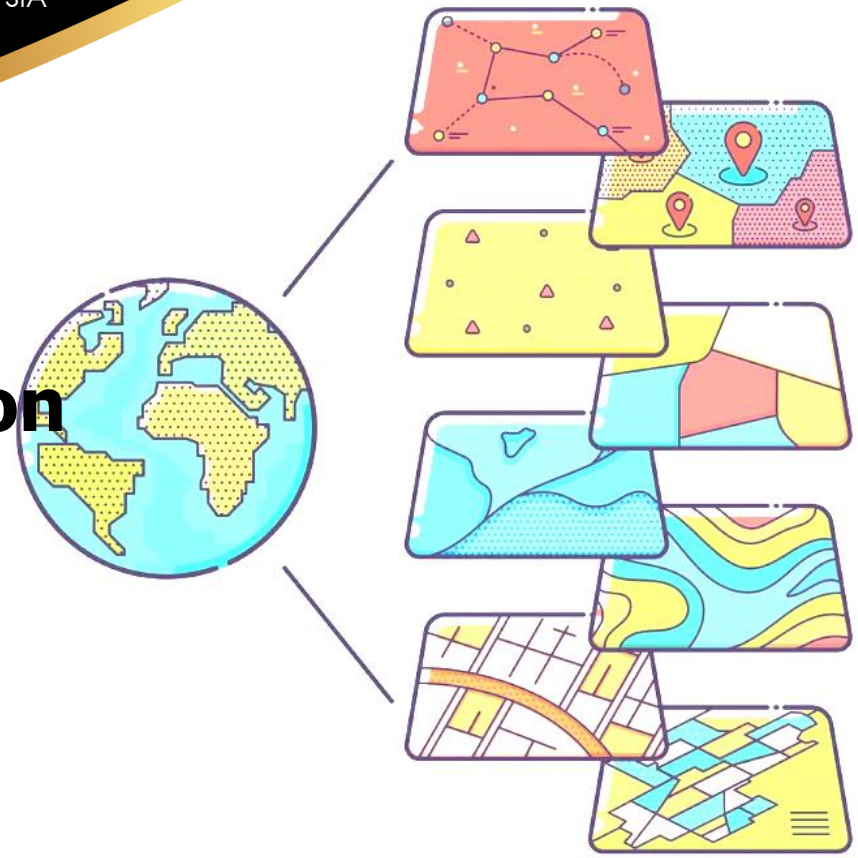
Application of seamless hybrid geocoding solution for business location using KAWASANKU API



***Ahmad Najmi ARIFFIN,** Mohamad Hamizan ABDULLAH

Core Team Big Data Analytics (CTADR),
Department of Statistics, Malaysia

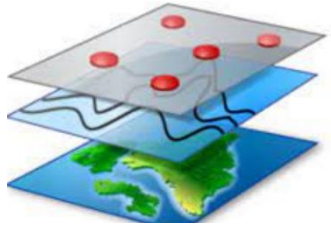
MyStats 2022
9TH MALAYSIA STATISTICS CONFERENCE
Department of Statistics Malaysia
Dealing with Uncertainties: Unearthing Measures for Recovery
4TH OCT. 2022 (VIRTUAL) & 5TH OCT. 2022 (ILSM, SUNGKAI, PERAK)
Organised by:
PRIME MINISTER'S DEPARTMENT DEPARTMENT OF STATISTICS MALAYSIA
BANK NEGARA MALAYSIA CENTRAL BANK OF MALAYSIA
MALAYSIA INSTITUTE OF STATISTICS

Problem Statement : Scenario ~ Solution

The purpose of our study was to verify the distributions of geocode location to consider when analysing geocoded address data, as well as to develop methods for enriching demographic databases and representing multiple levels - district, parliament, and state legislative assembly (Malay: Dewan Undangan Negeri, DUN) - using public repository **KAWASANKU API** from **Github platform**.

