



SALUD
SECRETARÍA DE SALUD



Instituto Nacional
de Salud Pública



CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES
CIENCIAS Y TECNOLOGÍAS



CentroGeo
Centro de Investigación en
Ciencias de Información Geoespacial, A.C.

Geocodificación de los datos de ESMaestras : Un enfoque basado en técnicas de Ciencias de Información Geoespacial

Alejandro Molina Villegas
CONAHCYT - CentroGeo
amolina@centrogeo.edu.mx

20 de Septiembre de 2023

Resumen

El **objetivo** de la investigación es obtener las **coordenadas** geográficas de los datos del Estudio de la Salud de las Maestras (BD versión anonimizada); a este proceso se le conoce como geocodificación*.

En un primer **diagnóstico** encontramos los datos muy **inconsistentes** para cubrir el objetivo al cien por ciento.

Sin embargo, gracias a la aplicación de diferentes **algoritmos** logramos obtener **coordenadas confiables para un volumen grande de datos** (66,887 registros georreferenciados con alta precisión).

* esri. (s.f.). [Geocoding]. En el *GIS Dictionary - Technical Support*. Recuperado el 20 de septiembre, 2023, en <https://support.esri.com/en-us/gis-dictionary/geocoding>

Resumen de la Problemática

- Georreferenciación informal*

Los datos proporcionados son textos con información del domicilio de las personas pero esto no son datos geográficos que pueda usarse directamente en aplicaciones de ciencia o tecnología.

- Big Data

104,003 entradas resultando inviable procesar manualmente la asignación de coordenadas.

- Inconsistencias

Llenado de encuestas con muchas inconsistencias (pej. DF, CDMX, CD MEXICO, CD DE MEXICO, DISTRITO FEDERAL...) originando 221 variantes de estados, 2902 variantes municipios, 1597 CPs inválidos, incontables nombres de calles y colonias alterados.

Resumen Metodológico

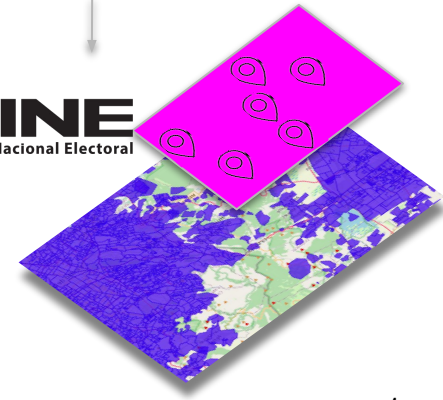
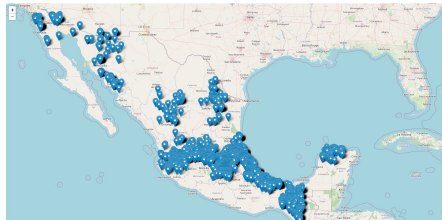
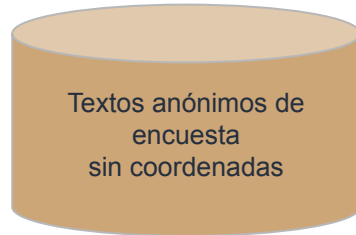
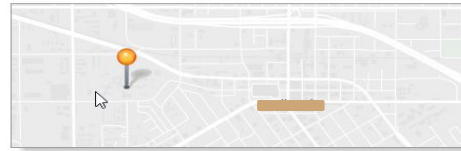


ArcGIS REST APIs

Av. El rosario 238, Santa Ana,
Tuxtla Gutiérrez, Chiapas,
29090

el rosario 238, Tuxtla Gutiérrez,
Chiapas

"x": -117.195665842, "y": 34.0564907277

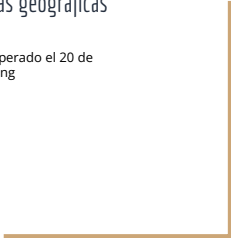




Geocodificar

Es el proceso que convierte direcciones en coordenadas geográficas

Fuente: esri. (s.f.). [Geocoding]. En el *GIS Dictionary - Technical Support*. Recuperado el 20 de septiembre, 2023, en <https://support.esri.com/en-us/gis-dictionary/geocoding>



