2020 Asia–Pacific Statistics Week
A decade of action for the 2030 Agenda: Statistics that leaves no one and nowhere behind
15-19 JUNE 2020 | Bangkok, Thailand

Big Data for Official Statistics: Administrative Area Identification from Plain Text Address

Wa Ode Zuhayeni Madjida*[1]; Takdir [2]

[1] Badan Pusat Statistik, Jakarta, Indonesia. Email: zuhayeni@bps.go.id
[2] Politeknik Statistika STIS, Jakarta, Indonesia. Email: takdir@stis.ac.id

Abstract:

The successful implementation of Sustainable Development Goals (SDGs) in Indonesia is inseparable from the problem of data availability. BPS - Statistics Indonesia has an important role in monitoring and evaluating SDGs accomplishment by providing data and information of SDGs indicators. To meet these needs, BPS conducts various modernizations in data collection including the use of big data as one of its data sources.

Utilization of big data can come with lower costs and can lead to more timely and granular statistics. One of the issues faced in using big data is integrating it with traditional sources, which are census and survey. Census and survey use Statistical Work Areas as geographical standards in collection process and dissemination of data. The statistical work areas are clustered by government administrative areas including transmigration settlement units, isolated tribal community settlements and Census Block enumeration areas. Data from other sources such as big data do not consider this form of categorizations. Administrative areas in big data are usually in the form of free text which, although referring to the same administrative area, can differ between one big data source and another. Identification or mapping of free text into an administrative area manually on large data sizes is impossible.

To answer the problem, this study proposes a text processing approach. We use Google Maps and websites that contain addresses in free text format as input of our simulation. Master File Desa (MFD), a standardized statistical area database provided by BPS, is used as a corpus for the area identification process. MFD is one part of the Statistical Work Area which is used as a reference for village names in official statistics census and survey activities. To measure the accuracy of the area identification process, a manual approach and coordinate overlay are used.

Keywords: statistical unit; text processing; geo-reference; geographic area; data matching

1. Introduction:

Data availability greatly influences the successful implementation of Sustainable Development Goals (SDGs). BPS – Statistics Indonesia as the agency that responsible for collecting official statistics in Indonesia also plays an important role in providing data and information for SDGs indicators. BPS routinely conducts various surveys and collaborates with and coordinates with the Ministry / Agency for the availability of the data. On the other hand, BPS also modernizes data collection, one of which is the use of big data as a new data source.

Big data is known for its large volume, high velocity, and high variety. The large volume of data makes it possible to produce better accuracy and detail, high velocity may cause statistical estimates can be done more often and more timely, and high variations provide opportunities in generating new statistics (Braaksma. B & Zeelenberg. K, 2020). However, big data also presents challenges in its use for official statistics.

The challenges of using big data for official statistics are revealed in (Hammer. C. L, et al, 2017), including quality assurance, accessing big data, the need for skills and the use of new technology. (Braaksma. B & Zeelenberg. K, 2020) also mentioned a number of challenges, namely highly volatile big data characteristics that make the cover age of population to which they may change day to day, leading to inexplicable jumps in time-series. In addition, there are uncontrolled changes in sources that threaten continuity and comparability. Another issue that arises is the problem of integration, especially with traditional data sources, particularly censuses and surveys. In fact, this integration can provide more detail in statistical series.

To facilitate integration with other sources such as big data, National Statistical Office (NSO) may apply standardized national addresses or ideally geo-referenced addresses (Stats Brief, 2019).

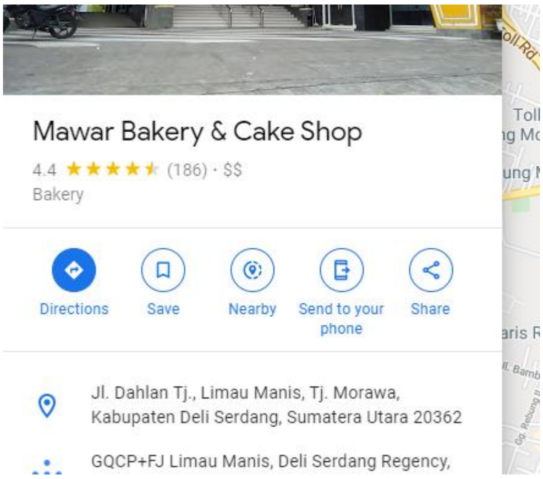
Census and survey use Statistical Work Areas as geographical standards in collection process and dissemination of data. The Statistical Work Areas are clustered by government administrative areas including transmigration settlement units, isolated tribal community settlements and Census Block enumeration areas. Data from other sources such as big data do not consider this form of categorizations. Administrative areas in big data are usually in the form of free text. Each administrative area information provided may differ from one another, although referring to the same administrative area. This structural difference makes it difficult to integrate or link big data with traditional sources. Identification or mapping of free text into a standard administrative area manually on large data sizes is impossible.

