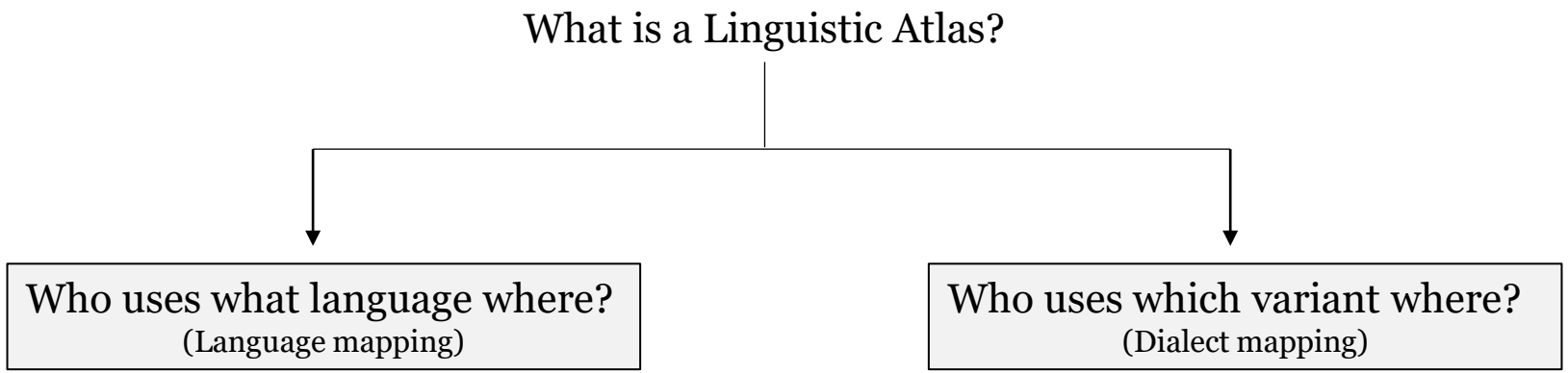

What is a Linguistic Atlas?
Who uses what language where?
(Language mapping)
Who uses which variant where?
(Dialect mapping)

Traditional approaches *ask* people to report what languages they use

What if we could **speed up** and **scale up** this survey process

by directly observing a population's language behaviour?

Machine Learning

- Data Cleaning
- Language Identification
- Grammar Induction
- Dialect Modelling

Machine Learning

- Data Cleaning
- Language Identification
- Grammar Induction
- Dialect Modelling

DATA

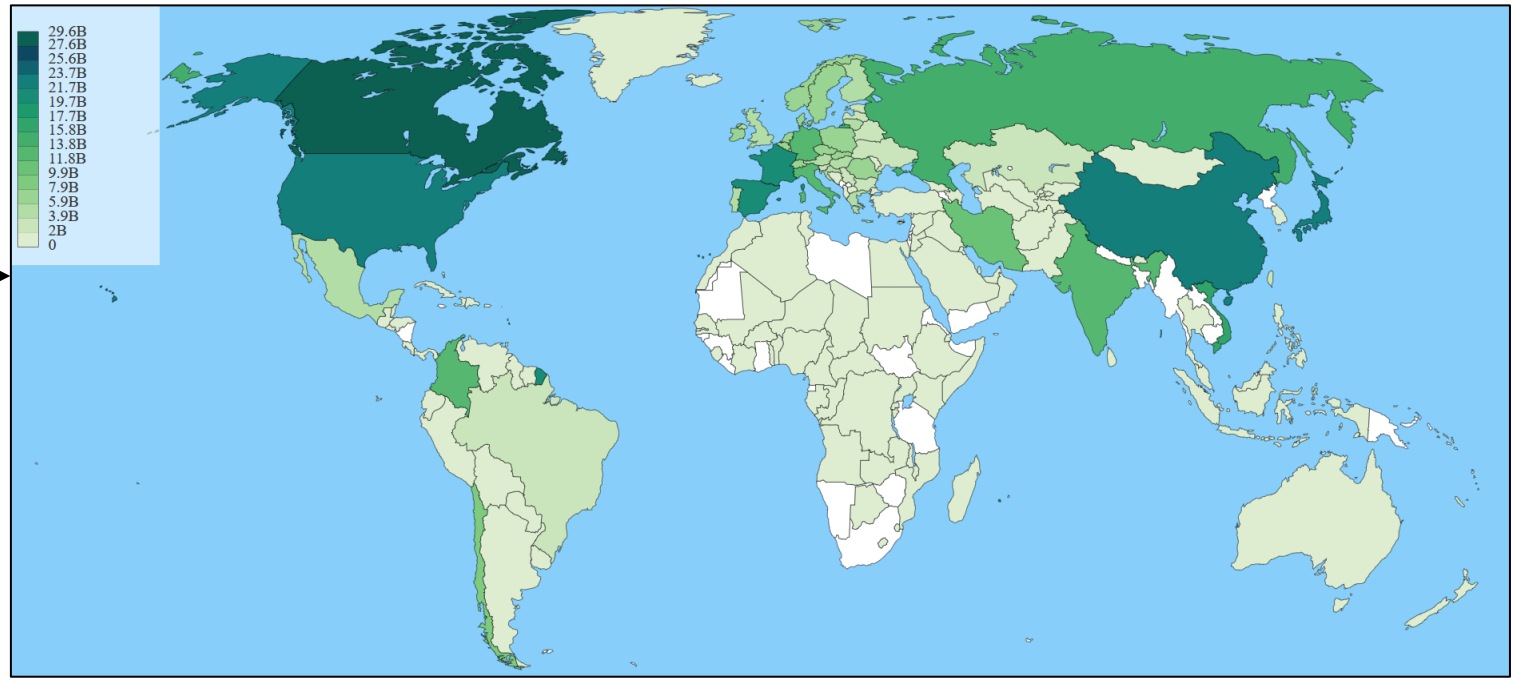
	Twitter Corpus (v 3.4) (Size in Words)	**Corpus of Global Language Use** (v 4.2) (Size in Words)
Africa, North	410,419,871	1,223,532,842
Africa, Southern	275,560,877	26,868,810
Africa, Sub	1,109,821,214	5,938,870,966
America, Brazil	220,184,927	2,265,386,107
America, Central	1,623,884,867	8,877,634,300
America, North	615,704,587	51,921,657,887
America, South	1,522,216,797	22,441,384,853
Asia, Central	439,151,317	17,069,517,255
Asia, East	622,728,293	49,521,933,987
Asia, South	993,107,732	15,147,872,671
Asia, Southeast	801,905,302	21,386,781,131
Europe, East	1,444,388,940	65,413,609,201
Europe, Russia	187,477,833	15,363,644,903
Europe, West	3,167,341,653	143,748,386,801
Middle East	749,192,209	1,721,856,657
Oceania	657,349,100	1,743,571,262
TOTAL	**14.84 billion words**	**423.81 billion words**

Machine Learning

- Data Cleaning
- Language Identification
- Grammar Induction
- Dialect Modelling

