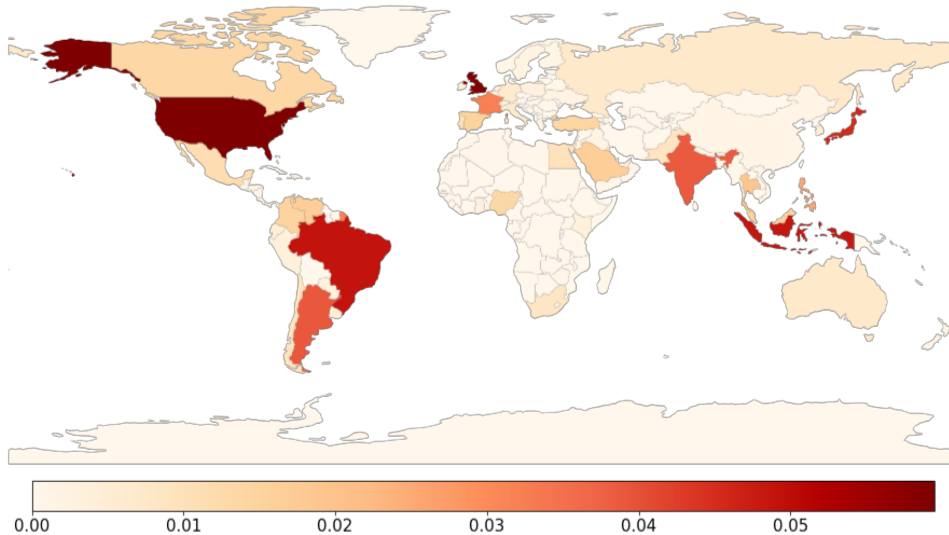


Figure 5: Distribution of country origin of tweets in TWITTER-RANDOM, a subset of the public Twitter API stream from 2013-2021. Locations are identified by Carmen. Scale is from 0 to 0.06 to show more detail. The United States represents 30% of tweets (0.3) and is capped to 0.06 for visualization purposes.

## 7 Ethical Concerns

Two issues that arise when geolocating users on social media: (1) privacy concerns and (2) consequences of incorrect predictions.

The privacy concerns are related to surveillance and revealing sensitive locations of users, such as their home address. Since Carmen only uses metadata provided by the user in the form of tagged places, coordinates, and profile location, it only infers locations readily shared by users. [Kruspe et al. \(2021\)](#) provide a helpful discussion of applications that require differing levels of location granularity, such as disaster relief or disease spread requiring more precise information (high granularity) versus marketing campaigns or opinion tracking (low granularity).

An issue with low granularity arises in high-risk applications where low precision is not helpful, such as tracking disease spread within a country. A possible solution for balancing higher granularity and user privacy is to map a user’s location to the largest city closest to the user. Carmen 2.0 does this automatically since the database only contains cities with population greater than 15000.

There can be negative consequences to using incorrectly inferred locations, such as in tracking high-risk emergencies like disease spread and civil unrest. Geolocation tool performance ablation over granularity, language, and country, is important for researchers to make informed decisions about location accuracy.

## 8 Conclusion

Geolocation tweets is useful for researchers that need to filter tweets to those originating in specific locations to study health, opinions, etc, by demographic. In this work we study and discuss the impact the factors of language, country origin, and time, can have on tweet geolocation.

To support our study we introduced Carmen 2.0, an updated version of geolocation tool Carmen ([Dredze et al., 2013](#)) backed by an open-source database, GeoNames. In addition to the tool, we introduced two datasets: (1) TWITTER-GLOBAL, a Twitter geolocation evaluation dataset for language, country, and time ablation studies, and (2) TWITTER-RANDOM, a sample of the worldwide Twitter stream from 2013-2021 for studying general country and language demographics and metadata availability over time.

We found a significant difference in performance in the ablation, with higher performance for English and US-based tweets. Also, we provided trends in metadata availability from 2013 to 2021, and discuss reasons for the decline in coordinates and place metadata. For future work in Twitter tweet geolocation, we encourage the use of content and metadata fields, such as user profile location. Focus on these consistently available metadata can make a tool robust to policy changes. Also, we encourage evaluating the geolocation model across language, time, and countries to ensure fair performance.

## Acknowledgment

We thank Arya McCarthy, Nathaniel Weir, and the anonymous reviewers for their time, helpful edits, and suggestions.