Text processing in big data is a very developed field of research nowadays. By using text processing, the free text analysis process can be done automatically for getting structured information. This paper uses a text processing approach to obtain a standardized Statistical Work Area from big data sources, so that the big data collected can be integrated or compared to traditional sources. In this study we present another proof that, using proper approaches, big data sources can be adapted and convverted into official statistics.

2. Methodology:

2.1. Plain Text Address in Big Data

Administrative area that available on big data are usually in the form of free text or plain text and vary greatly from one big data source to another source. The geographical level covered by the big data source between one text can also be different from the other address text. One source can give administrative area information to village level, while another is only provide at province level. The following are examples of free text from administrative areas or addresses obtained from Google Maps.



Mawar Bakery & Cake Shop

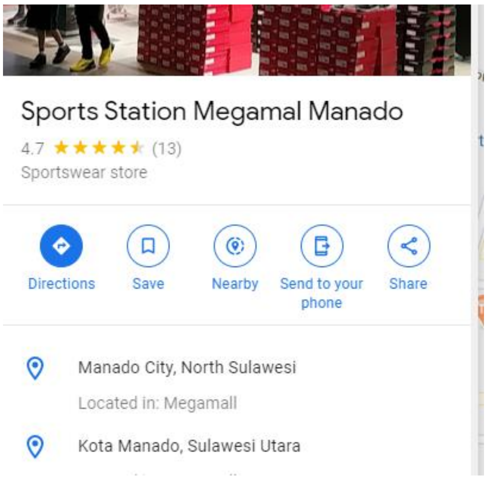
4.4 ★★★★½ (186) · $$

Bakery

Jl. Dahlan Tj., Limau Manis, Tj. Morawa, Kabupaten Deli Serdang, Sumatera Utara 20362

GQCP+FJ Limau Manis, Deli Serdang Regency,

Figure 1. Address on Google Map that Give Information Until Village Level



Sports Station Megamal Manado

4.7 ★★★★½ (13)

Sportswear store

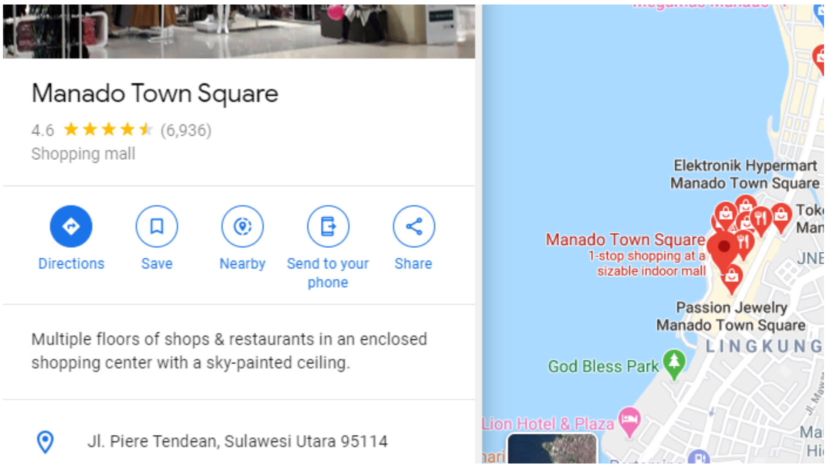
Manado City, North Sulawesi

Located in: Megamall

Kota Manado, Sulawesi Utara

Figure 2. Address on Google Map that Give Information Until Regency/District Level

We give some address examples from Google Map. It can be seen that the level of a given area can vary. Figure 1 offer the information to the village level and figure 2 gives area to regency/district level. While Figure 3 only provides geographic information at the province level.



Manado Town Square

4.6 ★★★★★ (6,936)

Shopping mall

Multiple floors of shops & restaurants in an enclosed shopping center with a sky-painted ceiling.