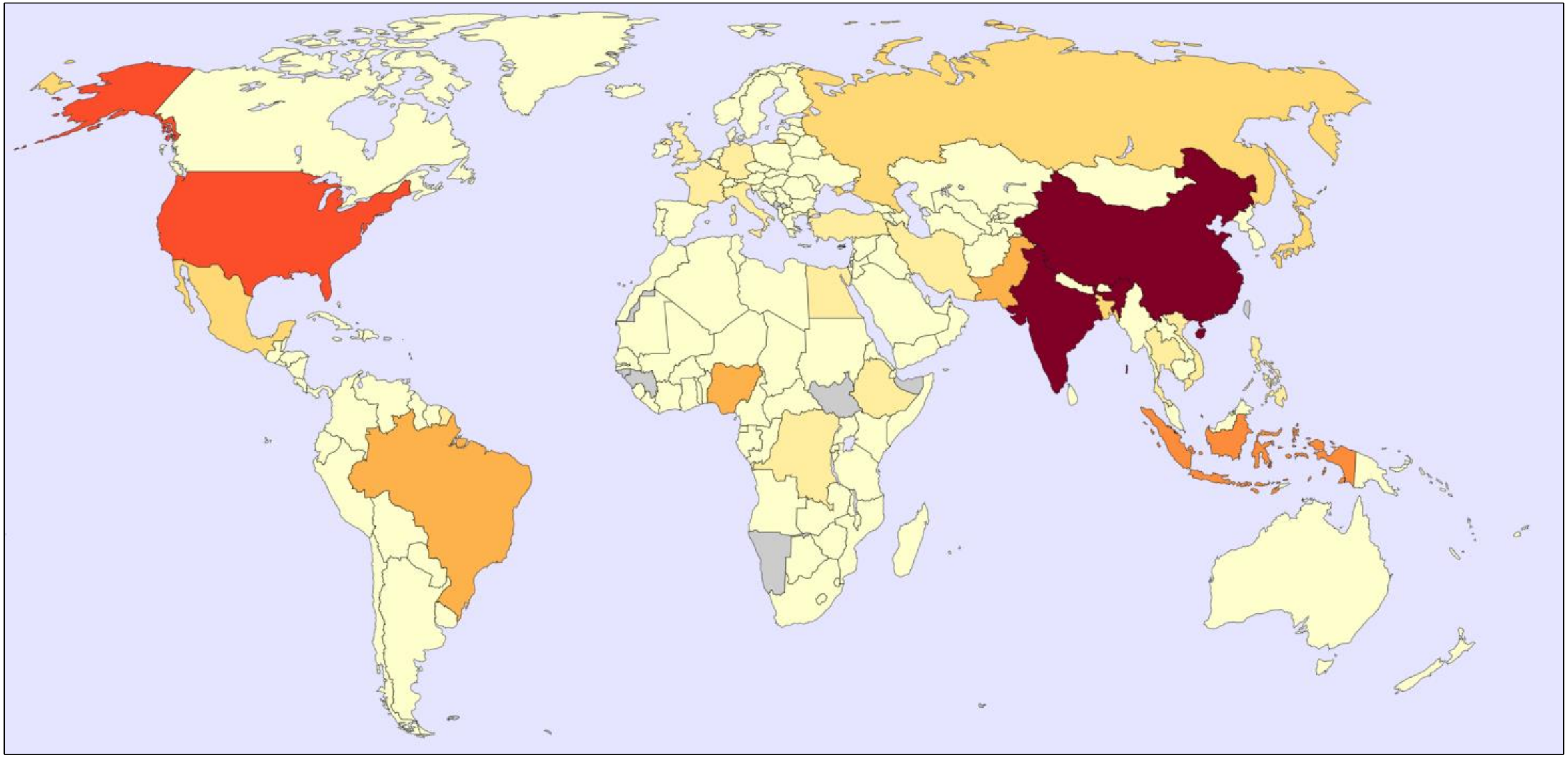
D A T A

Language and Dialect Maps

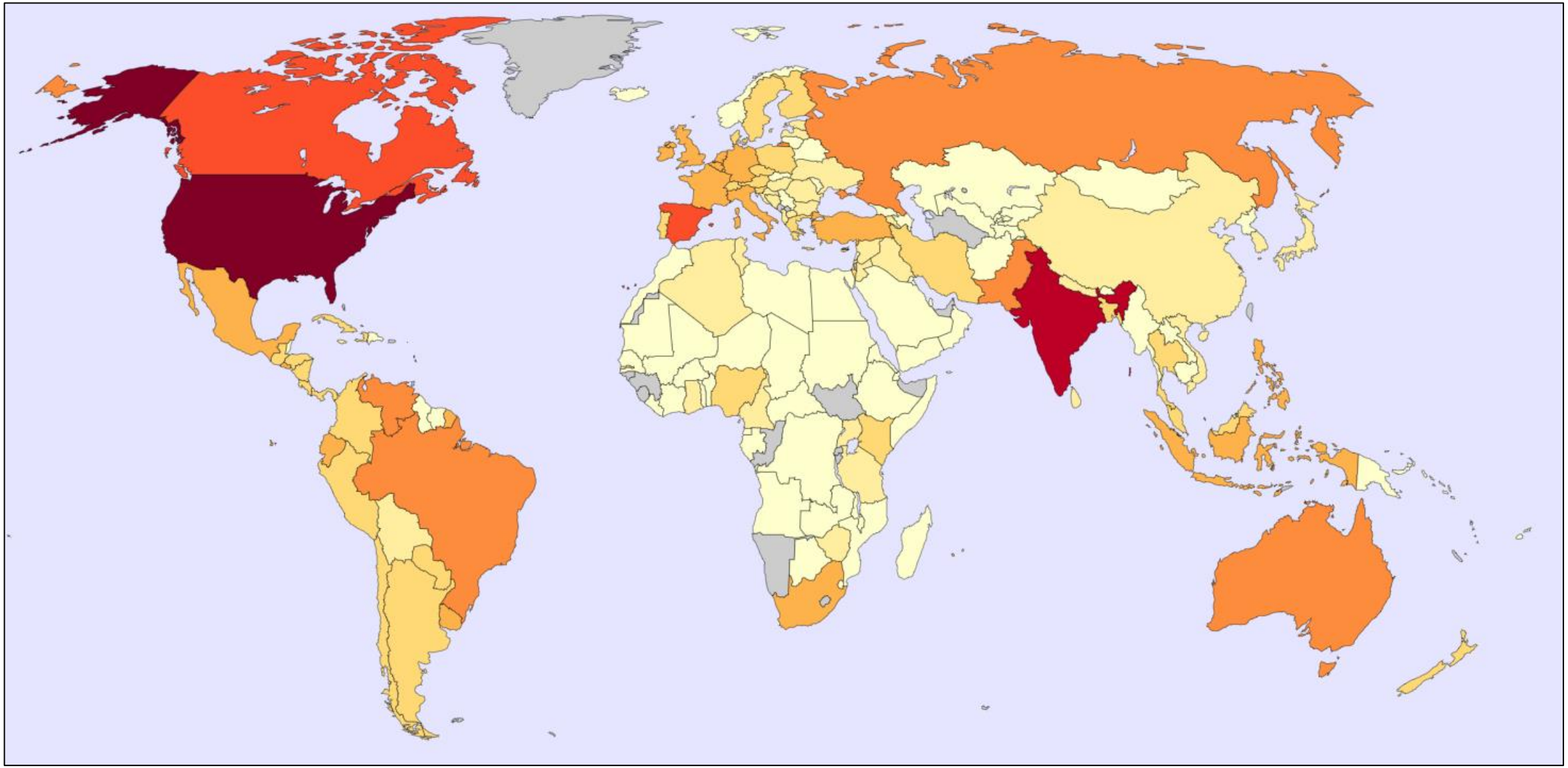


But we also need to systematically measure and adjust demographic bias in the data

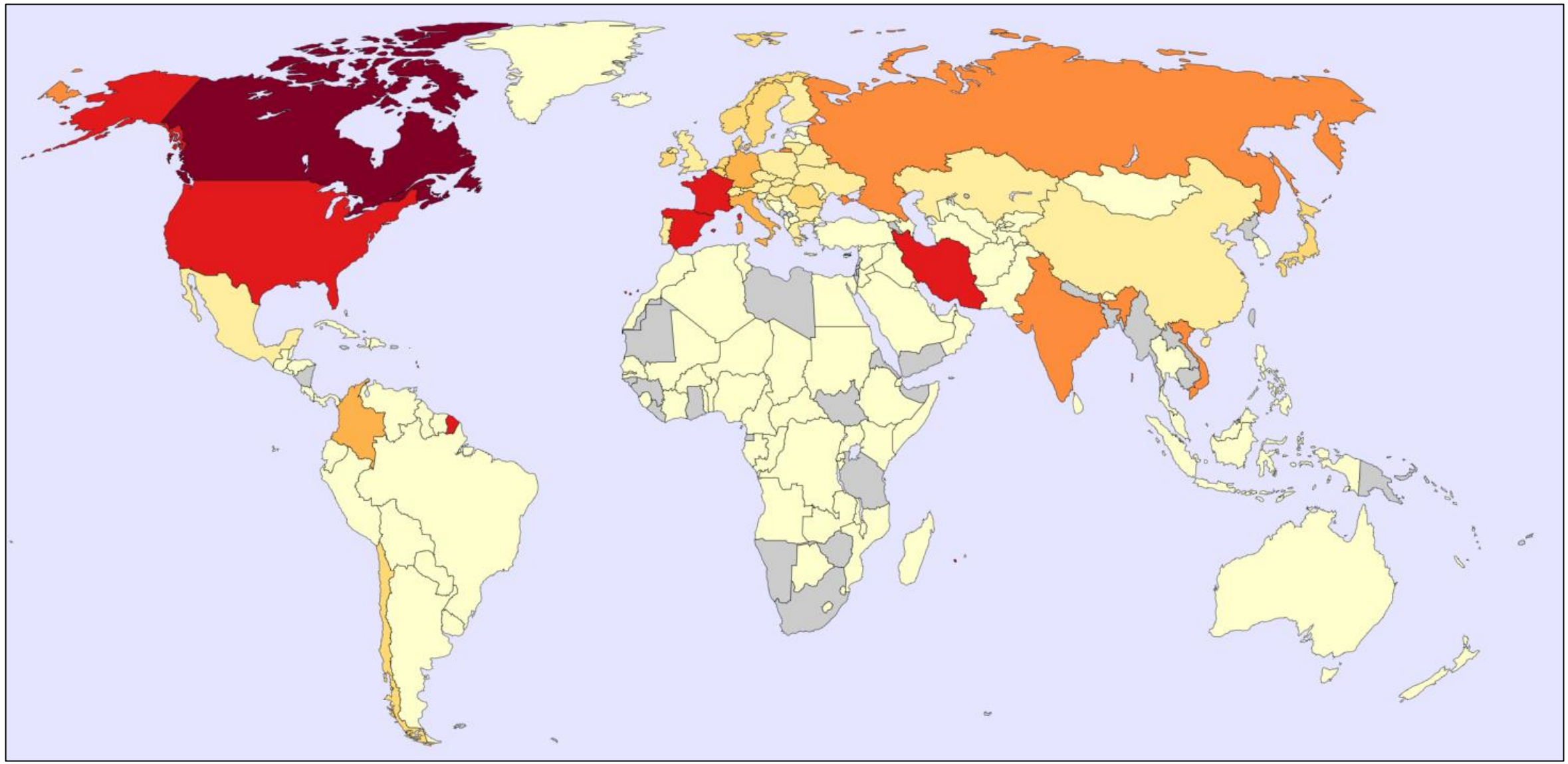
Population: Where do people live?



