COMMUNICATIONS IN SCIENCE AND TECHNOLOGY
Homepage: cst.kipmi.or.id

Using neighborhood association area as new spatial data infrastructure to link public administrative data with GIS in Indonesia

Muriadi Arip[a*], Yan Wanglin[a], Kazuyo Hirose[b]

[a] *Eco-GIS Laboratory, Graduate School of Media and Governance, SFC, Keio University, Fujisawa shi - Kanagawa ken, 252-0882, Japan*
[b] *Department of International Cooperation, Japan Space System, Minato ku - Tokyo, 105-0011, Japan*

Article history:
Received: 24 April 2019 / Received in revised form: 30 May 2019 / Accepted: 02 June 2019

Abstract

Despite its potential use for earth observation and GIS-based analysis, Public Administrative Data (PAD) has been neglected in the spatial big data discussions. For instance, linking unaggregated public databases to the smallest administrative units for mining spatial data currently absents from literature. In this study, a neighborhood association base map was developed and the usability as a platform for linking PAD in Indonesia was investigated. The base map is proposed as a new feature in Indonesia's SDI. A data model was developed, and data accuracy and reliability were assessed by a case study. Four unaggregated databases obtained from public institutions were examined using common structured query language. The results show that from 1.3 million records, more than 95% can be directly linked to the base map. Finally, it is concluded that despite the existence of challenges, linking PAD with the base map is feasible and beneficial for GIS-based analysis.

Keywords: Neighborhood Association, Public Administrative Data, Geographic Information System (GIS), Spatial Data Infrastructure (SDI).

1. Introduction

The advanced technology with extraordinary and exponential improvements in data storage and computing capacities makes it possible to collect, manage, and analyze data in magnitudes and in manners that would have been inconceivable just a short time ago, so have the world's governments developed large-scale, comprehensive data files on tax programs, workforce information, benefit programs, health, and education [1]. In short, advancements in technology have permitted statistical agencies to overcome many of the limitations caused by processing large datasets [2]. Moreover, government departments and agencies around the world routinely collect administrative data produced by citizen interaction with the state [3]. However, not only in the GIS domain, even in the mainstream big data discussions, (public) administrative data has been largely neglected [4]. Therefore, it is not surprising that many researchers, particularly those in spatial-based research, found the lack of theoretical frameworks in the use of PAD.

In principle, PAD is generated for service delivery purposes, thus may not be exactly aligned with statistical or research needs, but they can nevertheless be useful to statisticians or researchers [5]. Moreover, with appropriate theoretical frameworks, adequate data infrastructure, and linking between data sources, PAD can be a big data source for earth observation and spatial-based analysis. Because what the data actually depicting is human-environment interactions occurred in spatiotemporal dimensions.

Indonesia is a country for a long time struggling with the use of their PAD. Currently, a policy called One Data Policy (ODP) has been launched. It is an initiative to promote research-based policy in the country. It is also aimed to reduce confusing data management while fulfilling government's data requirements for policy purposes. This policy is not yet well established. Amongst the drawback is related to the lack of ability to link between existing data as well as with other data sources [6].

Geographical Information (GI) can be a proper solution to the drawback. GI can play an important role as a common link between data [7]. Therefore, incorporating PAD management to the Indonesia Spatial Data Infrastructure (ISDI) is important. In fact, at this point, Indonesia has established its data management on track. Because, apart from ODP, in 2011 Indonesia has launched what is called One Map Policy (OMP). The goals include on standardizing and unifying spatial data across the Indonesian archipelago, creating a base map for all agencies to use, and making spatial data free and readily accessible for Indonesian citizens [8]. Nevertheless, the implementation of the policy far from enough to help linking PAD for greater use. Some constraining factors to mention are including the availability of large scale basic geospatial information, the availability of mapping guidelines, and human resources [9].

On the other hand, GIS, as a powerful tool for analyzing PAD geographically, is still ignored. Despite the realization of the importance of GIS over two decades ago and the recent acceleration of software development, many people in Indonesia remain ignorant of GIS [8]. It is still an unfamiliar tool for most of the local government offices in the country. In

* Corresponding author. Email: muriadii.a7@keio.jp.

consequence, despite the launch of OMP years ago, the development of ISDI is not congruent with the need for GIS-based research and analysis.

Administrative base maps are among the key features in any National SDI. The availability is required to extract spatial data within the PAD. In a large extent, it can be simply done by linking aggregated data with available administrative units. However, for unaggregated data, a smaller administrative unit is required. In Indonesia, the smallest administrative unit commonly used in the administration affairs is Neighborhood Association (Rukun Tetangga, abbreviated as RT). Unfortunately, base maps for this administrative entity are currently absent in ISDI. Meanwhile, mapping RT is a challenging work considering the extent and uncertainty with the boundaries.

To contribute on finding solutions, the objective of this study is to explore the possibility of developing RT base map while examining its use for linking PAD as a prospective big data source for spatial based research and analysis. A source of data is proposed for developing RT base maps correspond to the need of linking PAD at a level of unaggregated databases. The data source is the village sketch map population census. That is a back-office data, used to support surveys in decennial agenda in the country.

2. Materials and Methods

This study uses multiple methods and multiple data. The analysis is divided into two main parts; base map development and linkability assessment. Data for the development of a digital metric base map of RT administrative boundary are village sketch maps of population census. The sketch maps were generated by Statistics Offices prior to decennial events of population census. Meanwhile, data for linkability analysis are unaggregated databases from local government offices.

Framework was developed based on understanding of Indonesia's data aggregation structure and their relationship to particular geographical area which can be understood from a three-dimensional relationship (see Figure 1). As depicted from the figure, hierarchy of government/public institutions downs from central to local governments. In some extent, they can reach the lowest level of government hierarchy; that is village level. In each level of hierarchy, sectors govern their jurisdictions within each determined area. However, there is also an institution specially established to cope with government's data. It is Indonesia Statistic Office. The institution has a structure vertically stretched from central level to sub-district level. Often, it has representatives working at a village level. The institution produces various types of data for national interest, but also not uncommon to cooperate with local governments on producing data for local needs.