Geocoding

Open data



ArcGIS API

KawasanKU API

Business Entity

1. Introduction and background

1.1 Benefits of Geocoding and Structuring Address Data
1.2 The advantages of geocoding for businesses
1.3 Availability of open data for business research

Figure 1: Smart Search Mechanism using ArcGIS API

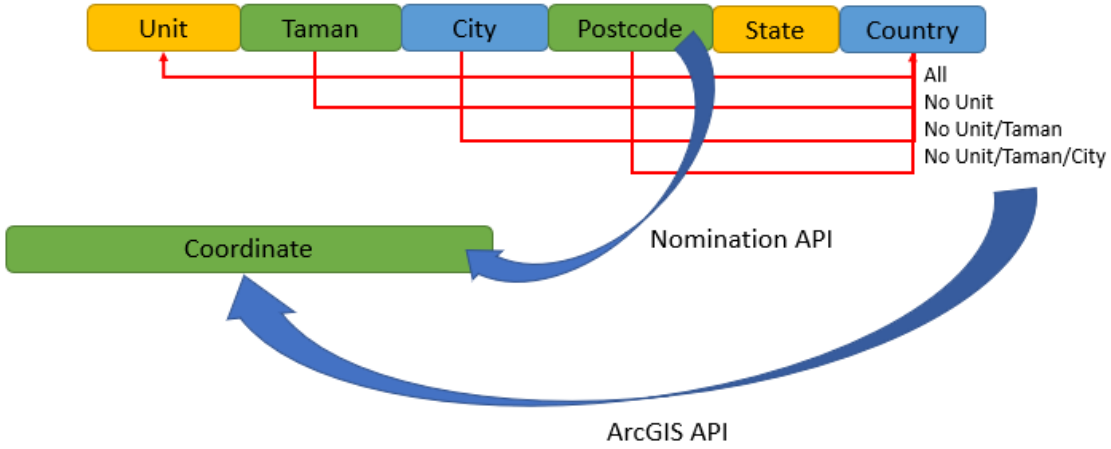


Table 1: A summary of the pros and cons of utilising online geocoding services.

Pros	Cons
1. Easy to use	1. No control over the reference database
2. Immediate coordinate results	2. No control over the parameter of geocoding process (e.g., match score, relaxation rules)
3. The user does not need to acquire, maintain, and update the reference database	3. Unknown quality of geocoded results
4. No software or tool is required on the user side	4. Relying on the Internet infrastructure

2. Methodology

Figure 2: Installation of the "geopy" module in Python IDE

```
from geopandas.tools import geocode, geocoding, reverse_geocode
```

```
type(geocode), type(reverse_geocode), type(geocoding)
```

```
(function, function, module)
```