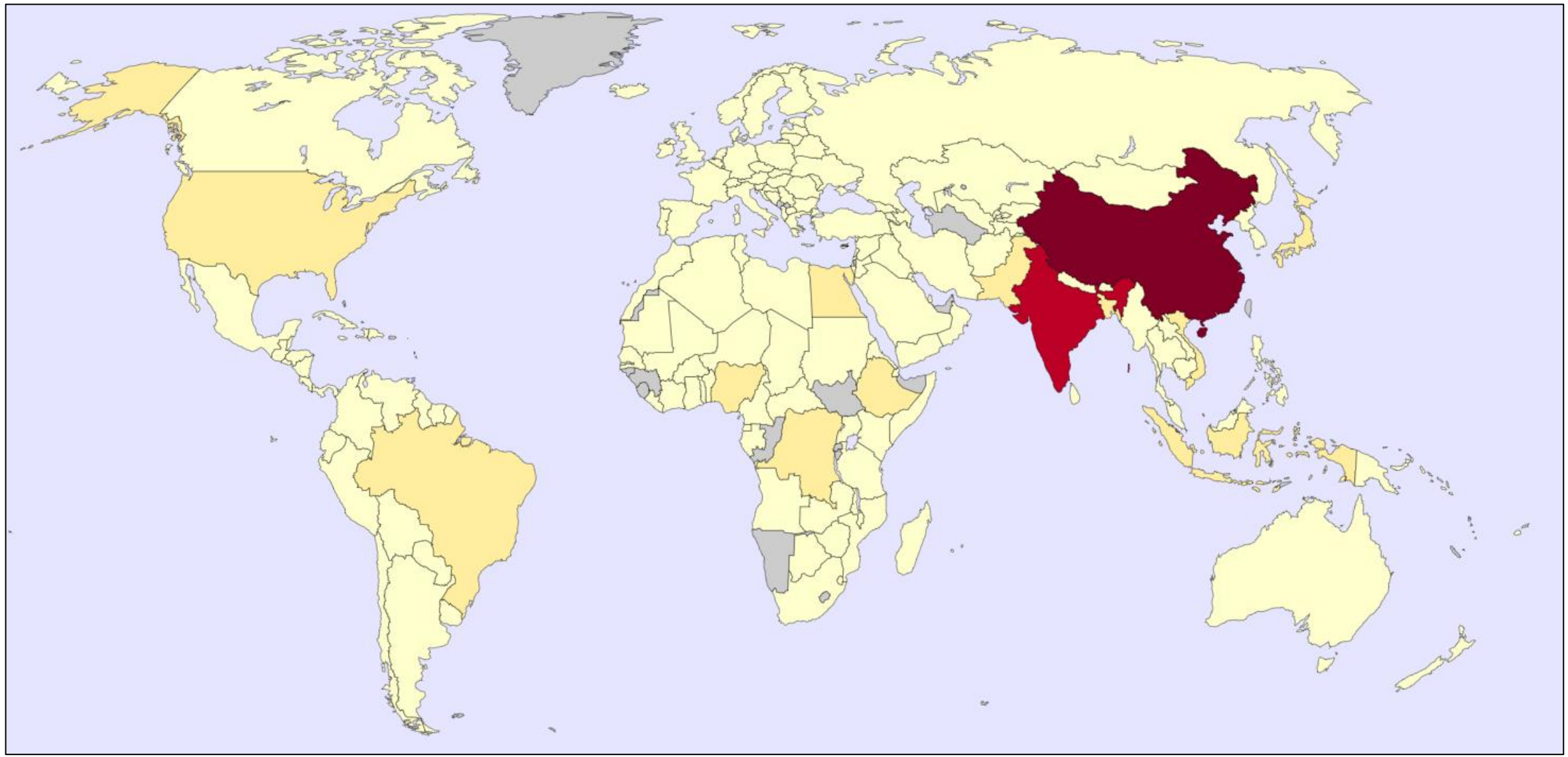
Twitter Density: Where do people Tweet?



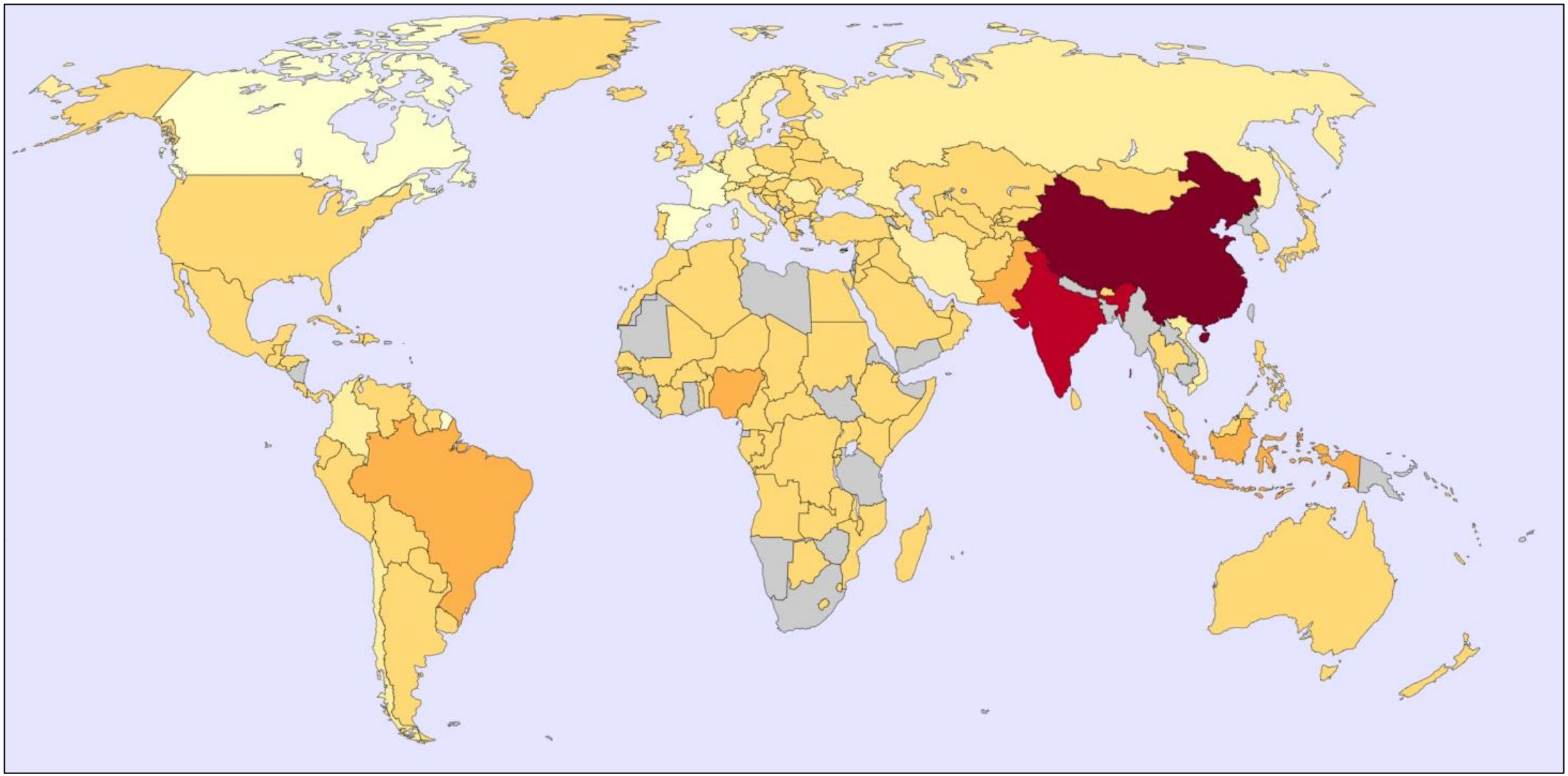
Web Density: Where do web documents come from?



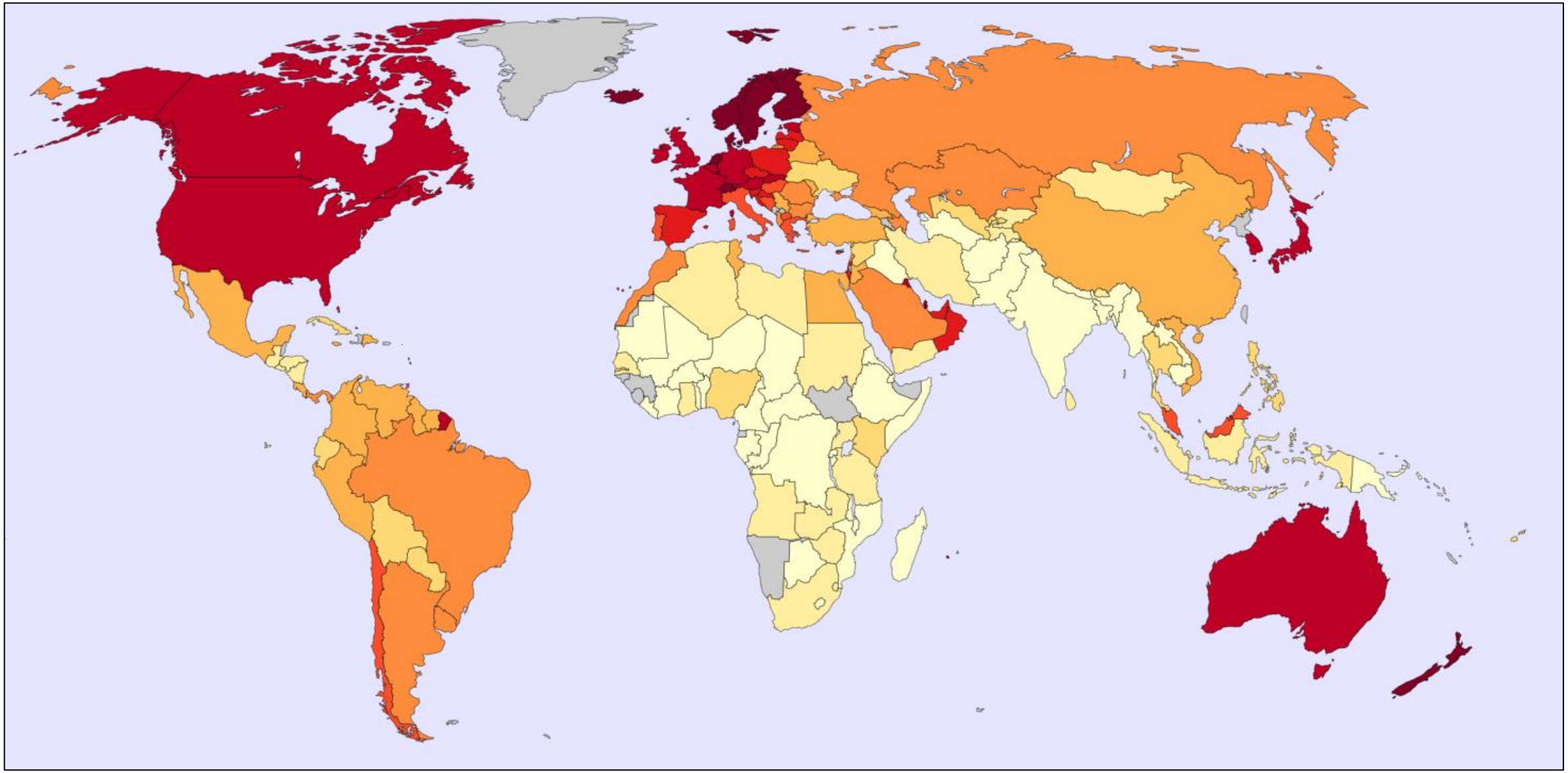
Where is Twitter not used? (i.e., under-represented countries)