Geocodificación Remota Asíncrona

Usamos los campos proporcionados para realizar consultas remotas a un servicio Web de la API de ArcGIS especializada en Geocodificación de direcciones. Para optimizar el tiempo de procesamiento usamos 105 máquinas virtuales de Amazon AWS que hacen mil queries cada una de manera asíncrona.



Geocodificación Remota Asíncrona (651,668 candidatos)

AV. DEL ROSARIO 2293 SANTA ANA Tuxtla Gutierrez 29090 CHIAPAS

[

```
{'address': 'Avenida El Rosario 2293, Caminera, Tuxtla Gutiérrez, Chiapas, 29090', 'location': {'x': -93.099509986629, 'y': 16.740710016685}, 'extent': {'xmin': -93.100509986629, 'ymin': 16.739710016685, 'xmax': -93.098509986629, 'ymax': 16.741710016685}}
```

```
{'address': 'Avenida Rosario 2293, Miravalle, Tuxtla Gutiérrez, Chiapas, 29039', 'location': {'x': -93.129425381397, 'y': 16.771478804309}, 'extent': {'xmin': -93.130425381397, 'ymin': 16.770478804309, 'xmax': -93.128425381397, 'ymax': 16.772478804309}}
```

```
{'address': 'Avenida El Rosario, Santa Ana, Tuxtla Gutiérrez, Chiapas, 29090', 'location': {'x': -93.098146514924, 'y': 16.740221469028}, 'extent': {'xmin': -93.099146514924, 'ymin': 16.739221469028, 'xmax': -93.097146514924, 'ymax': 16.741221469028}}
```

```
{'address': '29090, Santa Ana, Tuxtla Gutiérrez, Chiapas', 'location': {'x': -93.097572819099, 'y': 16.739326771929}, 'extent': {'xmin': -93.102572819099, 'ymin': 16.734326771929, 'xmax': -93.092572819099, 'ymax': 16.744326771929}}
```

]