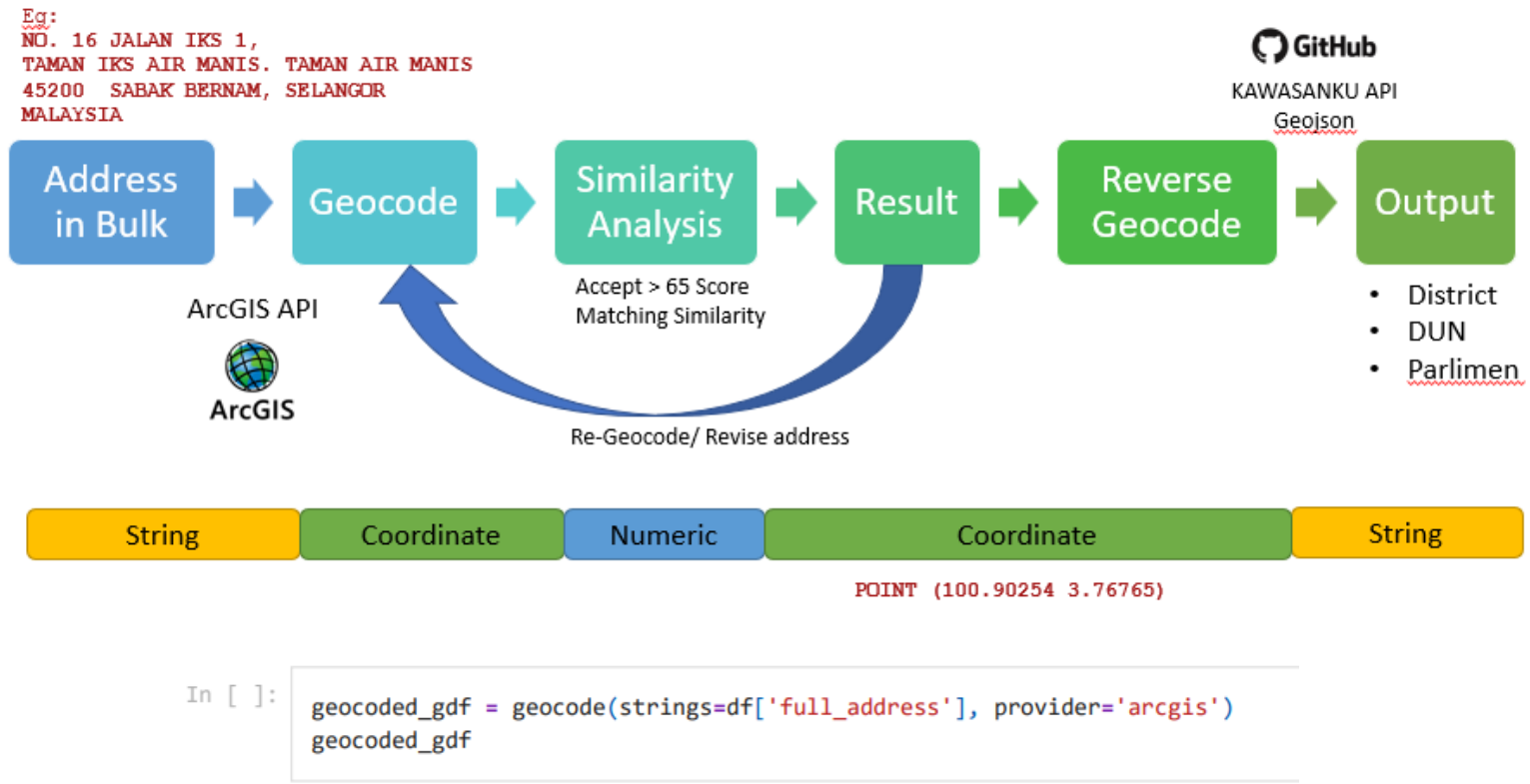
```
import geopy, inspect
print(geopy.__version__)
```

```
1.17.0
```

```
# use inspection, but limit to just classes
inspect.getmembers(geopy, predicate=inspect.isclass)
```

```
[('ArcGIS', geopy.geocoders.arcgis.ArcGIS),
 ('AzureMaps', geopy.geocoders.azure.AzureMaps),
 ('Baidu', geopy.geocoders.baidu.Baidu),
 ('Bing', geopy.geocoders.bing.Bing),
 ('DataBC', geopy.geocoders.databc.DataBC),
 ('GeoNames', geopy.geocoders.geonames.GeoNames),
 ('GeocodeEarth', geopy.geocoders.geocodeearth.GeocodeEarth),
```

Figure 3: Workflow for Geocoding Address and Reverse Geocode




```
In [ ]:
geocoded_gdf = geocode(strings=df['full_address'], provider='arcgis')
geocoded_gdf
```

2. Methodology

Figure 3 : Installation of the "fuzzywuzzy" module to Python Library

Finding strings that approximately match a pattern in your data using Python.

```
!pip install fuzzywuzzy python-Levenshtein -qq
```

```
from fuzzywuzzy import fuzz
from fuzzywuzzy import process
```

```
fuzz.ratio("Sankarshana Kadambari","Sankarsh Kadambari")
```

```
92
```

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j)+1 \\ \mathrm{lev}_{a,b}(i,j-1)+1 \\ \mathrm{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise} \end{cases} \quad (2)$$

where 1(ai=bj) denotes 0 when a = b and 1 otherwise. Finally, the Levenshtein similarity ratio is computed based on the Levenshtein distance, and is calculated using the formula in Equation. 3.

$$\frac{(|a|+|b|) - \mathrm{lev}_{a,b}(i,j)}{|a|+|b|} \quad (3)$$

where |a| and |b| are the lengths of sequence a and sequence b, respectively.