Data aggregation gives data users multi-scale perspective on data and information for users beyond levels of government. However, unaggregated database gives optimum advantage as it has the highest resolution. An unaggregated database is a database in its original form. Usually, the scale is in the smallest spatial units of the policy interest.

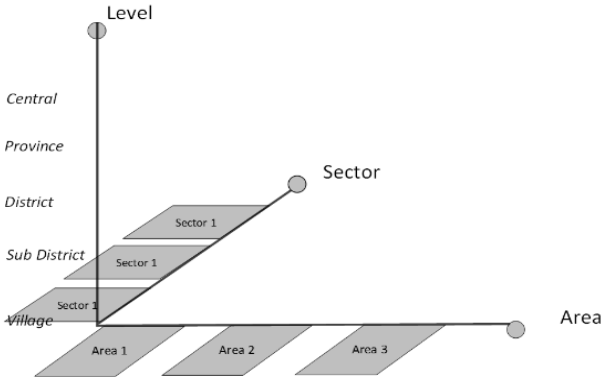

Fig. 1. PAD and geographical areas in a three-dimensional perspective

2.1. Framework

Two main challenges on spatially linking government's unaggregated databases are the absence of appropriate base maps and inadequate existence of geolocation identifiers in the databases. To address these challenges, this study investigates the linkability of government's unaggregated databases to the smallest administrative area in Indonesia (RT). The investigation is derived from a framework as shown in Figure 2. From the framework, it can be seen that GI is produced by linking PAD with SDI. This system downs from central government into lower levels of government structure until the community or the individual level.

The potentiality of using PAD for geographical based analysis by linking the records to RT's area was investigated and clarified. In line with the objective, a new base map of RT was developed and is proposed as a new feature within ISDI systems. Indonesia Population Census's village sketch maps are proposed as a data source for developing the base map.

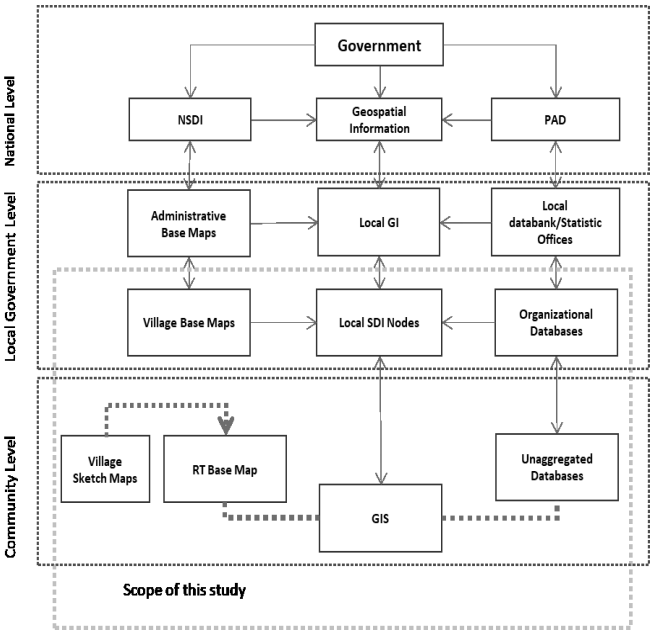

Fig. 2. framework of Linking PAD to the Neighborhood Association area and scope of this study

2.2. The Data

2.2.1. The village sketch maps of population census for developing Neighborhood Association's Administrative Boundary Metric Base Map

As aforementioned, developing metric base map of RT is a challenging task considering uncertainty with the delineation of the boundaries. However, for research purposes, this study proposes an alternative source of data. Village sketch maps of population census are expected containing such valuable information for developing the base map.

Village sketch maps in principle can be obtained from local Statistic offices which conducted population census. The sketch maps are usually generated nation-wide in line with the scale of population census. It means the data should be available for whole villages in the country. However, In Indonesia, the sketch maps were used as back-office data and therefore had never been published. They were used as a supporting tool in the survey of population census.

The sketch maps usually can be obtained in scanned image format. The population census is a decennial agenda in Indonesia. The last population census was conducted in 2010.

Guidelines were published prior to the census imposed to all statistic offices in the country. Mappers were trained and ensured to follow the guidelines in the mapping process. Based on this, it can be assumed that the data are uniformed, in which all scenarios of possible differences in the vast country are well accommodated. Therefore, one case study should be enough to make a generalization for the whole country.

Based on the guidelines, there are three scenarios in developing village sketch maps: first, a sketch map developed from satellite image with a coordinate system; second, sketch maps developed from a map with a coordinate system; and third, a sketch map developed from statistic sketch maps without coordinate system [10]. Despite having slight differences, in general, there are three components in the village sketch maps: the header and footer, sidebar, and the main map. All the components are important for the digitizing process.

On the header, there is information about the identity of the sketch map. The name of the village can be found on this part. Also, can be found there, is the identity code for the village and the sketch map. This information is helpful particularly for locating the sketch in the georeferencing process in which the sketch maps must be brought into the correct corresponding village in the existing administrative base maps. On the footer, validation sign can be seen. The information is required to ensure that the sketch map has been verified by the authorized officer which means the sketch map process has followed the guidelines.

On the sidebar, there are information sections including administrative structure section, scale and legendary section, statistic section, and validation section. All sections give valuable information for understanding the sketch map contents. In the statistic section, for example, there is information regarding existing features in the main map. While in the validation section, information about the person who drew the sketch maps are indicated.

On the main contents, the shape of the feature is drawn. There is information about coordinates, names of geographic features, and various administrative boundaries as well as important landmarks. This part is the main target in the digitization. Three elements in the main map are very important besides coordinate points. Topology names, boundary lines, and existing landmarks.