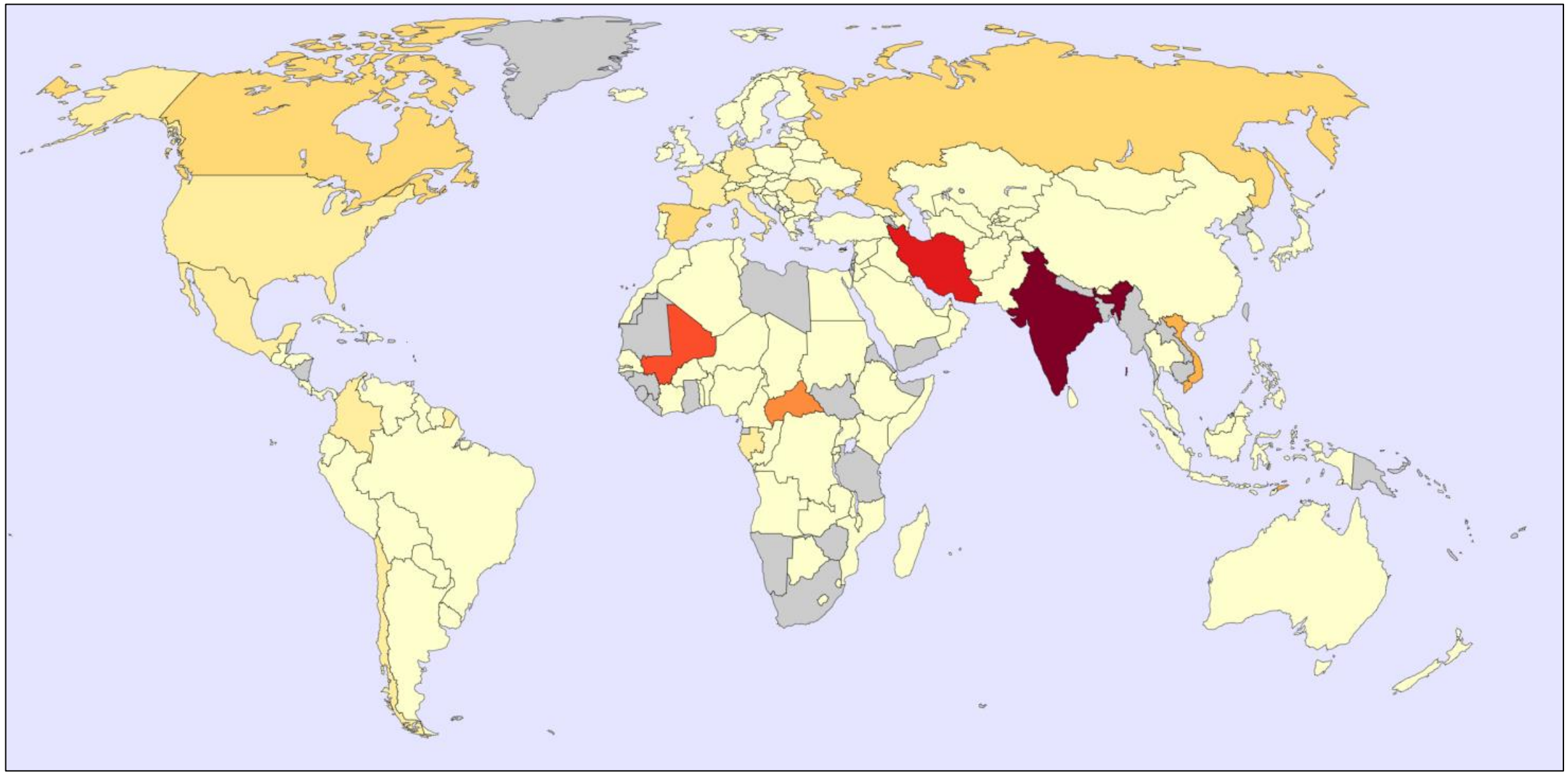
Where are web documents scarce? (i.e., under-represented countries)



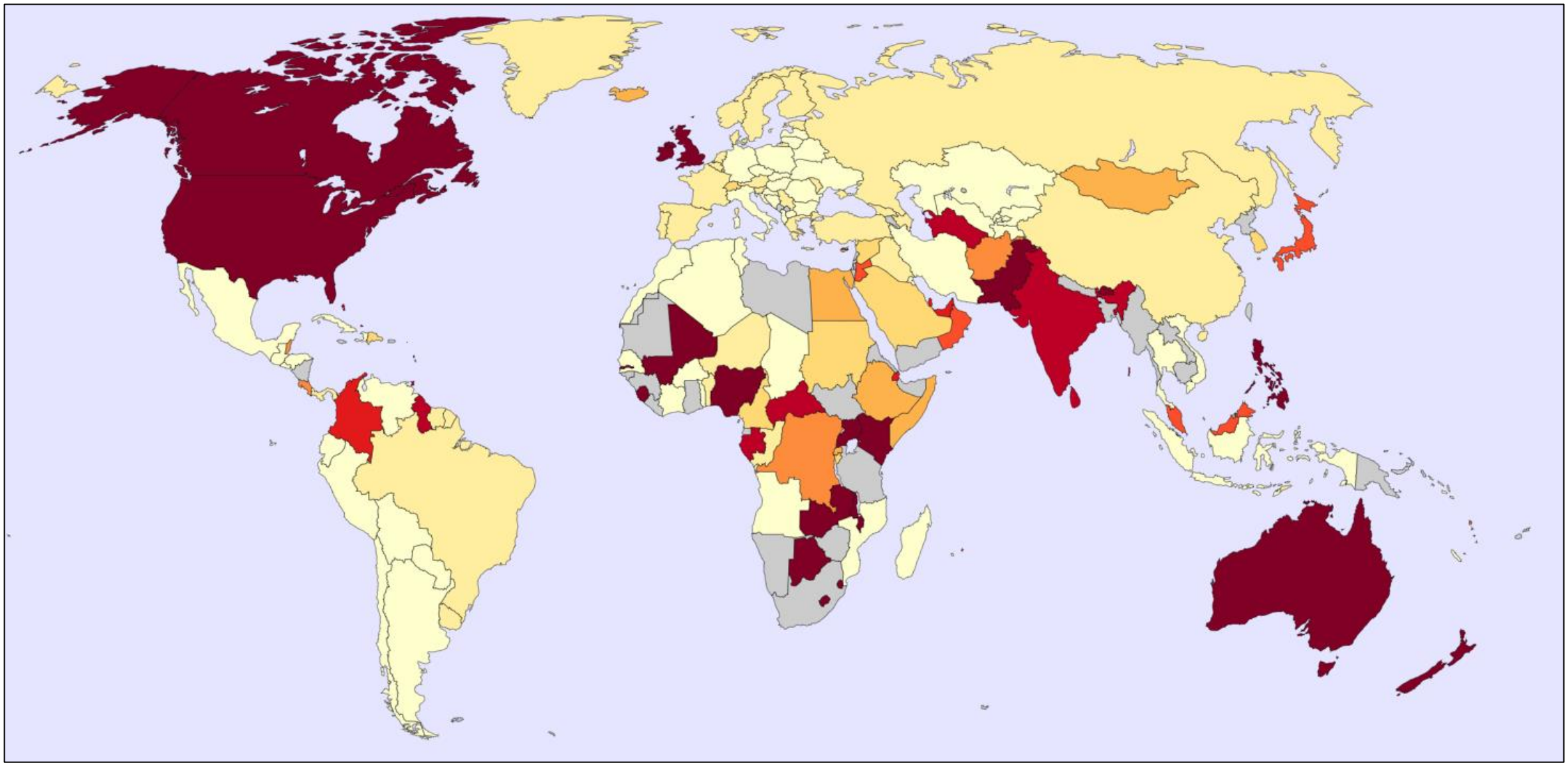
Internet Usage: Does the percentage of the population with internet access influence these datasets?