3. Result

3.1 Risk-assessment on valid geocode

Figure 5: Geocoded address(coordinate) and generated address from arcGIS API

In [ ]:
```
geocoded_gdf = geocode(strings=df['full_address'], provider='arcgis')
geocoded_gdf
```

Out[ ]:

	geometry	address

localhost:8891/lab/tree/Geocode_with_arcgis_and_Similarity_Score_Address.ipynb

9/16/22, 11:02 AM Geocode_with_arcgis_and_Similarity_Score_Address

	geometry	address
0	POINT (103.61336 1.66343)	411 Jalan Makmur 13, Taman Makmur, Kulai, 8100...
1	POINT (102.56284 2.14575)	Sungai Mati, Tangkak, Johor
2	POINT (103.31638 2.05099)	22 Jalan Cermai 2, Taman Suria, Kluang, 86000,...
3	POINT (103.67413 1.49669)	25 Jalan Uda Utama 1/1, Bandar Uda Utama, Joho...
4	POINT (102.81421 1.89754)	83600, Kampung Parit Guntong, Semerah, Batu Pa...
...	...	...
80	POINT (100.27700 6.41725)	Jalan Sanji, Taman Utara Guar Sanji, Arau, Kan...
81	POINT (100.26004 6.52359)	Lorong 4, Rancangan Perumahan Awam C, Chuping,...
82	POINT (100.26556 6.42183)	02600
83	POINT (100.26556 6.42183)	02600
84	POINT (100.24658 6.38267)	Jalan Behor Mentalon, Kurong Anai, Kangar, 026...

85 rows × 2 columns

1.Address in Bulk

↓

1.Geocode

3. Result

3.2 Risk-assessment on similarity analysis

Figure 6: Fuzzy matching analysis, Q-Ratio Score as Similarity Indicator

Geocode_with_arcgis_and_Similarity_Score_Address

	old_names	correct_names	correct_ratio	partial_ratio	ratio	QRatio	Wratio	
2	NO 22 JALAN CERMAI 2 TAMAN SURIA 86000 Johor, ...	22 Jalan Cermai 2, Taman Suria, Kluang, 86000,...	92	47	45	77	87	PASSED
3	25 JALAN UDA UATAMA 1 1 BANDAR UDA UTAMA 8120...	25 Jalan Uda Utama 1/1, Bandar Uda Utama, Joho...	93	41	46	80	88	PASSED
4	POS 67,LORONG HJ ANUAR, KG PT LUBOK DARAT, MUK...	83600, Kampung Parit Guntong, Semerah, Batu Pa...	57	30	26	44	54	FAILED
...	...	...	...	...	...	...	...	
80	NO. 463, KAMPUNG GUAR SANJI, JALAN RUMAH PAM A...	02600	100	100	11	11	60	FAILED
81	NO 11 LORONG 4 TAMAN EMAS BESERI 02450 PERLIS ...	Lorong 4, Rancangan Perumahan Awam C, Chuping,...	60	32	32	52	57	FAILED
82	451 KAMPUNG BARU PAUH 02600 PERLIS 02600 Per...	02600	100	100	16	16	60	FAILED

1.Geocode

Similarity Analysis

Result

3. Result

3.3 Optimisation - Data Enrichment using open data sharing source, extract data using KAWASANKU API

Figure 7: KAWASANKU API Matching - indicate geocode point within district, parliament and DUN

```
In [ ]:
levels = ['country','state','district','parlimen','dun']
for i in [1,2,3,4]:
    df[levels[i]] = df.progress_apply(lambda x: reverse_geocode(x['lon'], x['lat'], int
df.head()
```

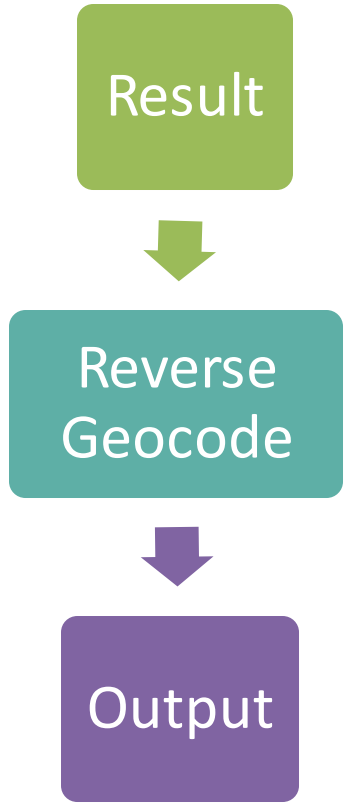
```
100%|██████████| 5/5 [00:00<00:00, 165.15it/s]
100%|██████████| 5/5 [00:00<00:00, 167.49it/s]
100%|██████████| 5/5 [00:00<00:00, 144.42it/s]
100%|██████████| 5/5 [00:00<00:00, 83.17it/s]
```

localhost:8891/lab/tree/Reverse_geocoderUpdate_toDUN_Parliment.ipynb#Bulk-Search 2/4