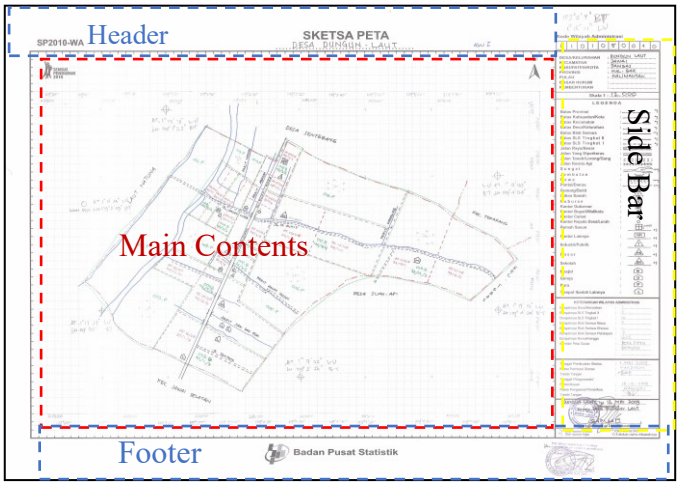

Fig. 3. One of village sketch maps of population census 2010

2.2.2. Unaggregated databases for linkability assessment.

An unaggregated database is a database in its original form. The records are not yet aggregated. Aggregation of PAD is usually based on administrative systems. Unaggregated databases can be collected from different local government offices.

Geolocation identifiers in unaggregated databases are usually in a very spesified location such as landmark positions, house numbers, road's names, and the smallest administrative units. In Indonesia, the smallest administrative units are the most common used and formally required in various administrative affairs in Indonesia.

In this study, because all collected databases are relational databases, therefore the examination and analysis of the databases are conducted using common Structured Query Language (SQL) process. The focus is to investigate how RT exists within the records. The linkability of the records to the RT, as a geolocation identifier, is determined by how RT can be appropriately related to the records in the databases. As a simple rule, a record is categorized as un-linkable to the RT if the data row finds no RT on its columns or RT cannot be appropriately identified whether because of mistyped, unknown RT identity (numbered 0), left blank or just unavailable (N/A).

RT as geolocation identifier usually stated following its identity which is always in a set of numbers. Example of RT's identity can be seen from an identification of a house in Figure 7. As it is shown, RT is always written in a three-digit numbers. Often it is written followed by its upper-level unit; RW (Rukun Warga/Community Association). So that it is commonly stated as RT/RW: 001/01, which means RT number 001 in RW number 01 (see Figure 7).

2.3. Method for developing New metric base map.

2.3.1. Georeferencing the village sketch maps without coordinates system

Village sketch maps are products of mappers' cognitive maps in which observation is used instead of measurement. Therefore, the information represented in sketch maps is typically distorted, schematized, incomplete, and generalized [11]. Therefore, before digitization, the sketch maps should be georeferenced to put it in the right position on existing metric base maps. As aforementioned, village sketch maps are made using pre-existing base maps: satellite image with coordinate systems, map with a coordinate system, and statistic sketch maps without coordinate system [10]. For the first and second type, the coordinates can help. Meanwhile, for village sketch maps without a coordinate system, they can be manually georeferenced based on shapes and segments in the sketch maps.

Appropriate transformation method is required to reach good accuracy in georeferencing. There are many methods for georeferencing. For example, in Quantum GIS (QGIS), there are seven transformation methods available in the application.

From all seven available transformation methods, Helmert method gives higher accuracy as well as simplicity. Therefore, this method is recommended for the transformation. Mathematically the Helmert 2D transformation described as:

(Equation 1)
$$\begin{bmatrix} X \\ Y \end{bmatrix}^B = \begin{bmatrix} T_x \\ T_y \end{bmatrix} + s \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}^A$$

Scale (s) in the transformation is unitless. This type of transformation method works better in this case because the method as seen from the mathematical description, only rescaling and rotating the images. Since the village sketch maps mostly developed from previous existing maps, the rescaling and rotating are the only needed techniques to fit with existing metric base maps.

2.3.2. Digitization and alignment.

Digitization is a process of transforming the sketch images into a metric geodatabase. Digitization should be completed one by one for each village sketch maps. Further, because the information represented in the village sketch maps are typically distorted and schematized, alignment is required. Otherwise the maps will highly deviated.

The first step for the alignment is placing all developed base maps into one bigger frame. It can be by sub-districts or multiple villages. In this case, because it is assumed that the villages boundaries are correct, the newly digitized base maps are combined on multiple villages. All RTs are firstly aligned into their corresponding villages and then aligned within each village according to the sketch maps proportion. This simply mosaicked all the digitized RTs boundaries.

Several geodatabases can be used for the alignment. Existing metric base maps obtained from local government, for instance, are very helpful. The newly digitized base maps aligned following the existing metric base maps and should be in this order in its priority. The existing metric base maps are:

a) Administrative boundary base maps, including village administrative boundaries, sub-district administrative boundaries, regency administrative boundaries.
b) Databases of the location of government buildings, facilities, offices, and other landmarks.
c) Geodatabase of roads, railways, and other infrastructures
d) Geodatabase of rivers and waterlines.

The order means, the digitized RT boundaries firstly aligned following existing administrative boundaries. Next, the boundaries must be aligned to each other to make sure all found landmarks such as education buildings and health facilities fall into corresponding RTs as indicated in the village sketch maps. After the that, the boundaries need to be aligned again following the roads and water lines.

The reason why the administrative boundary becomes the first alignment priority is, in order to make sure that the new base maps can be mosaicked into villages properly without serious overlaps. The village boundary is the frame for all of RT boundaries in a village. Meanwhile, landmarks are the most accurate way for mappers to locate the positions of RTs. Therefore, they should be in high order to locate RTs. Roads and other networks of infrastructures are objects for locating RTs position but tend to be distorted, schematized, incomplete, and generalized.

Rivers and waterlines are also geographical objects for locating RTs similarly to roads. However, it can be assumed that roads are more accessible to mappers. The shapes and positions of roads should be more familiar to mappers; therefore, they should be prioritized before rivers and waterlines.

Unless the village sketch maps are drawn carefully following accurate maps or satellite images/aerial images then the order is important to ease the process. Illustration of the alignment procedure can be seen from Figure 4.

