**Subnational administration shapefile matching: process documentation**

**Step 1: Cleaning data and initial data quality check**

We clean the datasets by replacing diacritics and removing capitalization.

Numerator/Denominator at admin 2 level add up to numerator/denominator at admin 1 level.

**Step 2: Comparing number of administrative units between JRF data and WHO/Polio shapefile**

We count the number of the most granular administrative units reported with the JRF for each country. The reported administrative unit counts are compared to corresponding administrative unit counts of WHO/Polio shapefile subset to the year in which the JRF data was collected. If they are different, we look at the surrounding years in the WHO shapefile (e.g. year before, year after, if available) to see if that improves the match. If they are still different, we record the difference and flag for WHO.

*Optional step:*

If the WHO/Polio shapefile unit counts comparison is different, The reported administrative unit counts are compared to corresponding administrative unit counts of every available IHME shapefiles library. The results are documented and flagged for WHO.

**Step 3: Matching administrative name between JRF data and WHO/Polio shapefile**

For countries where the unit counts are perfectly match, we match both the first- and second-level administrative unit names from the administrative data to WHO/Polio shapefile using WHO synonym library. All unsuccessful matchings/unmatched polygons are documented and flagged for WHO

*Optional step:*

If there is any difference in administrative unit names, we then use a fuzzy matching algorithm from the stringdist package in R. The stringdist package has multiple string distance methods to compare the difference between two string sequence:

|  |  |
| --- | --- |
| **Method** | **Description** |
| osa | Optimal string alignment, (restricted Damerau-Levenshtein distance) |
| lv | Levenshtein distance |
| dl | Full Damerau-Levenshtein distance |
| hamming | Hamming distance (string a and string b must have same number of character) |
| lcs | Longest common substring distance |
| qgram | 1-gram distance |
| cosine | Cosine distance between q-gram profiles |
| jaccard | Jaccard distance between q-gram profiles |
| jw | Jaro, or Jaro-Winker distance |

We try all matching methods listed above for each country and choose the one with highest number of matched administrative units, without additional units reported in the shapefile. If there is a tie between two or more matching methods, we investigate further using manual matching validation (below) to determine the better method.

All the matches derived from the fuzzy matching algorithm are manually verified using external sources (such as online mapping platforms, country GIS data, news articles, or other publicly-available resources describing the evolution of subnational administrative divisions over time). All the matchings using stringdist package are documented and flag for WHO

**Step 4: Updating synonym library**

Any differences in administrative unit names are validated by WHO and added to a synonym library which was used to facilitate matching of subsequent years of data.

**Subnational administrative data shapefile matching: flowchart**