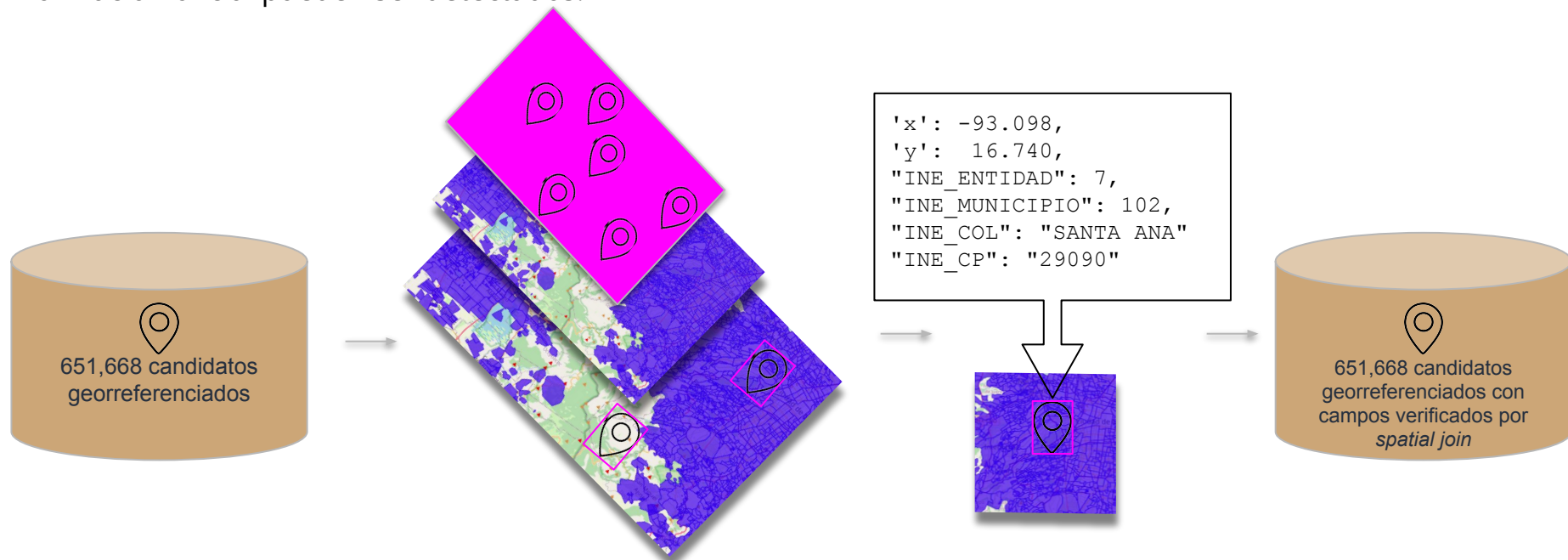
Verificación geoespacial

Comprobar o examinar las coordenadas obtenidas con fuentes oficiales



Verificación Geoespacial con Datos Oficiales

Usando datos oficiales de México* proyectamos los puntos de los candidatos sobre los polígonos de colonias mediante la operación de *spatial join* con lo cual se puede validar la coincidencia entre direcciones y puntos con los polígonos oficiales a nivel colonia, CP, municipio y entidad. Análogamente, los puntos sin congruencia con la información oficial pueden ser detectados.



Verificación Geoespacial con Datos Oficiales

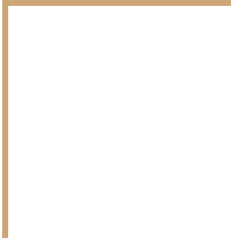
AV. DEL ROSARIO 2293 SANTA ANA Tuxtla Gutierrez 29090 CHIAPAS

```
[
  {'address': 'Avenida El Rosario 2293, Caminera, Tuxtla Gutiérrez, Chiapas, 29090',
'location': {'x': -93.099509986629, 'y': 16.740710016685}, 'extent': {'xmin': -93.100509986629,
'ymin': 16.739710016685, 'xmax': -93.098509986629, 'ymax': 16.741710016685}} U {"INE_ENTIDAD": 7,
"INE_MUNICIPIO": 102, "INE_COL": "CAMINERA", "INE_CP": "29090", "areakm2": 0.047}

  {'address': 'Avenida Rosario 2293, Miravalle, Tuxtla Gutiérrez, Chiapas, 29039', 'location':
{'x': -93.129425381397, 'y': 16.771478804309}, 'extent': {'xmin': -93.130425381397, 'ymin':
16.770478804309, 'xmax': -93.128425381397, 'ymax': 16.772478804309}} U {"INE_ENTIDAD": 7,
"INE_MUNICIPIO": 102, "INE_COL": "MIRAVALLE", "INE_CP": "29039", "areakm2": 0.047}

  {'address': 'Avenida El Rosario, Santa Ana, Tuxtla Gutiérrez, Chiapas, 29090', 'location':
{'x': -93.098146514924, 'y': 16.740221469028}, 'extent': {'xmin': -93.099146514924, 'ymin':
16.739221469028, 'xmax': -93.097146514924, 'ymax': 16.741221469028}} U {"INE_ENTIDAD": 7,
"INE_MUNICIPIO": 102, "INE_COL": "SANTA ANA", "INE_CP": "29090", "areakm2": 0.047}

  {'address': '29090, Santa Ana, Tuxtla Gutiérrez, Chiapas', 'location': {'x':
-93.097572819099, 'y': 16.739326771929}, 'extent': {'xmin': -93.102572819099, 'ymin':
16.734326771929, 'xmax': -93.092572819099, 'ymax': 16.744326771929}} U {"INE_ENTIDAD": 7,
"INE_MUNICIPIO": 102, "INE_COL": "SANTA ANA", "INE_CP": "29090", "areakm2": 1.183}
]
```