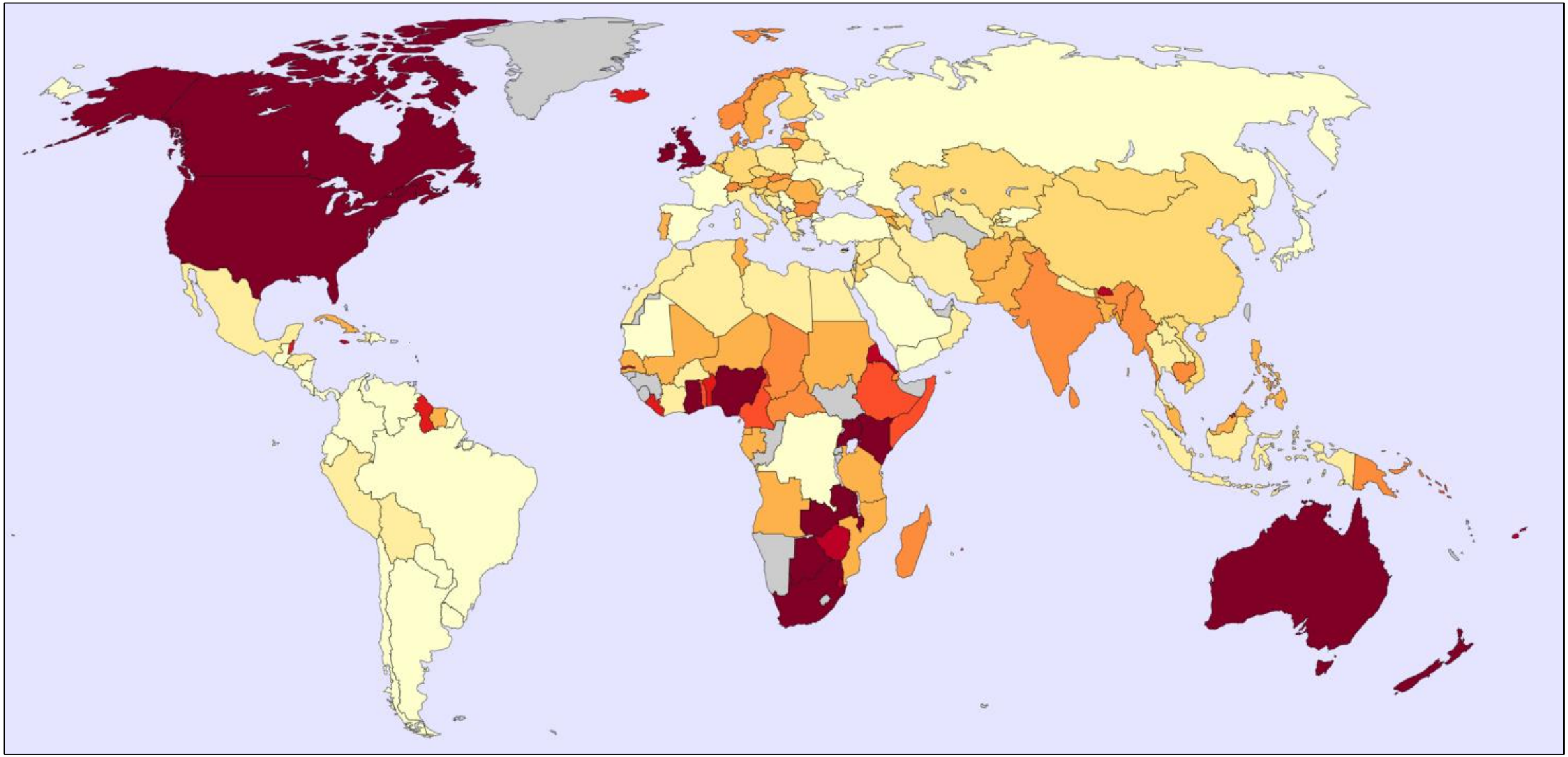
What would the web corpus look like if everyone had internet access?



Countries by their percent usage of English (web): Darker red means more monolingual usage

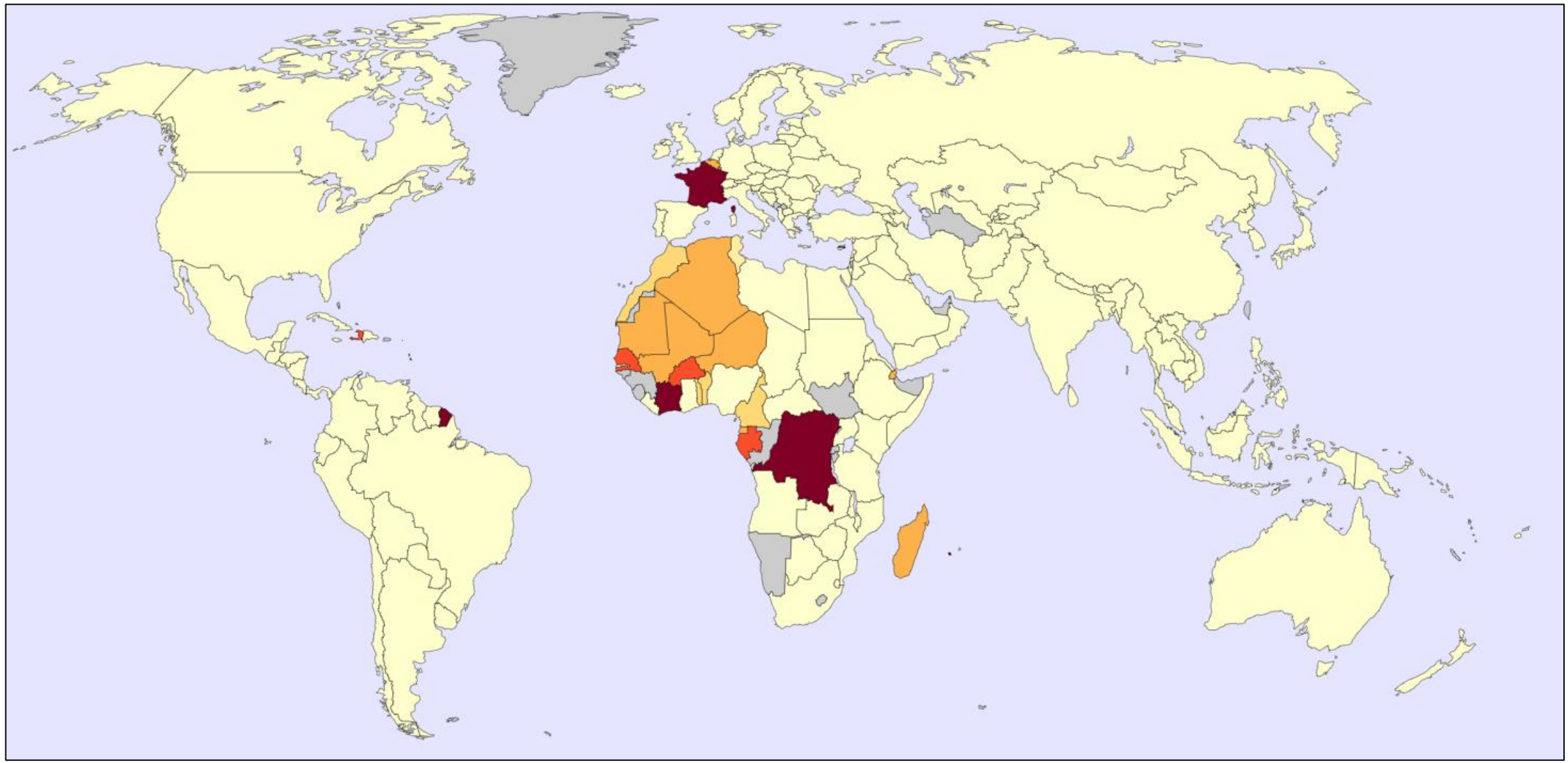


Countries by their percent usage of English (Twitter): Do the datasets capture different populations?