## References

- Nasser Alsaedi, Pete Burnap, and Omer Rana. 2017. [Can We Predict a Riot? Disruptive Event Detection Using Twitter](#). *ACM Transactions on Internet Technology*, 17(2):18:1–18:26.
- Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. 2013. [Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1019, Atlanta, Georgia. Association for Computational Linguistics.
- Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. [You are where you tweet: A content-based approach to geo-locating twitter users](#). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, page 759–768, New York, NY, USA. Association for Computing Machinery.
- Abhinav Chinta, Jingyu Zhang, Alexandra DeLucia, Mark Dredze, and Anna L. Buczak. 2021. [Study of Manifestation of Civil Unrest on Twitter](#). In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 396–409, Online. Association for Computational Linguistics.
- Mark Dredze, Miles Osborne, and Prabhajan Kambaradur. 2016. [Geolocation for Twitter: Timing Matters](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1064–1069, San Diego, California. Association for Computational Linguistics.
- Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. 2013. [Carmen: A Twitter Geolocation System with Applications to Public Health](#). Association for the Advancement of Artificial Intelligence.
- Jacob Eisenstein, Brendan O’Connor, Noah A. Smith, and Eric P. Xing. 2010. [A Latent Variable Model for Geographic Lexical Variation](#). In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA. Association for Computational Linguistics.
- Andrew Halterman. 2017. [Mordecai: Full Text Geoparsing and Event Geocoding](#). *Journal of Open Source Software*, 2(9):91.
- B. Han, P. Cook, and T. Baldwin. 2014. [Text-Based Twitter User Geolocation Prediction](#). *Journal of Artificial Intelligence Research*, 49:451–500.
- Bo Han, Paul Cook, and Timothy Baldwin. 2012. [Geolocation Prediction in Social Media Data by Finding Location Indicative Words](#). In *Proceedings of COLING 2012*, pages 1045–1062, Mumbai, India. The COLING 2012 Organizing Committee.
- Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. [Twitter Geolocation Prediction Shared Task of the 2016 Workshop on Noisy User-generated Text](#). In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 213–217, Osaka, Japan. The COLING 2016 Organizing Committee.
- Keith Harrigian. 2018. [Geocoding Without Geotags: A Text-based Approach for reddit](#). In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 17–27, Brussels, Belgium. Association for Computational Linguistics.
- Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. [Tweets from Justin Bieber’s heart: the dynamics of the location field in user profiles](#). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. Association for Computing Machinery, New York, NY, USA.
- Binxuan Huang and Kathleen Carley. 2019. [A Hierarchical Location Prediction Neural Network for Twitter User Geolocation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4732–4742, Hong Kong, China. Association for Computational Linguistics.
- Mike Izbicki, Vagelis Papalexakis, and Vassilis Tsotras. 2019. [Geolocating Tweets in any Language at any Location](#). In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19*, pages 89–98, New York, NY, USA. Association for Computing Machinery.
- David Jurgens. 2013. [That’s What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):273–282. Number: 1.
- Anna Kruspe, Matthias Häberle, Eike J. Hoffmann, Samyo Rode-Hasinger, Karam Abdulahhad, and Xiao Xiang Zhu. 2021. [Changes in Twitter geolocations: Insights and suggestions for future usage](#). In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 212–221, Online. Association for Computational Linguistics.
- Justin Littman. 2018. [Charlottesville Tweet Ids](#). Publisher: Harvard Dataverse type: dataset.
- Umashanthi Pavalanathan and Jacob Eisenstein. 2015. [Confounds and Consequences in Geotagged Twitter Data](#). In *Proceedings of the 2015 Conference on*

- Empirical Methods in Natural Language Processing*, pages 2138–2148, Lisbon, Portugal. Association for Computational Linguistics.
- Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2016. [pigeo: A Python Geotagging Tool](#). In *Proceedings of ACL-2016 System Demonstrations*, pages 127–132, Berlin, Germany. Association for Computational Linguistics.
- S. Ribeiro and G. Pappa. 2017. [Strategies for combining Twitter users geo-location methods](#). *GeoInformatica*.
- Dominic Rout, Kalina Bontcheva, Daniel Preotjiuc-Pietro, and Trevor Cohn. 2013. [Where’s @wally? a classification approach to geolocating users based on their social ties](#). In *Proceedings of the 24th ACM Conference on Hypertext and Social Media, HT ’13*, pages 11–20, New York, NY, USA. Association for Computing Machinery.
- Justin Sech, Alexandra DeLucia, Anna L. Buczak, and Mark Dredze. 2020. [Civil Unrest on Twitter \(CUT\): A Dataset of Tweets to Support Research on Civil Unrest](#). In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 215–221, Online. Association for Computational Linguistics.
- Haoyu Wang, E. Hovy, and Mark Dredze. 2015. [The Hurricane Sandy Twitter Corpus](#). In *AAAI Workshop: WWW and Public Health Intelligence*.
- Congyu Wu and Matthew S. Gerber. 2018. [Forecasting Civil Unrest Using Social Media and Protest Participation Theory](#). *IEEE Transactions on Computational Social Systems*, 5(1):82–94. Conference Name: IEEE Transactions on Computational Social Systems.
- Paiheng Xu, Mark Dredze, and David A. Broniatowski. 2020. [The Twitter Social Mobility Index: Measuring Social Distancing Practices With Geolocated Tweets](#). *Journal of Medical Internet Research*, 22(12):e21499.