A is village sketch maps after digitization into a geodatabase where RTs boundary just follow the sketch map. B is RTs boundary after alignment where four required existing metric base maps are followed. RT 001 becomes larger as it follows the position of a landmark in the existing metric base map. RT 002 move to the right to follow the road's edges, and RT 003 narrowed as the consequence.

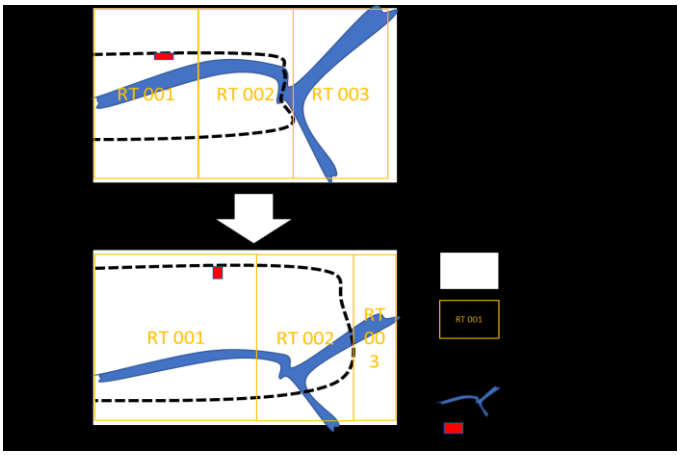

Fig. 4. Illustration of alignment with landmark, road, and waterlines

In a case where roads are unavailable, the alignment should follow the next order (rivers). As illustrated in Map C, RT 002 is narrowed to follow the river's junction, and the rest of the area is given to RT 003. Meanwhile, if a landmark is the only alignment tool, the first step is to make sure that the landmark falls proportionally into the correct RT as indicated by the village sketch map. Later the other RTs will proportionally share the rest of the area.

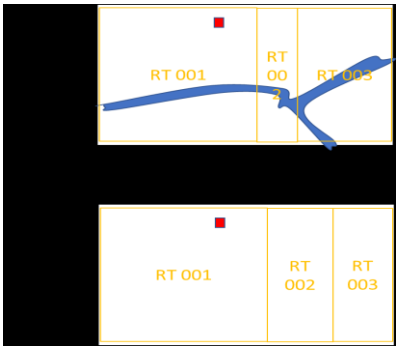

Fig. 5. Illustration of alignment with landmark and road (C) and with landmark only (D)

3. Results and Discussion

3.1. Case Study in Sambas Regency of West Kalimantan Province, Indonesia

To test the applicability of the framework and method, a case study was conducted in Sambas Regency. It is a third level sub-government in Indonesia. This Regency currently has 193 villages as its lower sub-government. However, in this case study, only 30 village sketch maps were processed for base map development. These villages were chosen because they are related to our current and future researches.

Sambas Regency is located in the northern part of West Kalimantan Province between 0°57'29,8° to 2°04'53,1° North Latitude and 108°54'17,0° to 109°45'7,56° East Longitude. Sambas Regency has 6,395.75 square kilometer area, which is 4.36 percent of West Kalimantan total area.

Table 1: Collected databases from local government in Sambas Regency

Database Name	Original Administrator	Database Platform
Health Insurance for the poor	Dinkes	Microsoft Excel
Population Registry 2015	Disduk	Oracle Database
Development proposal 2015	Bappeda	Microsoft
Land Taxation 1988-2017	Bakuda	Oracle Database

For linkability assessment, four unaggregated databases were collected and investigated. The four databases are a representation of PAD categories as described by Wallgren and Wallgren [12]. First, Population registry and Health insurance database are actually statistical data produced by an authority for their own purposes. Land Taxation is a database containing variables which are legally important. Meanwhile, development planning proposals are a category of variables representing decisions.

From the four databases, only one database (database of development proposal 2015) is categorized as open public database. Meanwhile, all other databases are protected by privacy and therefore the handling is carefully carried out.

Health Insurance for the poor is a database administered by the local health department. The database contains the identity of the beneficiaries of free health insurance for the poor program. Population Registry is a database from famous Information System for Population Administration. The database is a copied version obtained from the local Registry Office. In this database personal identities including the address of each individual registered as the citizen of Sambas Regency are recorded. Meanwhile, the development proposal database is a database contains the information of development projects proposed by citizens to the local government of Sambas Regency.

The three collected databases are for the year 2015 only. Meanwhile, one database named the Land Taxation Database has accumulated records for the year 1988 until 2017. However, the records are obtained for 4 subdistricts only out of 19 in the Regency.

Five geodatabases were used for the alignment. Existing metric base maps obtained from local public institutions. Following are the existing metric base maps in priority order:

a) Administrative boundary base maps, including village administrative boundaries, sub-district administrative boundaries, regency administrative boundaries. These databases developed in 2006. There is no clear explanation on the definitive year to refer for these base maps.
b) Education buildings and facilities database. This database was developed in 2008 until 2009
c) Health facilities, this database also developed from 2008 until 2009.
d) Geodatabase of roads
e) Geodatabase of rivers and waterlines.

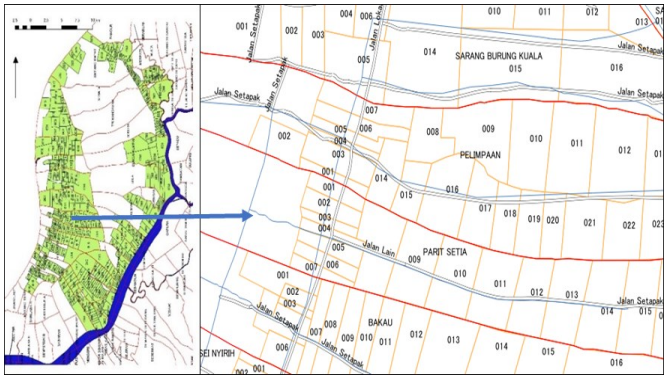

Fig. 6. New developed RT base map after integration to the existing local government's base map.

3.2. Results

3.2.1. New RT Base Map base on Village Sketch Maps of Population Census.

New Developed RT base map as a result of the process in this study has been published as an open dataset on Mendeley Data [13].