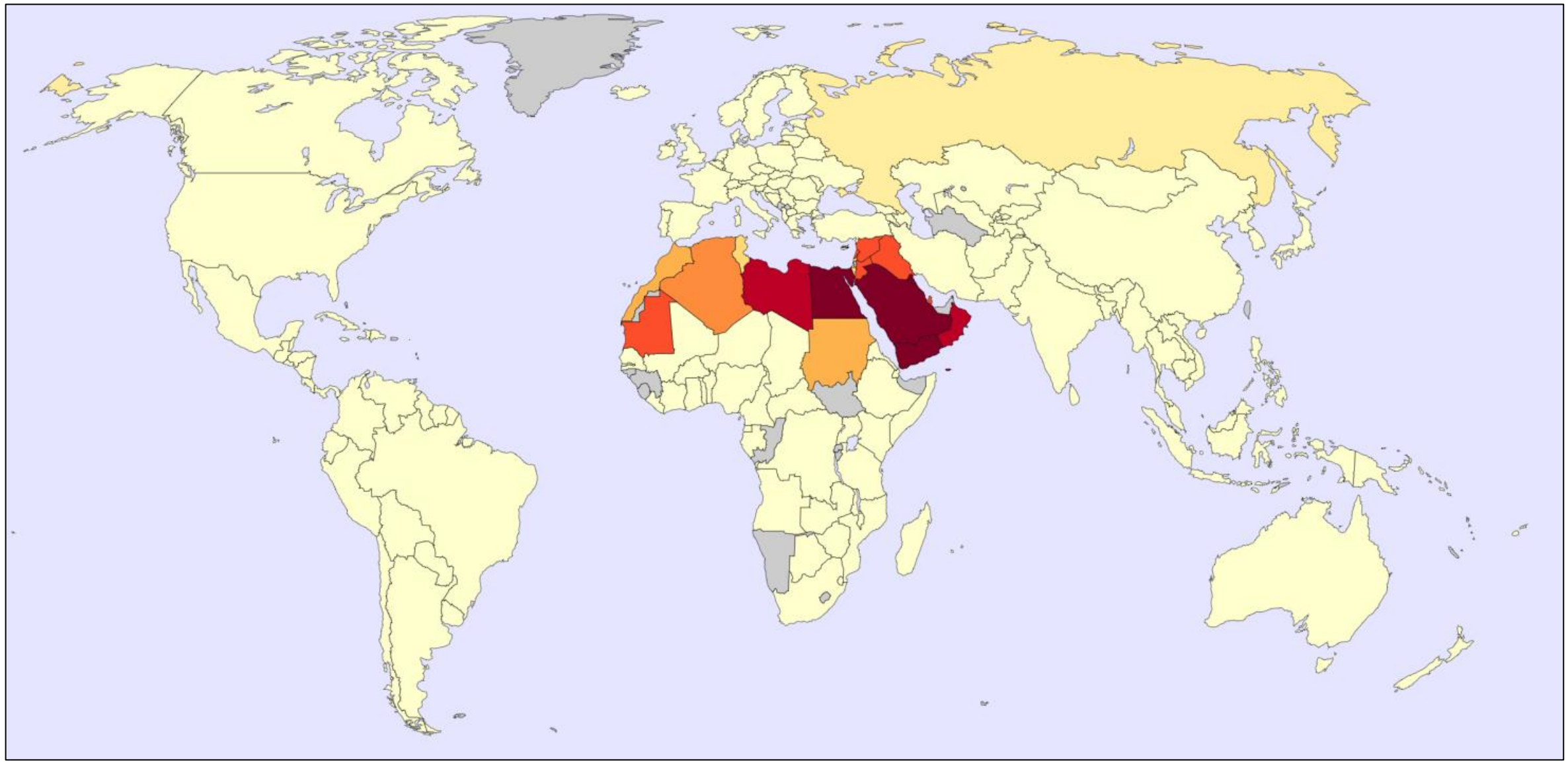


Countries by their percent usage of Spanish (Twitter): English is a global language, but not Spanish

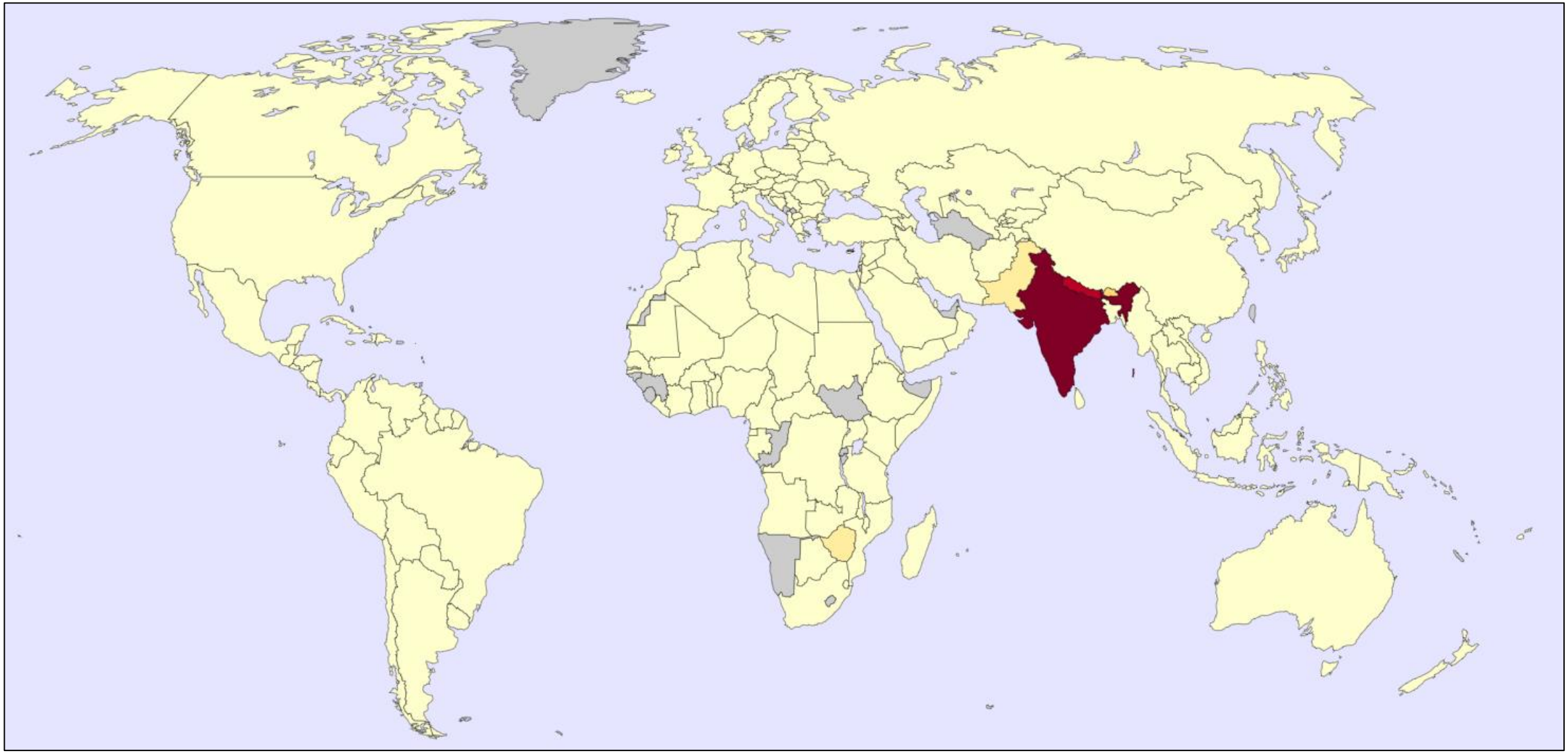
Countries by their percent usage of French (Twitter): French is also widely used



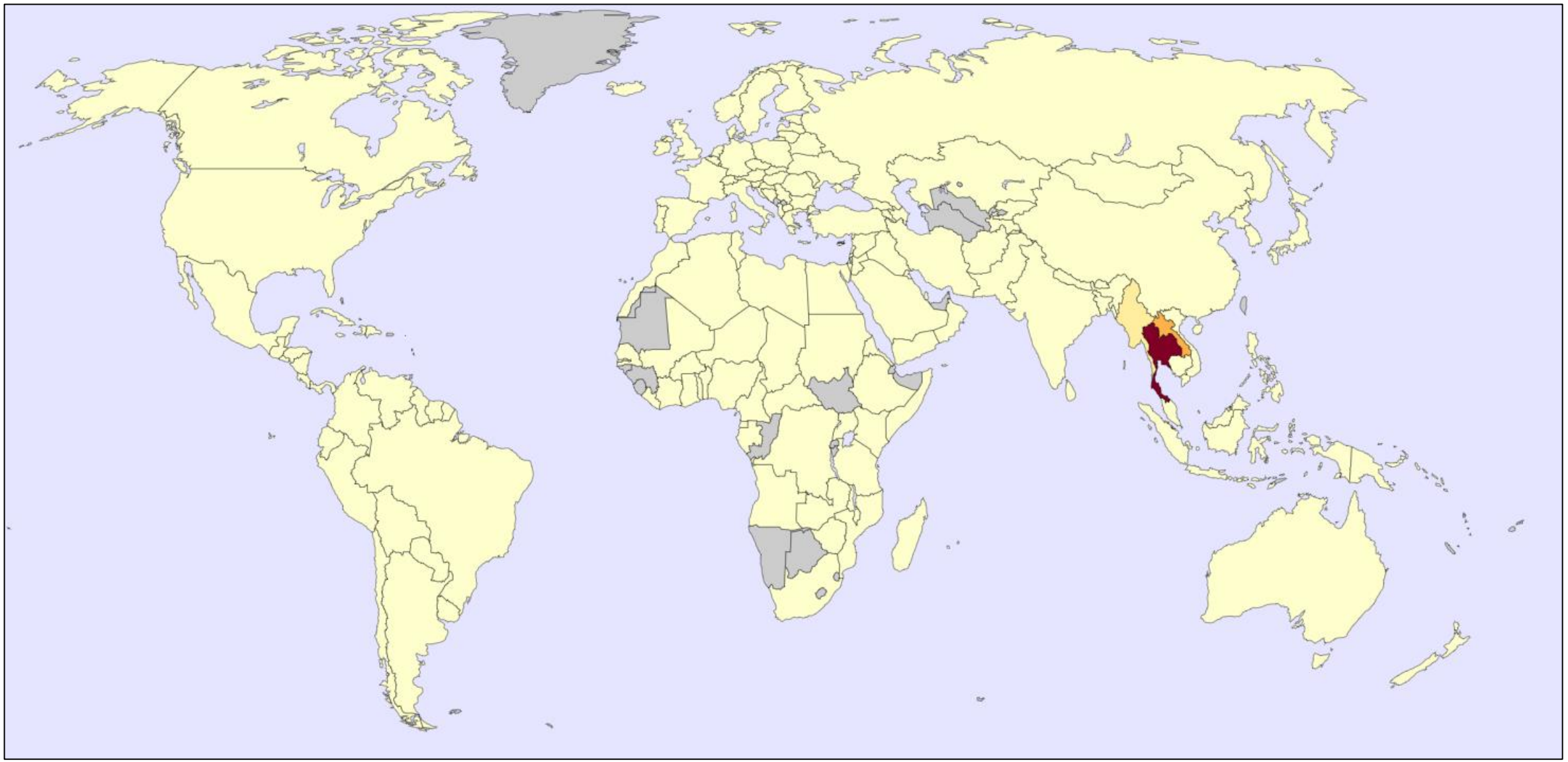
Countries by their percent usage of Arabic (Twitter): Some languages move with along moving populations