## A Carmen Review

### A.1 Aliases

The alias list was constructed through two methods: (1) manually filtering resolved user location strings, and (2) using the user clustering method from Bergsma et al. (2013). For (1), common user location strings were resolved with Yahoo’s PlaceFinder API and then manually filtered and merged. In (2), users were clustered based on social network, fullnames, and the profile location strings. This process discovered that “balto” is an alias for “Baltimore”, based on the frequency that users with “balto” in their profile location communicate with “Baltimore” users.

### A.2 Resolvers

Carmen includes three location resolvers to map from the tweet to a location in the internal database. The default settings are to use the resolvers in the following order, but this is user configurable: geocode (coordinates), place, and profile.

**Geocode** Some tweets (before 2019) contain exact coordinates, and we use these coordinates to find the closest location in our internal database. The distance threshold between our internal location and the coordinates is user configurable.

**Place** A Twitter Place object is a JSON that is returned with the tweet, but only 2% of tweets contain a place (Kruspe et al., 2021). The object is in a different location in API v2 and must be specifically requested, but the object itself has not significantly changed. The place includes an ID that refers to a Twitter Places database, the place type (neighborhood, city, admin), name, fullname, country code, country name, and a bounding box.<sup>9</sup> Twitter Places are supported by Foursquare and Yelp (Kruspe et al., 2021).

**Profile** If a tweet does not contain place or coordinate information, the user profile resolver is used. As reported by Kruspe et al. (2021), only 30-40% of tweets contain user profile location information. While more users have their profile location filled in, the information is a free-text field completed by the user and is not restricted, thus some users put jokes or made-up locations (Hecht et al., 2011).

<sup>9</sup><https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/geo>

The profile resolver matches the string to the internal database by normalizing it (e.g., removing punctuation), identifying state or country names with regular expressions, and then matching the string, along with the state/country, against the location database.

Twitter introduced similar functionality with their Profile Geo Enrichment in the paid Enterprise API, but not all user location strings can be geocoded.<sup>10</sup>

Like the place object, accessing the user location string is different in API v2, and needs to be requested separately from the tweet object.

## B Carmen 2.0 Updates

### B.1 Functionality Updates

The Carmen code was updated to be compatible with tweets in Twitter API v2 format. As mentioned in §A, the placement of some metadata has changed in the new API. In Carmen 2.0, besides obtaining coordinates from the “coordinates” field of a Tweet object, we also obtain coordinates from the bounding box coordinates from the place object, if it exists. We use the average of all bounding box coordinates as the coordinates used for the geocode resolver. Although this is less accurate and is not an exact coordinate compared to the “coordinates” field, it still serves as a reliable source of location metadata.

Another improvement is a faster geocode resolver. The algorithm uses coordinates from the internal location database to group the known locations into *cells*. The default cell is size 0.5, which groups location within 0.5 degrees of each other, or 34.5 miles for latitude and 27.3 miles for longitude. For example, a coordinate of (1.2, 1.3) will be mapped to a cell containing all coordinates within the interval  $[0.75, 1.25) \times [1.25, 1.75)$ . The grouping is performed at Carmen initialization, so inference is a limited linear search over all locations in the database that are in the same cells as the query coordinates. Because different gazetteers might select different coordinate points for the same location, the design of cells gives a margin of error and allow the correct location entry to be mapped even if the coordinates does not match exactly.

<sup>10</sup><https://developer.twitter.com/en/docs/twitter-api/enterprise/enrichments/overview/profile-geo>

	Resolved/s	Tweets/s
Original	263.03	655.41
GeoNames-only	120.14	297.20
GeoNames-combined	140.51	311.13

Table 4: Processing speed for different Carmen 2.0 models. Resolved/s is the average number of resolved tweets per second, and Tweets/s is average number of processed tweets per second.

## B.2 Processing Speed

Table 4 shows processing speed of Carmen 2.0 with the different databases. To measure speed we use two metrics: (1) resolved tweets per second, the average number of tweets that Carmen resolves per second, and (2) tweets per second, which is the average number of processed tweets per second. The Original database, with only 7K locations, is faster than the GeoNames-only and GeoNames-combined databases, which have 74K locations. Despite this 10x increase in database size, the speed does not reduce linearly with the number of locations. This sublinear scaling is important for addition of new locations, such as incorporating cities in GeoNames with a population under 15K.