9/12/22, 10:36 AM Reverse_geocoderUpdate_toDUN_Parliment

Out[ ]:

	name	lon	lat	state	district	parlimen	dun
0	Pulai Chondong	102.246508	5.808972	Kelantan	Machang	P.029 Machang	N.33 Pulai Chondong
1	Pendang	100.474091	5.986576	Kedah	Pendang	P.011 Pendang	N.18 Tokai
2	Taiping	100.736561	4.849122	Perak	Larut Dan Matang	P.060 Taiping	N.17 Pokok Assam
3	Padang Tengku	101.981106	4.230687	Pahang	Lipis	P.079 Lipis	N.03 Padang Tengku
4	Kinabatangan	117.861843	5.587962	Sabah	Kinabatangan	P.187 Kinabatangan	N.58 Lamag



DEMOSTRATION STEP BY STEP

4. Discussion and Conclusion

- This framework to suggest a **seamless and less-dependent** workflow.
- The benefits of Open Data can be **increased if both private industry and public agencies advocate for the Open Data sharing platform** and mindset, thereby fostering a thriving open data ecosystem.
- The federal and state levels of government are also able to use a geospatial approach to **plan for better strategies to enhance new uncertainty business entities**.
- **Planning a more advantageous location** for an entrepreneur's business based on the distribution network using a map.
- Measuring the **geographical distribution of economic activity** is essential for scientific research and policy formation.

Limitation

- Unstructured address data is a **common obstacle.**
- Geocoding API's address geocoding has **significantly higher latency** and produces **less accurate results for incomplete** or ambiguous queries;
- Not recommended for real-time user input-responsive applications.
- In the **future, we plan to employ better model** in order to comprehend the qualitative similarities between two datasets more thoroughly.

References

Veregin, H. (1999). Data quality parameters. *Geographical information systems*, *1*, 177-189.

MacEachren, A. M. (2017). Leveraging big (geo) data with (geo) visual analytics: Place as the next frontier. In *Spatial data handling in big data era* (pp. 139-155). Springer, Singapore.

Kirby, R. S., Delmelle, E., & Eberth, J. M. (2017). Advances in spatial epidemiology and geographic information systems. *Annals of epidemiology*, *27*(1),1-9.

Cheng, Z., Caverlee, J., & Lee, K. (2010, October). 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* (pp.759-768).

Location Intelligence Drives Competitive Edge In The Digital Age, July 2018, A Forrester Consulting Thought Leadership Paper Commissioned By Loqate, A GBG solution, https://info.loqate.com/hubfs/Loqate%202018/Reports/Location%20Intelligence%20Drives%20Competitive%20Edge%20In%20The%20Digital%20Age.pdf

A study on the Impact of Re-use of Public Data Resources, November 2015, Wendy Carrara, Wae San Chan, Sander Fischer, Eva van Steenbergen (Capgemini Consulting), https://data.europa.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf

Khayyat, M., & Bannister, F. (2015). Open data licensing: more than meets the eye. *Information Polity*, *20*(4), 231-252.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, *29*(4), 258-268.

Dawes, S. S., Vidiasova, L., & Parkhimovich, O. (2016). Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly*, *33*(1), 15-27.

GeoPy, Welcome to GeoPy documentation! Retrieved on Sep. 2022, From https://geopy.readthedocs.io/en/stable/.

Nominatim Documentation, Nominatim API, Retrieved on Sep. 2022, From https://nominatim.org/release-docs/develop/api/Overview/.

OSM's Nominatim Service, Nominatim Usage Policy, Retrieved on Sep. 2022, From https://operations.osmfoundation.org/policies/nominatim/.

Krieger, N., Waterman, P., Lemieux, K., Zierler, S., & Hogan, J. W. (2001). On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *American journal of public health*, *91*(7), 1114.

Roongpiboonsopit, D., & Karimi, H. A. (2010). Comparative evaluation and analysis of online geocoding services. *International Journal of Geographical Information Science*, *24*(7),1081-1100.

TheFuzz documentation - Github repository, Retrieved on Sep. 2022, From https://github.com/seatgeek/fuzzywuzzy.

THANK YOU

StatsMalaysia

www.dosm.gov.my