As it was expected, at first stage, digitized village sketch maps resulted in highly deviated features. There are three

sources of deviation. First, the deviation comes from transformation. The second, the deviation comes from changing village boundaries itself, that is when the village boundaries changed and cause differences between the sketch maps and the metric base maps. The consequence of the deviation is overlapping boundaries. However, it is found that not all of the overlapping comes from digitization deviation. Some may come from territorial disputes

Quantum GIS (QGIS) was used as the main application for processing the sketch maps. This application was chosen because it is free and open-source software. Hopefully, other researchers or local government officials can easily download it from the internet thus making them easy for replicating this work.

Statistical Summary of the base map can be seen from an output of the Basic Statistic process in QGIS as follow:

Table **3.** Statistical Results from QGIS Application

1	Count	458
2	CV	1.247
3	Empty	0
4	Filled	458
5	First quartile	20.9
6	IQR	66.3
7	Majority	10.1
8	Max	710.16
9	Mean	76.001
10	Median	47.44
11	Min	1.01
12	Minority	1.01
13	Range	709.15
14	Std_Dev	94.794
15	Sum	34808.470
16	Third quartile	87.2
17	Unique	456

For 30 village sketch maps digitized in this study, as many as 458 RTs were found. On average, the area of RTs is around 76 hectares per RT. With the smallest area is around 1 hectare only with the majority has 20 to 90 hectares area. The largest RT in this study reaches an extent of more than 700 hectares. This size is similar to a village in some densely populated areas.

The total area of all 30 villages is 59.280 hectares, but the total area of RTs in those villages is around 34.808 hectares. Therefore, there are more than 24.000 hectares area of the villages which are not associated with any particular RT. The blank areas can be a forest, agricultural fields, or just bare lands. For example, a big blank area surrounded by the digitized RTs shown on the map (Figure 6) is actually a protected forest. This forest has a size of around 15.000 hectares.

3.2.2. Linkability of PAD to RT as geolocation identifier

There are 1,337,028 rows of records found from the four collected databases (see Table 4 below). In total, 94.63 percent of the rows are linkable to RT as geolocation identifier. Population Registry 2015 is the highest among others in linkability. 99.48 percent of the rows in the database directly linkable to corresponding RT. Health Insurance for the poor 2015 which is the second highest, has 99.22 percent linkability. Land taxation database which is a collection of data from the year 1988 until 2017 has almost 83 percent linkability. Meanwhile, development planning proposal database for the year 2015 has only 27.06 percent.

Table 4: Percentage of linkable records to RT as geolocation identifier

No	Name of the Database	Records	Missing link	Percentage
1	Health Insurance for the poor 2015	374,577	2,935	99.22%
2	Population Registry 2015	599,919	3,122	99.48%
3	Dev. Planning Proposal 2015	6,902	5,034	27.06%
4	Land Taxation 1988-2017 (4 sub districts)	355,630	60,750	82.92%
	Total	1,337,028	71,841	94.63%

3.2.3. Field surveys for accuracy assessment of the new developed base map.

In order to understand the accuracy and applicability of the newly developed base map, a field survey was conducted in which exact ground objects coordinates were collected. Coordinate of ground objects where RT's identity can be clearly defined such as house number and statistics stickers, border statues, and any other signs were recorded (Eg. Figure 7).

Fig. 7. RT's identity on a house wall

Sample are categorized into three types:

1. Random points of RT identification,
2. Exact points of village border,
3. and the exact point of RT border.

For the first type, as many as 53 samples were collected from 15 out of 30 villages. The results show that against digitized base map (before alignment), 51% of the samples fall within false RT meanwhile 49% fall into correct RT. The same samples overlayed against aligned base map show that 75% fall into correct RT which is significant accuracy improvement.

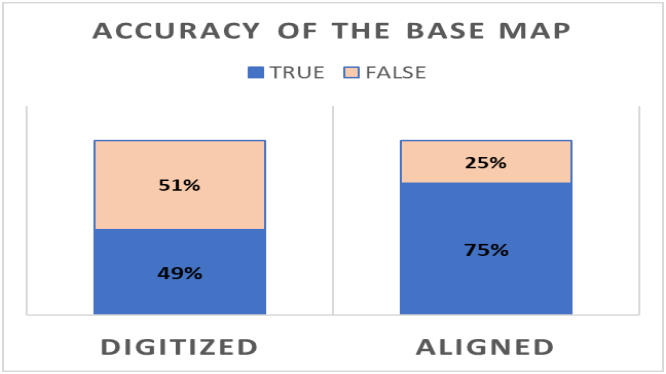

Fig. 8. Improvement of the accuracy after alignment

For the second type, 24 samples were collected from 19 villages. Averagely, the borders on the map deviate around 529 meters from the sample points taken on the field.

Meanwhile, for the third type, ten samples were collected from eight villages and the measurement show averagely 45 meters deviation of RT borders on the map with a coordinate of collected points. All of the samples for exact point show highly accurate relative position to landmarks, roads, and water lines (below 10-meter accuracy which is the precision of used GPS tool).

3.3. Discussion

3.3.1. Spatial accuracy of the new developed base map.

The new developed base map shows that RT averagely gives much smaller spatial unit of analysis than the village level. However, despite significant improvement after the alignment, the accuracy still has a lot of space for improvement. As mentioned on the results of the field survey, 24 percent of randomly sampled points for signs of RT on the ground fell into wrong RTs. In fact, the problem of accuracy is on the existing village metric base map obtained from the local government. The deviation of the borderlines as figured out on the field study is 529 meters which is quite much. The new developed base map reached such level of accuracy was helped by the accuracy of landmarks, roads, and water lines. It is likely that if the village metric base map was improved, the results should be more accurate.

However, with an average deviation of RT's boundaries around 45 meters from the exact location, the base map can be used for the social or spatial analysis requiring fair accuracy. which is good, considering the accuracy of the existing village metric base maps is low.

Most of the RTs' area are in almost perfect square indicating artificial shapes. It means that natural elements are not much used in determining RT's boundaries. But also, can be understood as a lack of certainty with the borderlines as straight lines might be a result of schematization in the drawing of the village sketch maps.