Alineamiento de Secuencias

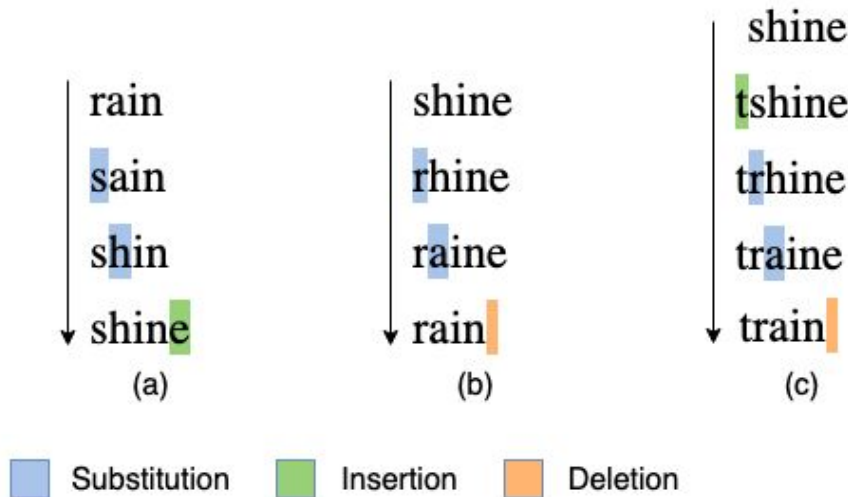
Comparar direcciones con respecto a su escritura



Algoritmos de Alineamiento de Secuencias

Para comparar la similitud entre direcciones usamos algoritmos de alineamiento de secuencias (*String Alignment*), los cuales son muy utilizados en genética y sistemas de voz y texto*.

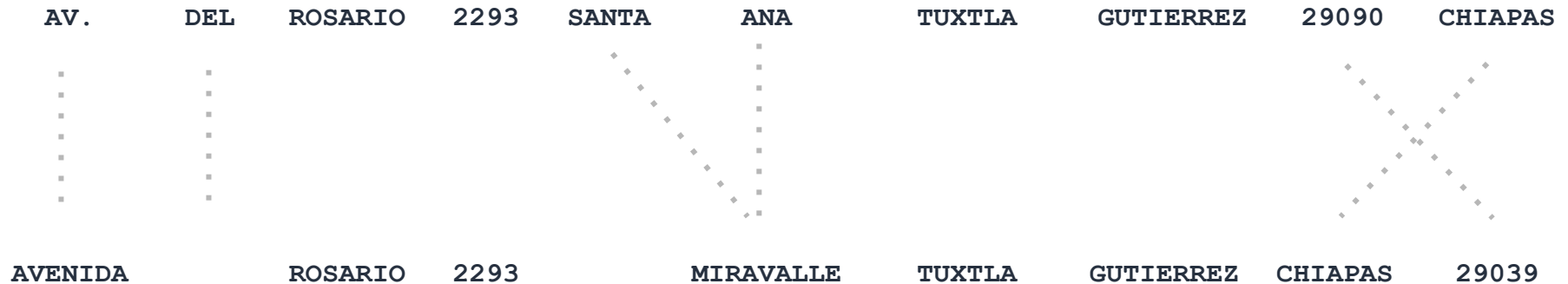
La idea básica es calcular el número mínimo de cambios para transformar una secuencia en la otra.



Algoritmo de Levenshtein

Para comparar la similitud entre las direcciones de ESMaestras y las de los candidatos usamos la **distancia de Levenshtein**. Las direcciones con valores de similitud más altos son consideradas las más precisas.