Jl. Piere Tendean, Sulawesi Utara 95114

Figure 3. Address on Google Map that Give Information in Province Level

2.2. Proposed Text Processing Algorithm

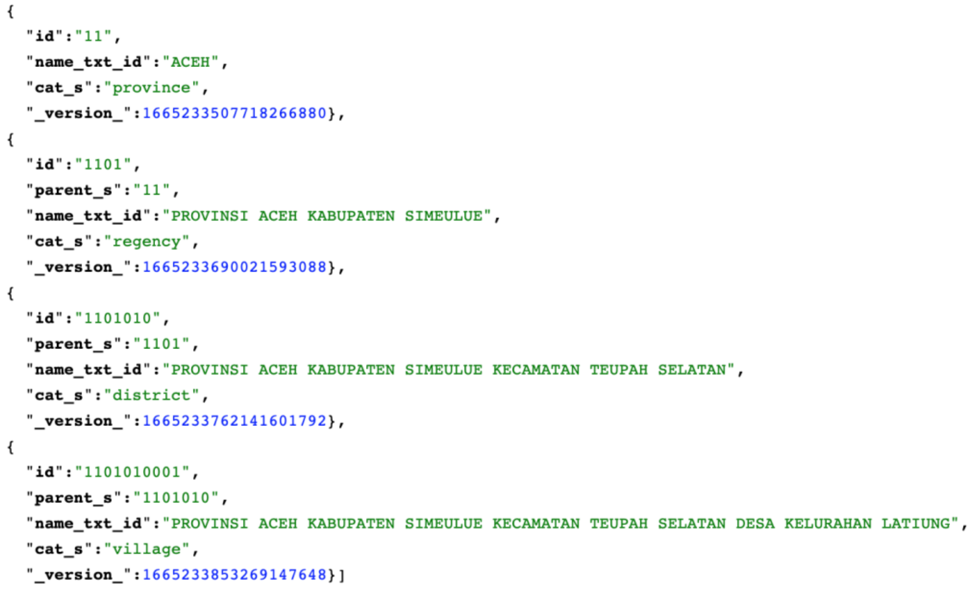
We use Master File Desa (MFD) database, a standardized statistical area database provided by BPS, as a corpus for the area identification process. MFD is one part of the Statistical Work Area which is used as a reference for village names in official statistics census and survey activities. The MFD used are at the province, district, sub-district level, up to the village level, both the area code and the name of the region. Table 1 shows an example from MFD:

Table 1. Example of MFD

Province Code	**Province Name**	**Regency code**	**Regency Name**	**Sub-district Code**	**Sub-district Name**	**Village Code**	**Village Name**
11	Aceh	16	Aceh Jaya	030	Krueng Sabee	010	Kabong
72	Sulawesi Tengah	09	Tojo Una-Una	070	Togean	002	Awo
94	Papua	01	Merauke	051	Eligobel	011	Tof-Tof

MFD names, e.g. province name and regency name, are indexed using Apache Solr. Apache Solr is a search platform integrated with scalable storage system (Kumar. J, 2015; Serafini. A, 2013). Records in search platform are represented as *documents* which are similar with rows in database system. Each document contains fields with associated field type such as `StrField`, `FloatPointField`, and `TextField`. MFD names are stored as `TextField` which can be indexed and queried with various text processing algorithm, while MFD codes, e.g. province code and regency code, are in `StrField` which is targetted to store single word or unique key of document.

MFD names are concatenated into one text with "space" character as separator before indexing. Area type, e.g. "province" and "regency", is also added in front of each MFD names to construct the text. This approach is intended to improve the accuracy of searching where address text may contain area types. The concatenated text is shown as the value of `name_txt_id` field in figure 5.



```
{
  "id":"11",
  "name_txt_id":"ACEH",
  "cat_s":"province",
  "_version_":1665233507718266880},
{
  "id":"1101",
  "parent_s":"11",
  "name_txt_id":"PROVINSI ACEH KABUPATEN SIMEULUE",
  "cat_s":"regency",
  "_version_":1665233690021593088},
{
  "id":"1101010",
  "parent_s":"1101",
  "name_txt_id":"PROVINSI ACEH KABUPATEN SIMEULUE KECAMATAN TEUPAH SELATAN",
  "cat_s":"district",
  "_version_":1665233762141601792},
{
  "id":"1101010001",
  "parent_s":"1101010",
  "name_txt_id":"PROVINSI ACEH KABUPATEN SIMEULUE KECAMATAN TEUPAH SELATAN DESA KELURAHAN LATIUNG",
  "cat_s":"village",
  "_version_":1665233853269147648}]
```

Figure 4. Documents format stored in search platform

In general, the algorithm consists of two major processes, they are indexing and searching. Indexing process make data can be searchable in machine (Shahi. D, 2015).