Moreover, from the above statistical summary (Table3), it can be figured out that the size of RT's areas varies with almost 125% variability as shown by the Coefficient of Variance (CV). The variation means that in terms of geographical area, RT cannot be seen as a custom size.

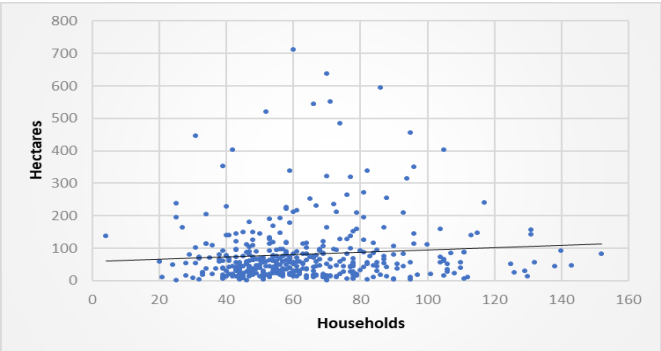

Fig. 9. The relationship between RT's Area and its households

Relationship between the number of households of RTs with their area can be seen from Figure 9. From 30 studied villages, their RTs have averagely 39 to 87 households for each. Some RTs have more than 100 households. Considering the random relationship of the area and number of households, using RT as data aggregation can help deidentifying individual records, which is important for protecting personal privacy. On the other hand, its average size which is 1/26 times smaller than the village average size as shown in this study, can give a balanced value between spatial accuracy and personal privacy protection.

Meanwhile, despite in principle, RT is a population-centered administrative unit instead of territorial one, the relationship between population/social components and geographic/spatial components of an RT can be drawn as figure 10.

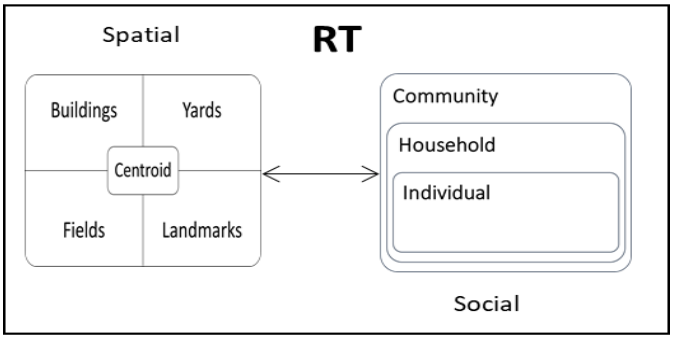

Fig. 10. The relationship between spatial and social components of an RT

RT is a polygon in shape; therefore, a centroid exists. Centroids are helpful for spatial analysis considering uncertain legal boundaries. Buffering, for example, can be done using centroids of the RTs.

Buildings are houses and other constructions where people live or do their activities. Some buildings have a yard which is open space complementary to the buildings. Sometimes, open space such as agricultural fields closely accessible to the population in an RT identified as a territorial area of corresponding RT. While landmarks can be used as unique objects to identify an RT territory.

The population is the members of an RT that can be analyzed in three levels; individual level, household level, or community level. In an analysis, spatial aspects and social aspects of an RT are exchangeable depend on the focus of analysis. For example, RT's area (centroid) can be used as a representation of its population position against a deforested area. Meanwhile, individual health records can be used as a representation of water pollution in an RT's territory where she/he lives.

3.3.2. Linkability and approaches for using RT as a unit of analysis.

High linkability in the population registry is a very good opportunity for linking PAD. Because, it means all other population-based databases, as far as containing population-based records such as individual identification number or household identity number, can be linked to RT. This also means that RT and population registry can be an

interchangeable platform for human-spatial/human geography analysis.

Another good point about population registry is that, despite administered by a local government institution, the database actually product of a national-scale information system. The system is known as SIAK (Sistem Informasi Administrasi Kependudukan/Information System for Population Administration).

Development planning proposals for the year 2015 has the lowest linkability because RT is not required in its application form. In this database, most of the records un-linkable to RT as geolocation identifier. Yet, around 27 percent of the records are linkable. With the percentage, there are a lot of spatial information can be gained from the database. Because the database is updated on an annual basis and contains a lot of records. With purposive policy, the percentage should be increasable.

Based on the result of the database analysis, there are two ways to cope with the un-linkable records. Distribute the value of the records to other linkable ones using a population-based approach is the first one. In this approach, the un-linkable records are proportionally distributed into other records based on population distribution. High linkability in the population registry as found in this study is a good point for this case. Because there are so many databases that actually linked to population data. As a consequence, even with the absent of RT information in the datasheet, a database can be linked to RT Base maps, as far as it is linkable to population database. Furthermore, the database can be spatially analyzed.

The second approach is the spatial distribution approach. In this approach, the value of un-linkable records is proportionally distributed to the other records based on their spatial share of RTs in their higher-level administrative unit. The second approach is basically the same approach as the normal linking of PAD using administrative boundary as the

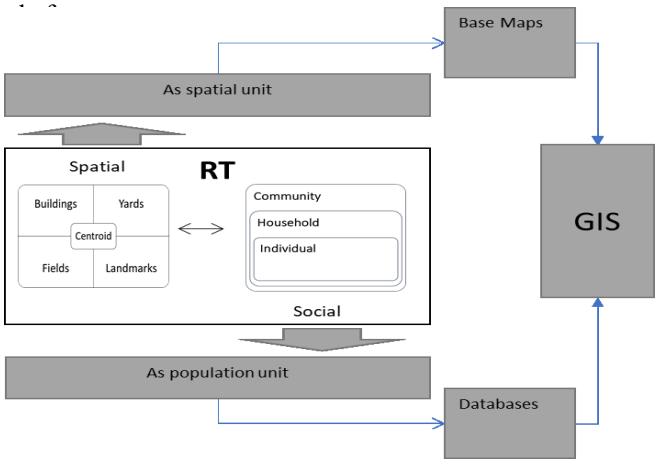

Fig. 11. Two approaches for using RT as a unit of analysis in GIS-based analysis

It is suggested that the best approach for analysis depends on the main interest of the research. If the research is more focus on human or population, then population-based distribution should be better. Meanwhile, if the research is more focus on spatial dimension or geographical issues, then spatial distribution should be better.