Por ejemplo, las siguientes direcciones tienen 78% de similitud:



Similitud de Direcciones Basada en Algoritmo de Levenshtein

AV. DEL ROSARIO 2293 SANTA ANA Tuxtla Gutierrez 29090 CHIAPAS

```
[
  {'address': 'Avenida El Rosario 2293, Caminera, Tuxtla Gutiérrez, Chiapas, 29090', 'location': {'x':
-93.099509986629, 'y': 16.740710016685}, 'extent': {'xmin': -93.100509986629, 'ymin': 16.739710016685,
'xmax': -93.098509986629, 'ymax': 16.741710016685}} U {"INE_ENTIDAD": 7, "INE_MUNICIPIO": 102, "INE_COL":
"CAMINERA", "INE_CP": "29090", "areakm2": 0.047} U {"comp-cp": 100, "comp-col": 35, "score_address": 83.8}

  {'address': 'Avenida Rosario 2293, Miravalle, Tuxtla Gutiérrez, Chiapas, 29039', 'location': {'x':
-93.129425381397, 'y': 16.771478804309}, 'extent': {'xmin': -93.130425381397, 'ymin': 16.770478804309,
'xmax': -93.128425381397, 'ymax': 16.772478804309}} U {"INE_ENTIDAD": 7, "INE_MUNICIPIO": 102, "INE_COL":
"MIRAVALLE", "INE_CP": "29039", "areakm2": 0.047} U {"comp-cp": 80, "comp-col": 22, "score_address": 75.6}

  {'address': 'Avenida El Rosario, Santa Ana, Tuxtla Gutiérrez, Chiapas, 29090', 'location': {'x':
-93.098146514924, 'y': 16.740221469028}, 'extent': {'xmin': -93.099146514924, 'ymin': 16.739221469028,
'xmax': -93.097146514924, 'ymax': 16.741221469028}} U {"INE_ENTIDAD": 7, "INE_MUNICIPIO": 102, "INE_COL":
"SANTA ANA", "INE_CP": "29090", "areakm2": 0.047} U {"comp-cp": 100, "comp-col": 100, "score_address": 89.4}

  {'address': '29090, Santa Ana, Tuxtla Gutiérrez, Chiapas', 'location': {'x': -93.097572819099, 'y':
16.739326771929}, 'extent': {'xmin': -93.102572819099, 'ymin': 16.734326771929, 'xmax': -93.092572819099,
'ymax': 16.744326771929}} U {"INE_ENTIDAD": 7, "INE_MUNICIPIO": 102, "INE_COL": "SANTA ANA", "INE_CP":
"29090", "areakm2": 1.183} U {"comp-cp": 100, "comp-col": 100, "score_address": 77.0}
]
```



Filtrado

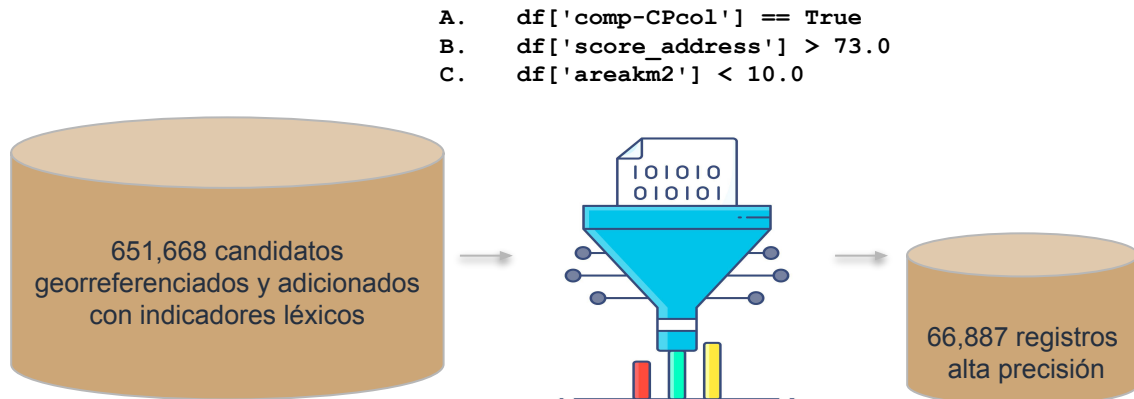
Aplicar reglas sobre los datos para separarlos