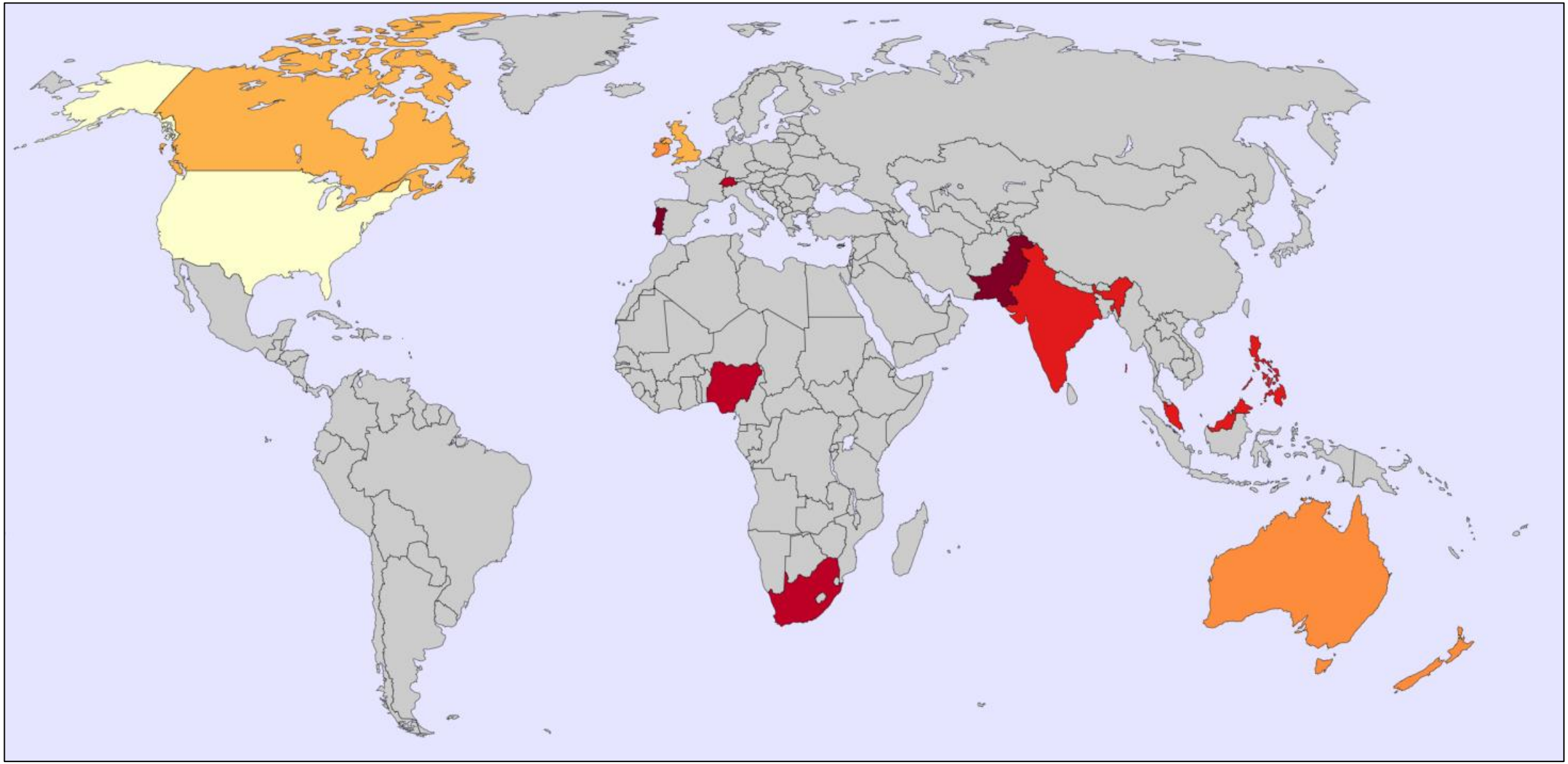
Countries by their percent usage of Hindi (Twitter): But others are restricted geographically



Countries by their percent usage of Thai (Twitter): Here, Thai is well-represented... but only in a narrow region

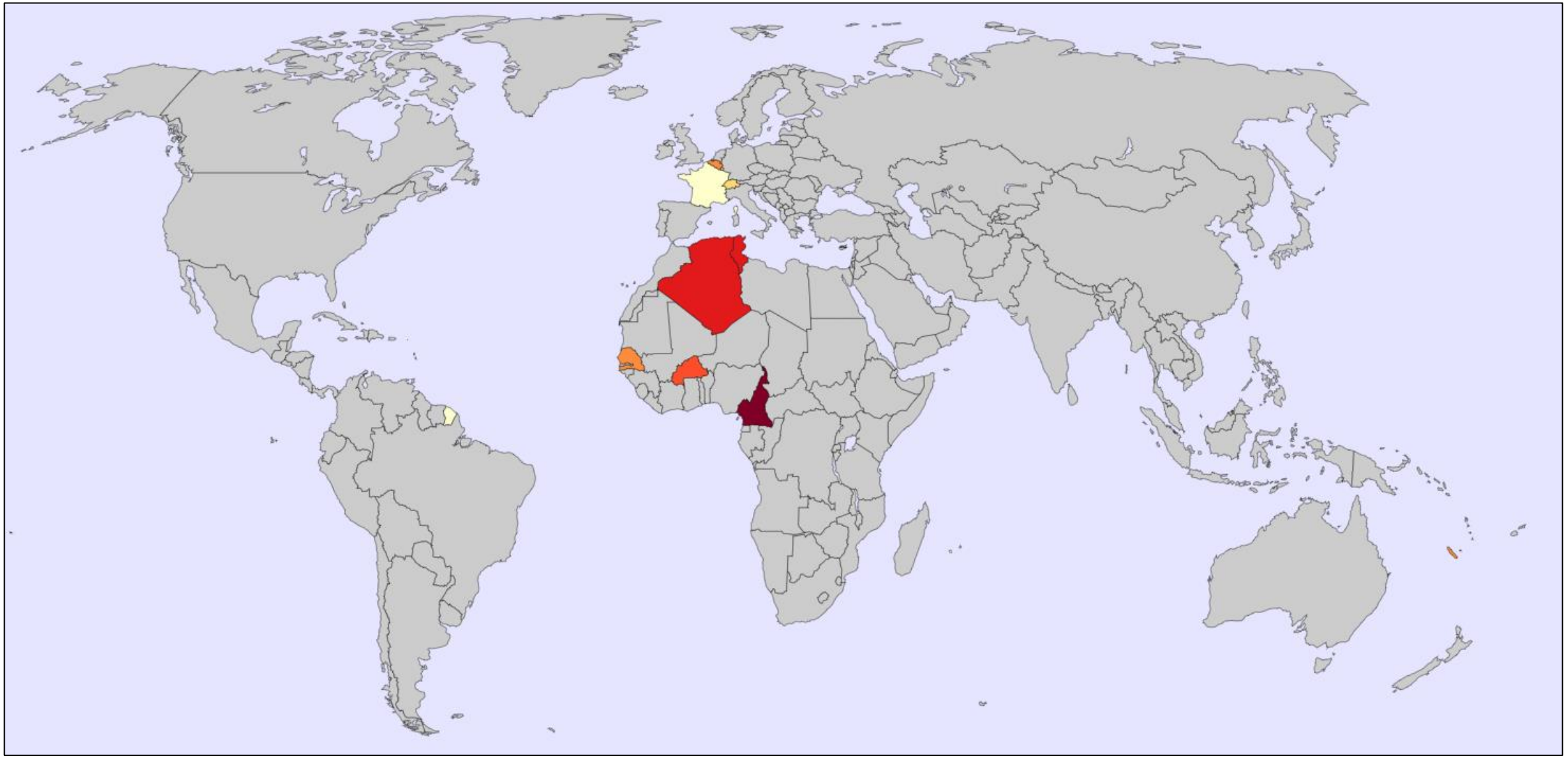


Uniqueness of English Dialects (Web): We can go beyond surveys by modelling the data from each country



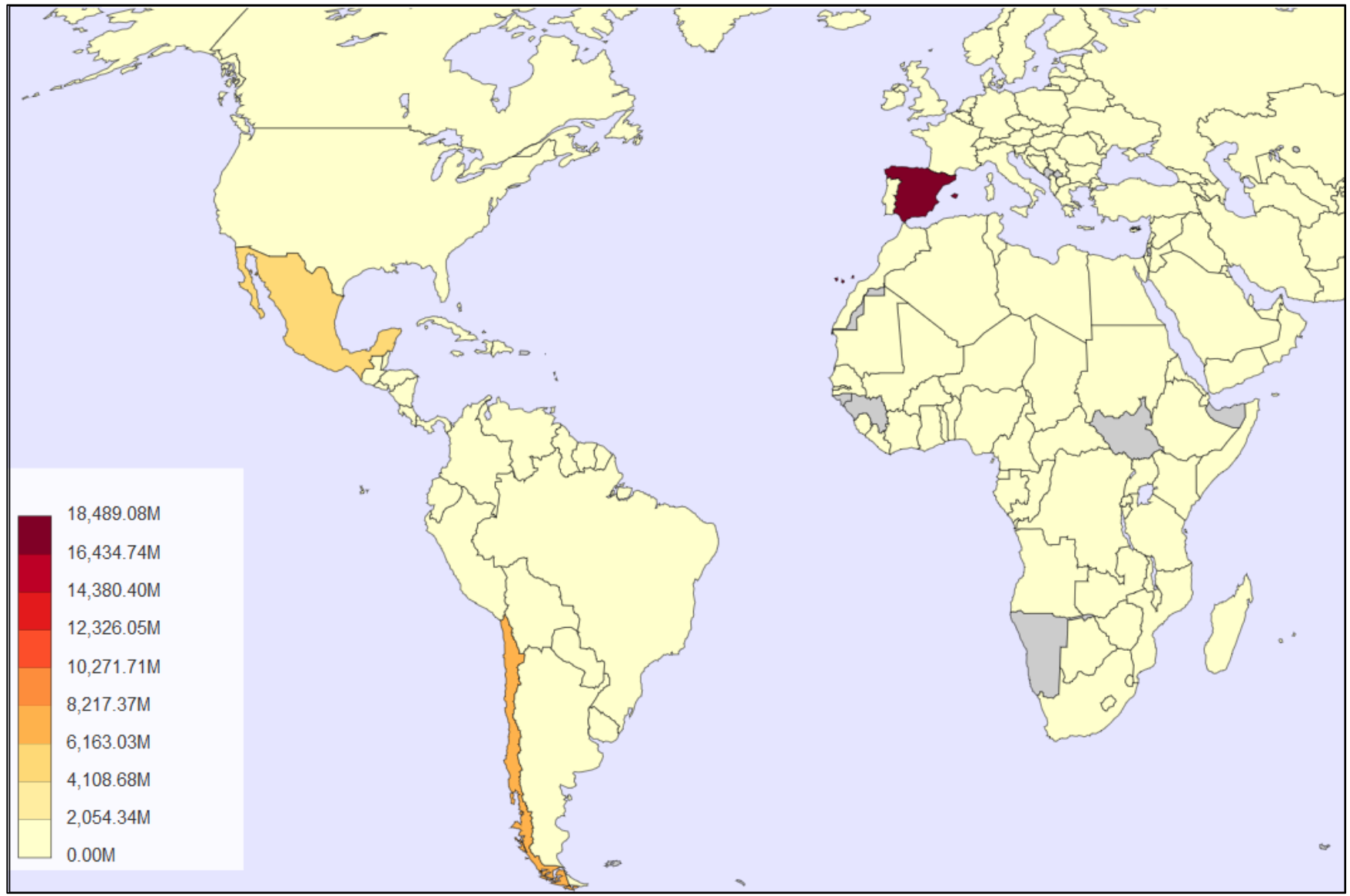
Uniqueness of Spanish Dialects (Web): The dialect model shows the Spain is the central variety of Spanish