3.3.3. Using neighborhood association area as new spatial data infrastructure to link public administrative data with GIS and lessons learned.

In this study, the new developed RT base map as a spatial data infrastructure was applied for mining and analyzing spatial data from the collected databases. The relational model of the data can be seen from below diagram. The data were collected and compiled as a new database in which RT's unique identification numbers were used as a primary key.

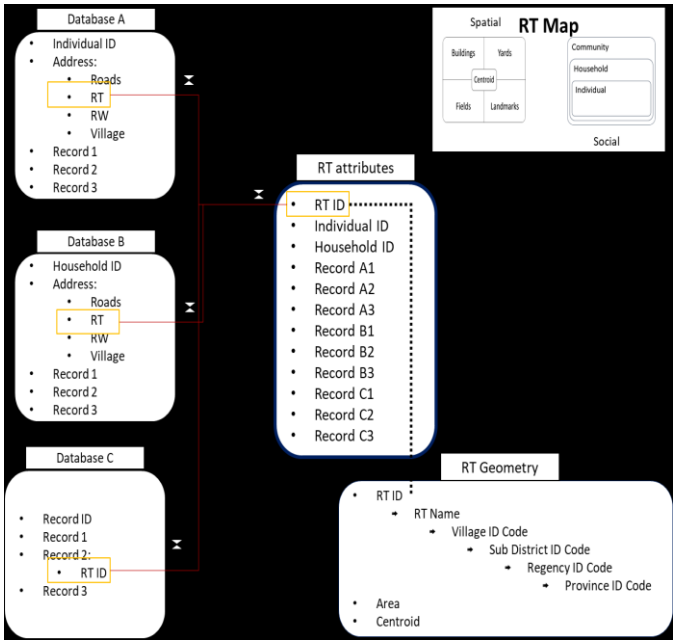

Fig. 12. Data model used for linking the databases and the base map.

Example of the data visualization can be seen from Figure 13 and Figure 14. Figure 13 below shows the distribution of land registration tax in 2011. Data from databases were directly linked to the base map following the data model. Map B in Figure 13 is the distribution of land registration by RT, and Map A is the same data aggregated and linked to Village Boundary Base Map from Local Government.

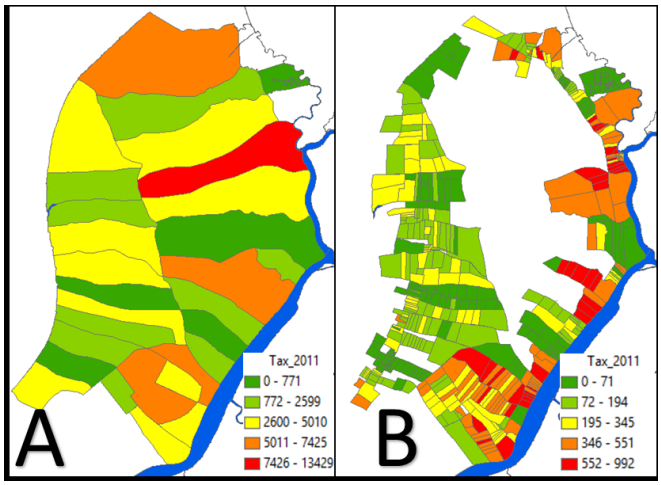

Fig. 13. Tax Registration database visualized in village boundary base map (A) and RT boundary base map (B)

As aforementioned, the smallest administrative unit boundary base map such as RT Base Map gives a higher

spatial resolution. Figure 13 shows that using the base map, a different perspective of spatial visualization of data can be presented. Therefore, it is useful for coping with Modifiable Areal Unit Problem (MAUP). MAUP is a biasing effect in density analysis commonly happened on a larger scale of GIS analysis.

On the other hand, the land tax registration combined with population distribution presented in density maps give another understanding as can be seen from Figure 14. RT base map (Map B) shows a more realistic distribution than village base map (Map A) considering the existence of protected forest (red line) in the area. Therefore, using RT base map is more realistic for analyzing the relationship between social variables (Eg. population and tax) with deforestation in this case.

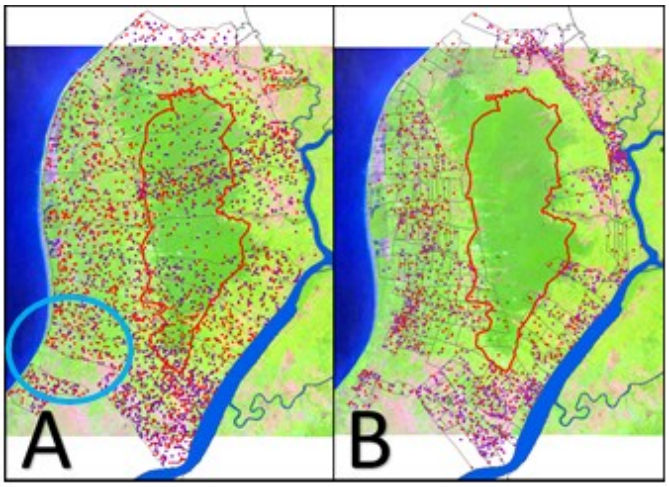

Fig. 14. Population density and Tax registration density visualized in village boundary base map (A) and RT boundary base map (B)

Better visualization of population distribution pattern in a spatial extent can also be depicted from Figure 13. For instance, how population tends to disperse over villages on the west part of the area and tend to concentrate on the east side, can be apparently distinguished from Map B in Figure 13. Using village as the unit of analysis, such pattern is hardly detected (Map A). The capability of detecting such tendency is important in studying various ecological dan spatial phenomena.