Filtrado de Datos Mediante Aplicación de Criterios

Los candidatos georreferenciados y verificados por *spatial join* son filtrados mediante criterios basados en los atributos agregados en los pasos previos.

- A. **Verificación de coordenadas** (latitud,longitud) con datos oficiales de entidad, municipio, colonia y CP;
- B. **Alta similitud** entre la escritura de la dirección del dato original y la dirección obtenida de la API;
- C. **Área pequeña** de incertidumbre.



AV. DEL ROSARIO 2293 SANTA ANA Tuxtla Gutierrez 29090 CHIAPAS

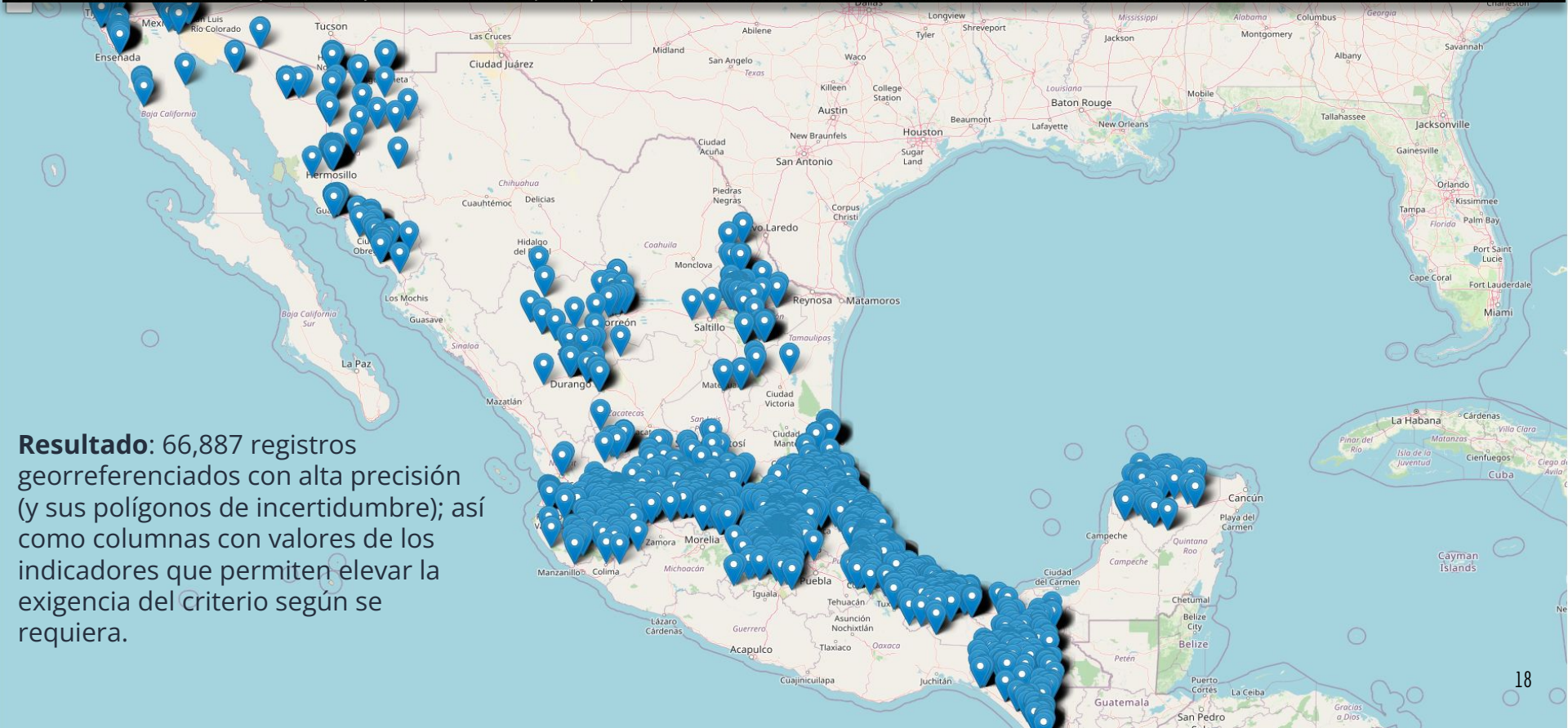
```
{'address': 'Avenida El Rosario 2293, Caminera, Tuxtla Gutiérrez, Chiapas, 29090',  
'location': {'x': -93.099509986629, 'y': 16.740710016685}, 'extent': {'xmin': -93.100509986629,  
'ymin': 16.739710016685, 'xmax': -93.098509986629, 'ymax': 16.741710016685}} U {"INE_ENTIDAD": 7,  
"INE_MUNICIPIO": 102, "INE_COL": "CAMINERA", "INE_CP": "29090", "areakm2": 0.047} U {"comp-cp": 100,  
"comp-col": 35, "score_address": 83.8}
```

```
{'address': 'Avenida Rosario 2293, Miravalle, Tuxtla Gutiérrez, Chiapas, 29039', 'location':  