## C Evaluation Metric Details

### C.1 Coverage

Given Twitter dataset  $D$ , coverage is formally defined in Equation (1)

$$coverage(D) = \frac{|\{t \in D \mid t \text{ is resolved}\}|}{|\{t \in D \mid t \text{ is geotagged}\}|} \quad (1)$$

### C.2 Accuracy

**Match Ratio** Match ratio on level  $L$ , denoted  $mr_L$ , is the number of resolved tweets such that the name matches the ground truth on level  $L$  over the number of resolved tweets that have location information on level  $L$ , where  $L \in \{\text{country, admin, city}\}$ .

$$mr_L(D) = \frac{|\{t \in D' \mid x_L(t) = y_L(t)\}|}{|\{t \in D' \mid y_L(t) \neq \text{null}\}|} \quad (2)$$

**Distance** Using similar notation as Equation (2), let  $x_c(t)$  denote the Carmen resolved geo-coordinates of tweet  $t$  and  $y_c(t)$  denote the ground

truth geo-coordinates of tweet  $t$ . We define the mapping distance of a tweet,  $d(t)$  as the geodesic distance provided in the `geopy` package.<sup>11</sup> The distance accuracy over all tweets in  $D$  is the average of mapping distance of all resolved tweet:

$$d(D) = \frac{1}{|D'|} \sum_{t \in D'} d(t) \quad (3)$$

In addition to the average distance (Equation (3)), we also consider  $Acc@K(D)$ , the ratio of resolved tweets such that the distance error does not exceed  $K$  miles. This metric removes outlier influence possibly present in  $d(D)$ .

$$Acc@K(D) = \frac{|\{t \in D' \mid d(t) \leq K\}|}{|D'|} \quad (4)$$

$Acc@K(D)$  can be easily retrieved from a percentile plot of the mapping distances.

We exclude other commonly used metrics such as classification accuracy (Eisenstein et al., 2010), since it is relatively weak metric because the proposed method uses either 4-way or 49-way classification, much less granular than the entries in Carmen or GeoNames gazetteer. Cheng et al. (2010) use  $Acc@K$  as a ranking metric, which is not applicable to models that only return one prediction, like Carmen.

## D Carmen 2.0 Comparison

**TWITTER-WORLD** This frequently used dataset was collected via the Twitter Streaming API over a span of 5 months (September 21 2011 to February 29 2012). It was filtered to English tweets, non-duplicate tweets, and tweets from users with at least 10 geo-tagged tweets. Locations are assigned on a per-user basis, where the ‘‘ground truth’’ is the city where the majority of a user’s tweets originate. Since Carmen does not require training, we only use the test split of 0.45M tweets.<sup>12</sup>

The coverage and accuracy metrics are shown in Table 5. Before performing ablations, we evaluate all versions of Carmen on TWITTER-WORLD and TWITTER-GLOBAL.

In general, all versions of Carmen perform significantly better on TWITTER-WORLD than TWITTER-GLOBAL. We believe this is

<sup>11</sup><https://geopy.readthedocs.io>

<sup>12</sup>Available for download from author’s website <http://tq010or.github.io/research.html>

Database	Dataset	Coverage	$mr_{country}$	$mr_{admin}$	$mr_{city}$	$d$	Acc@10	Acc@100	Acc@1000
GeoNames-only	TWITTER-WORLD	93.82%	97.42%	73.59%	48.66%	522.6	0.866	0.906	0.907
	TWITTER-GEO-STREAM	45.45%	99.37%	83.87%	32.69%	653.0	0.823	0.867	0.869
GeoNames-combined	TWITTER-WORLD	95.34%	97.73%	56.08%	49.07%	19.2	0.866	0.947	0.999
	TWITTER-GEO-STREAM	45.48%	99.37%	83.30%	32.86%	83.6	0.824	0.902	0.989
Original	TWITTER-WORLD	91.54%	97.45%	49.12%	50.04%	40.3	0.796	0.929	0.995
	TWITTER-GEO-STREAM	39.35%	99.16%	88.75%	32.70%	75.0	0.724	0.890	0.992

Table 5: Ablation over Carmen location database and performance on popular geolocation dataset TWITTER-WORLD and new dataset, TWITTER-GLOBAL. “Acc@ $K$ ” represents the ratio of tweets predicted within  $K$  miles of the ground truth. Higher values are best for all metrics except distance ( $d$ ).

due to the higher availability of metadata in TWITTER-WORLD, since the data is from 2011-2012. This change in metadata availability is discussed more in §5.3 and §6.

Within each dataset, we see a clear trend in GeoNames-combined performing better than GeoNames-only, and Original, with respect to coverage.