Uniqueness of French Dialects (Web): There aren't that many country-level dialects of French

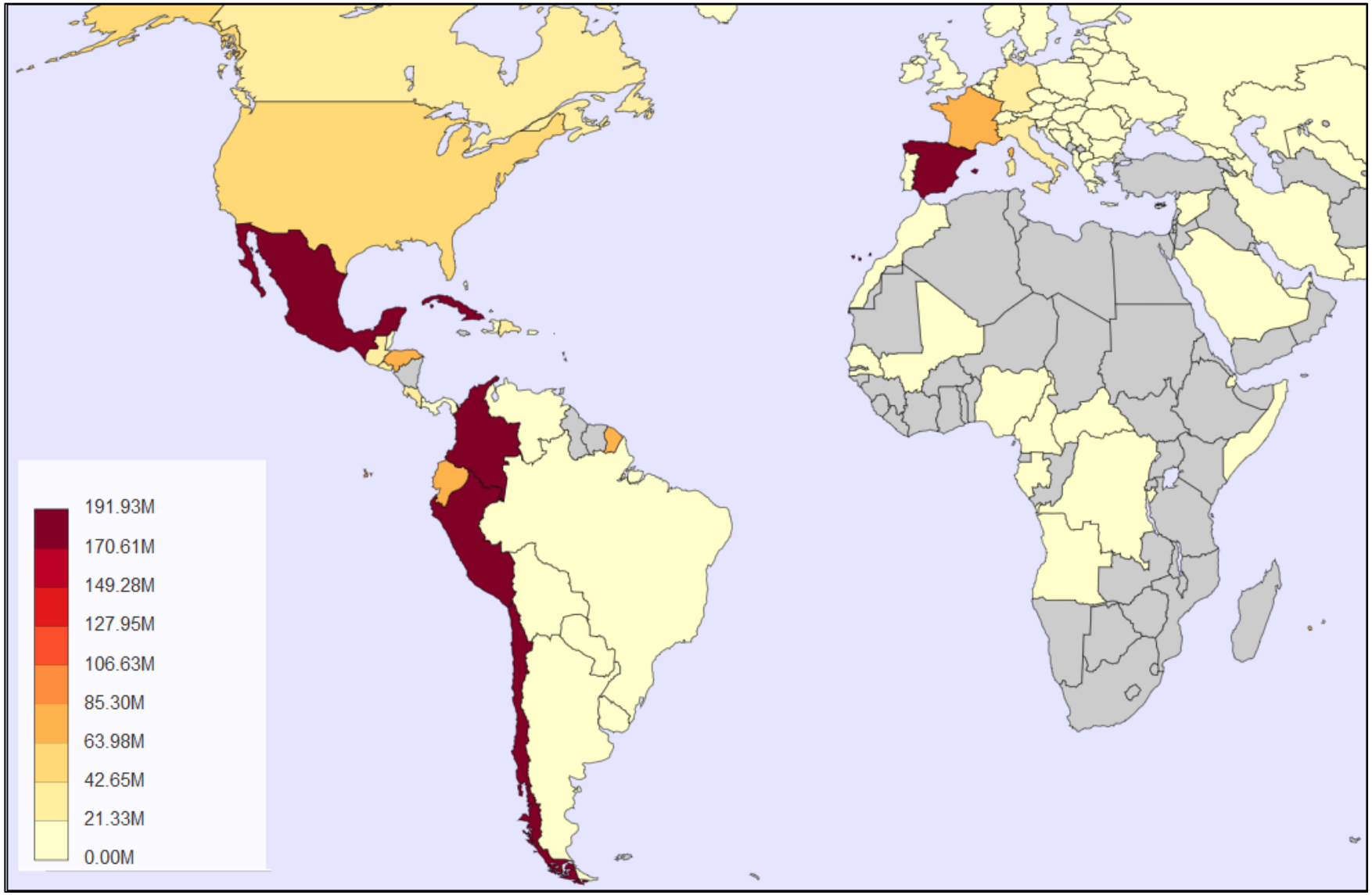


Uniqueness of Russian Dialects (Web): Unlike other major languages, Russian is restricted to a (large) contiguous region

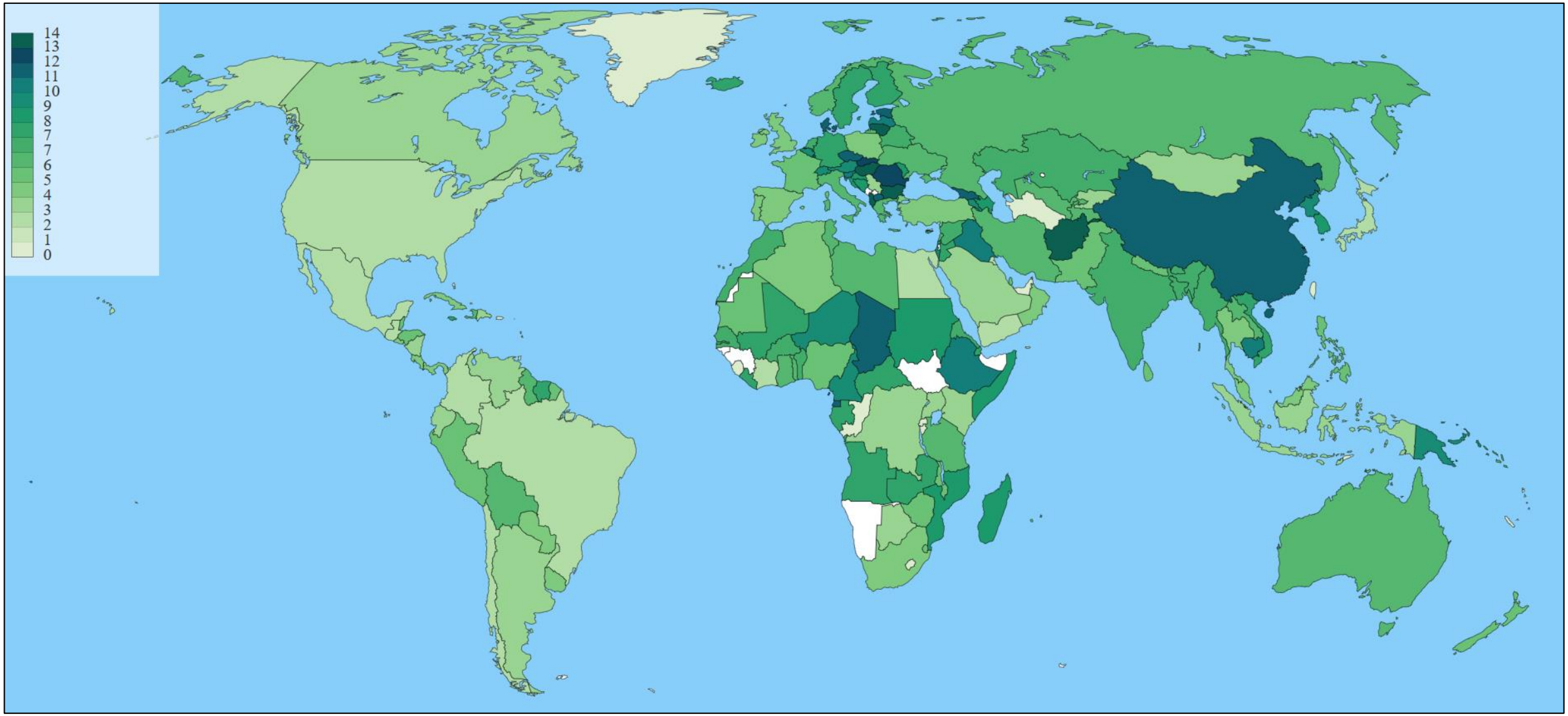
Removing demographic bias (Web): Use of Spanish by country without removing bias



Removing demographic bias (Web): Use of Spanish with population-based sampling (GeoWAC)



Linguistic Diversity (Twitter): Which countries have a more diverse linguistic landscape? (Darker means more languages)



Explore the data at earthLings.io