{ 'x': -93.129425381397, 'y': 16.771478804309}, 'extent': {'xmin': -93.130425381397, 'ymin':  
16.770478804309, 'xmax': -93.128425381397, 'ymax': 16.772478804309}} U {"INE_ENTIDAD": 7,  
"INE_MUNICIPIO": 102, "INE_COL": "MIRAVALLE", "INE_CP": "29039", "areakm2": 0.047} U {"comp-cp": 80,  
"comp-col": 22, "score_address": 75.6}
```

```
{'address': 'Avenida El Rosario, Santa Ana, Tuxtla Gutiérrez, Chiapas, 29090', 'location':  
{ 'x': -93.098146514924, 'y': 16.740221469028}, 'extent': {'xmin': -93.099146514924, 'ymin':  
16.739221469028, 'xmax': -93.097146514924, 'ymax': 16.741221469028}} U {"INE_ENTIDAD": 7,  
"INE_MUNICIPIO": 102, "INE_COL": "SANTA ANA", "INE_CP": "29090", "areakm2": 0.047} U {"comp-cp":  
100, "comp-col": 100, "score_address": 89.4}
```

```
{'address': '29090, Santa Ana, Tuxtla Gutiérrez, Chiapas', 'location': {'x':  
-93.097572819099, 'y': 16.739326771929}, 'extent': {'xmin': -93.102572819099, 'ymin':  
16.734326771929, 'xmax': -93.092572819099, 'ymax': 16.744326771929}} U {"INE_ENTIDAD": 7,  
"INE_MUNICIPIO": 102, "INE_COL": "SANTA ANA", "INE_CP": "29090", "areakm2": 1.183} U {"comp-cp":  
100, "comp-col": 100, "score_address": 77.0}
```