Indexing

At index time, when a field is being created, `name_txt_id` is breaked into lexical units, called *tokens*, using `LetterTokenizedFactory`. It creates tokens by cutting text when non-letter character found, such as space and period, and discarding that character. Additionally, `LowerCaseFilterFactory` is applied which will convert all tokens into lower case letters in order to ignore the case when searching performed.

These tokens are called *terms* in scoring. Therefore, index file in search platform contains the set of terms including information about associated document, occurrence of term in that document, positions, and so on.

Searching

At query time, the values being searched for are analyzed similar with indexing mechanism above and the terms that result are matched against those that are stored in the field's index. In searching, `SynonymGraphFilterFactory` is used to map the synonym of words to the associated ones in comma separated format as shown below.

```
prov,prov.,prop,prov.,provinsi
kab,kab.,kabupaten
kec,kec.,kecamatan
kel,kel.,kelurahan
```

Score of each search result is calculated usign term frequency–inverse document frequency (TFIDF). TFIDF score of each token/term in text being seached for computed as the multiplication of *tf* and *idf* variables in `name_txt_id` field, where:

$$\boldsymbol{tf = \frac{freq}{\left(freq + k1 * \left(1 - b + b * \frac{dl}{avgdl}\right)\right)}}$$

$$\boldsymbol{idf = \mathrm{lo\ g}\left(1 + \frac{N - n + 0.5}{n + 0.5}\right)}$$

from:

freq: occurrences of term within document
k1: term saturation parameter
b: length normalization parameter
dl: length of field
avgdl: average length of field
n: number of documents containing term
*N: t*otal number of documents with field.

Document's score is the sum of every tokens/terms that matched with the queried text tokens. The greates score will become the selected estimated Statistical Work Areas.

2.3. Simulation and Testing

The following are the simulation and testing steps:

1. Crawling simulation data on Google Map by using 'market' keyword. Some variables that has been collected are name of location, Universal Resource Locator (URL), address string, coordinates, and type of place. We get 17,935 records of data.
2. Doing overlay of MFD polygon with coordinates obtained from crawling results for getting MFD location of each location.
3. To process each address string that has been collected by using algorithm that was designed in section 2.2. The return value is estimated area level (province / regency / sub-district / village) and administrative area based on MFD at the level that can be estimated.
4. To test the accuracy of the designed text processing method, we compare MFD results from step 3 with the results in step 2.
5. Evaluating results.

3. Result:

From 17,935 data processed, the algorithm can provide accuracy at village level of 73.58%, MFD from the algorithm give the same results as the MFD from overlayed. The algorithm does not give the same results as the MFD overlay of 26.42%, where 31 records or 0.0065% of which the proposed algorithm cannot issue its estimation (NULL value). This is caused by anomaly data where the crawler does not get the address string.

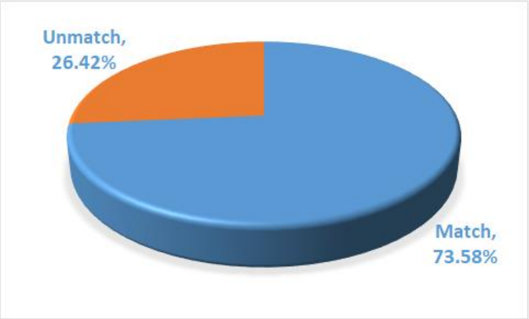


Figure 5. Matching results

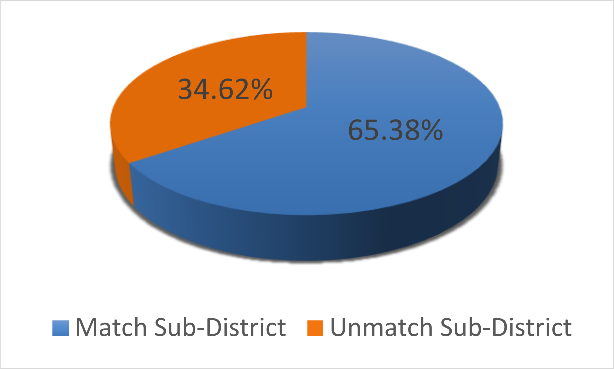


Figure 6. Matching results in Sub-District Level

From records that did not match, but still released estimates of their area, there were 65.38% which although estimates at the village level did not match, estimates at the sub-district level were matches. Some of the causes are:

- If we look at the location address string, the estimated area given by the algorithm actually matches with MFD database. What could possibly happen is expansion of regions, but the address on Google Map does not adjust the expansion. So, based on overlay, the area obtained is the village resulting from the expansion. While address string in Google Map is still based on village name before expansion.

- The words of address on Google Map were not found in MFD village name.

Records that do not match at the village and sub-district levels are then matched again at the next level, which is district level. 82.39% of the algorithm results can match at the district level. One cause of matching that is only at the district level is the use of abbreviations in writing districts on Google Map, for example "Jambi Tim." which shows the district of East Jambi (Jambi Timur). This makes the algorithm can not find the MFD area.

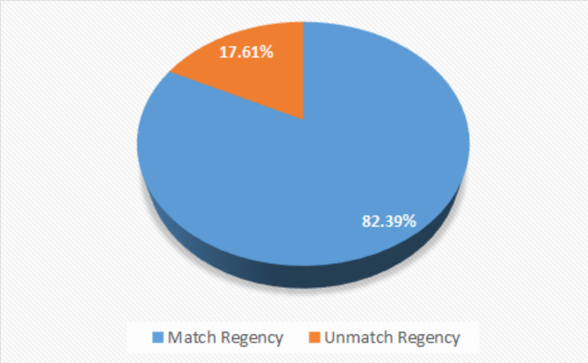


Figure 7. Matching results in Regency Level

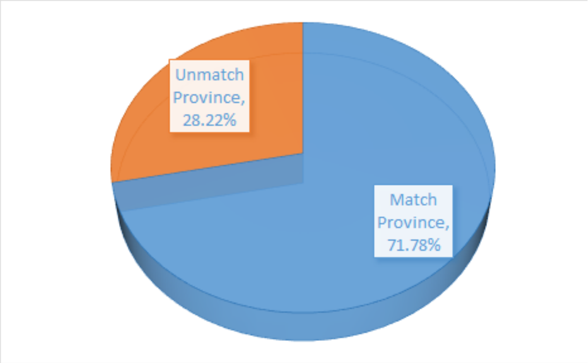


Figure 8. Matching Results in Province Level

Furthermore, from 17.61% of records that did not match at the district level, match accuracy was then observed at the provincial level. There are 71.78% of records that do not match at the village to district level, but match at the province level. This is partly due to incomplete address presented on Google Map, which only contains provincial information. On the other hand, it was found 28.22% which did not match to the province level. Regional expansion and lack of information up to the province level on Google Map are the cause of inaccuracies.

4. Discussion and Conclusion:

Text processing is a field of research that is very developed at this time, especially its implementation on big data. Big data as a new data source for official statistics still presents many challenges, one of which is integration and links with traditional sources of censuses and surveys. Administrative area or geographical location is a variable that can be used. However, the location provided by the big data source is usually in free text. This paper proposes a text processing approach to get a standardized administrative area from big data source. The accuracy of the proposed approach is 73.58%. In next studies, some improvements can be made to the algorithm, namely processing the address string in which there is an abbreviation. This improvement is expected to improve the accuracy of identifying administrative areas in big data.

References:

Apache Solr Reference Guide. https://lucene.apache.org/solr/guide/8_5/index.html

Braaksma, B. & Zeelenberg, K. (2020). Big Data in Official Statistics. Centraal Bureau Voor De Statistiek.

Hammer, C. L., et al. (2017). Big Data: Potential, Challenges, and Statistical Implications. International Monetary Fund.

Kumar, J. (2015). Apache Solr Search Pattern. Packt Publishing.

Serafini, A. (2013). Apache Solr Beginner’s Guide. Packt Publishing.

Shahi, D. (2015). Apache Solr : A Practical Approach to Enterprise Search. Appress.

Stats Brief. (2019). Integrated Statistics: A Journey Worthwhile. United Nations ESCAP.