Folio	score_address	address	latitude	longitude	ine_colname	ine_coltype	ine_cp	areakm2	score_compcp	score_comcol	geometry
1	80	Calle Las Flores, Piedras Negras, Tlaxiucayan, Veracruz de Ignacio de la Llave, 95226	18.766385312806	-96.1769724417	30	180	PIEDRAS NEGRAS	27	95226		
3	86	Calle Abasolo, El Calvario, Huichapan, Hidalgo, 42404	20.377094967763	-99.648200407038	13	29	EL CALVARIO	7	42400	0.046	100 100 POLYGON ((-99.64
4	91	Calle Nunkini 476, Héroes de Padierna, Tlalpan, Ciudad de México, 14200	19.276823537927	-99.213034821536	9	12	HEROES DE PADIERNA	1	14200	0.046	
5	90	Calle Santa Genoveva 1227, La Purísima, Guadalupe, Nuevo León, 67129	25.706879997489	-100.234310021733	19	26	LA PURISIMA	1	67129	0.044	100 100
6	87	Avenida El Rosario, Santa Ana, Tuxtla Gutiérrez, Chiapas, 29090	16.740221469028	-93.098146514924	7	102	SANTA ANA	1	29090	0.047	100 100 POLYGON



Resultado: 66,887 registros georreferenciados con alta precisión (y sus polígonos de incertidumbre); así como columnas con valores de los indicadores que permiten elevar la exigencia del criterio según se requiera.

Conclusion, Recomendaciones y Notas

Con la metodología presentada se pueden obtener coordenadas con buena precisión para un volumen grande de datos de ESMaestras (66,887) pero no para todos.

- La metodología es repetible y adaptable a nuevos criterios o adición de indicadores para filtrado.
- Los indicadores presentados ya son persistentes en las tablas intermedias del proceso lo que permite cambiar el filtrado según se requiera (directorios "*candidates*" y "*joined*" del *Google Drive*).
- Los criterios aplicados estuvieron basados en propiedades estadísticas como la distribución de valores de los indicadores (pej. similitud de direcciones en Q3-Q4).
- El nivel de exigencia en los criterios de filtrado determina la precisión del conjunto resultante pero también una reducción considerable en los datos finalmente obtenidos.



SALUD
SECRETARÍA DE SALUD



Instituto Nacional
de Salud Pública



CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES
CIENCIAS Y TECNOLOGÍAS



CentroGeo
Centro de Investigación en
Ciencias de Información Geoespacial, A.C.

Geocodificación de los datos de ES Maestras : Un enfoque basado en técnicas de Ciencias de Información Geoespacial

GRACIAS

Alejandro Molina Villegas
CONAHCYT - CentroGeo
amolina@centrogeo.edu.mx
(55) 4050 8741
20 de Septiembre de 2023



ANEXOS

ANEXO Referencias

- Hill, L. L. 2006. Georeferencing: The Geographic Associations of Information, 2. Cambridge, MA: MIT Press.
- Fiscus, J. G., Ajot, J., Radde, N., & Laprun, C. (2006, May). Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech. In LREC (pp. 803-808).
- <https://developers.arcgis.com/rest/geocode/api-reference/geocoding-find-address-candidates.htm>
- <https://www.geeksforgeeks.org/sequence-alignment-problem/>
- https://geopandas.org/en/stable/gallery/spatial_joins.html
- https://idegeo.centrogeo.org.mx/layers/geonode:ine2010_colonias_areas
- https://idegeo.centrogeo.org.mx/layers/geonode%3Aine2010_colonias_areas/pdf_metadata_layer

ANEXO Versionado y Configuración del Código Fuente y Programación

- Python 3.8.0
- requests==2.27.1
- geopandas==0.8.1
- thefuzz==0.19.0

ANEXO Definición de Levenshtein

Definition [\[edit \]](#)

The Levenshtein distance between two strings a , b (of length $|a|$ and $|b|$ respectively) is given by $\text{lev}(a, b)$ where

$$\text{lev}(a, b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \text{lev}(\text{tail}(a), \text{tail}(b)) & \text{if } a[0] = b[0], \\ 1 + \min \begin{cases} \text{lev}(\text{tail}(a), b) \\ \text{lev}(a, \text{tail}(b)) \\ \text{lev}(\text{tail}(a), \text{tail}(b)) \end{cases} & \text{otherwise} \end{cases}$$