Meanwhile, despite method for the development of the base map from this case study is expected to be applicable nationwide considering the data uniformity, in some extent, use of the base map is unique for this case study. Use in other areas may need some modifications depends on local circumstances and research uniqueness. However, some lessons were learned from using the base map which is important. One of them is that there is a village where the RT's identity cannot be linked to the databases (see blue circle in Map A on Figure 14). The reason is that the RTs' numbering system on the base map is not in line with the databases. Names of RTs in the village were unique to RW (One level upper administrative unit). It should be unique to their Village which is used by the databases. It seems that, while other villages adopted the numbering system before the made of the village sketch maps, the village did it later. The key lesson on this point is that the linkability of the data to the base map is not merely determined by the existence of geolocation identifiers in the databases but also by the base map itself. This is in line with missing links in the databases likely resulted from the establishment of new RTs. This case, despite unique for this study, may also happen in different areas in the country considering less attention given to RT mapping. In fact, this case supports the third condition for the usability of the base map as discussed in Chapter 2.2 above.

The other lesson learned is on using confidential data. Tax registration database and population registry database used in this study are public databases containing personal records protected by the privacy act. Using RT base map as shown in Figure 13 and Figure 14 indicates that the personal data can be de-identified and analyzed with less compensation of the spatial accuracy. In this study, the average size of RT is 0.04 % or 1/26 times smaller than the average size of villages. In other areas, it can be different. But it can be safe to assume that generally, the size can give a balanced value between spatial accuracy and personal privacy protection.

Finally, using the base map, there are many public databases containing confidential records can be mined and analyzed for various spatial-based research and analysis, as shown using the approach in mining tax and population registry of this study. Rapid improvement on data management adopted by government institutions in Indonesia is a prospective spatial data mining opportunity. The base map is a geocoding instrument for mining such data, which is one of the essences of administrative boundaries base map in the National Spatial Data Infrastructure

4. Conclusions

Despite the uncertainty with their legal boundaries, developing metric base maps for neighborhood associations' administrative boundaries in Indonesia using village sketch maps of population census as a data source is feasible, at least, as far as it is aimed for research purposes. The proposed method of georeferencing and alignment for transforming village sketch maps of population census into a metric base map, smoothly resulted in a new developed RT metric base map with a fair spatial accuracy which is improvable should the accuracy of other existing geodatabases used in the alignment is higher.

This study also finds that despite in principle RT is designed as a population-centered administrative unit instead of spatial one, the relationship between the population/social components and geographic/spatial components of an RT can be drawn for GIS-based analysis (Figure 10). Based on the spatial and social components, the RT base map has potential use as a feature for linking and mining spatial data from unaggregated databases of government institutions (PAD). The high presence of RT as geolocation identifier in some databases is a good opportunity for such purpose. On the other hand, using RT as a base for data aggregation can help deidentifying individual records, which is important for protecting personal privacy.

Furthermore, the usability of PAD for GIS-based research and analysis in Indonesia is highly prospective

considering the high linkability as found in this study. Current improvement on data management adopted by most of government institution in Indonesia is additional advantage. If the databases can be mined then it can be used for monitoring human-environment reciprocal impacts. However, several challenges are there. Data quality is one of the obvious challenges besides the availability. The quality here means reliability and validity of the data. Therefore, it is recommended that Public institutions multiply efforts to improve their data quality. Standardizing how the data generated can be one of the viable options. Including on how to increase the presence of RT as geolocation identifier in the PAD. One Map Policy and One Data Policy can be prospective policies for linking PAD geographically.

Improving data or record quality alone will not enough to enhance usability. Maintaining appropriate base maps as features of SDI is as important as improving the data/record quality. As proposed in this study, RT base map can be included as part of the ISDI's feature for better linking and mining spatial data from PAD. For such purpose, RT as the smallest administrative entities need to be mapped properly and regularly following the changes of data aggregation in the administrative system.

Acknowledgments

Our sincere gratitude to the Taikichiro Mori Memorial Research Fund 2016 and Keio University Doctorate Student Grant-in-Aid Program 2017 for supporting this study including several fieldworks and surveys.

References

1. D. Figlio, K. Karbownik, KG. Salvanes. *Education Research and Administrative Data*. In: Handbook of the Economics of Education. 2016.
2. Canada Statistics, "*Use of administrative data.*" Internet: http://www.statcan.gc.ca/pub/12-539-x/2009001/administrative-administratives-eng.htm.
3. A. Sexton, E. Shepherd, O. Duke-Williams, A. Eveleigh, *A balance of trust in the use of government administrative data*. Arch Sci. 17 (2017) 305–330.
4. R. Connelly, CJ. Playford, V. Gayle, C. Dibben, *The role of administrative data in the big data revolution in social science research*. Soc Sci Res. Elsevier Ltd. 59 (2016) 1–12.
5. L. Rivas, J. Crowley, "*Using Administrative Data to Enhance Policymaking in Developing Countries: Tax Data and the National Accounts*." 2018. Report No.: WP/18/175.
6. D. Gunawan, A. Amalia, *The Implementation of open data in Indonesia*, *Proc 2016 Int Conf Data Softw Eng ICoDSE, Denpasar, Indonesia, 2016, pp. 1-6*
7. J. Goodwin, C. Dolbear, G. Hart, *Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web*. Trans GIS. 12 (2008) 19–30.
8. K. Bretz, *Indonesia'S One Map Policy: a Critical Look At the Social Implications of a "Mess.",* Senior Thesis, University of South Carolina, USA, 2017.
9. F. Hasyim, H. Subagio, M. Darmawan, *One map policy (OMP) implementation strategy to accelerate mapping of regional spatial planing (RTRW) in Indonesia*, IOP Conf Ser Earth Environ Sci. IOP Publishing, Kuala Lumpur, Malaysia, 2016.
10. BPS, *Pemetaan SP2010: Pedoman Pemeta Desa*, Jakarta, Indonesia: BPS, 2009.
11. A. Schwering, J. Wang, M. Chipofya, S. Jan, R. Li, K. Broelemann, *SketchMapia: Qualitative Representations for the Alignment of Sketch and Metric Maps*. Spat Cogn Comput. 14 (2014) 220–254.
12. A. Wallgren, B. Wallgren, *Register-based Statistics: Statistical Methods for Administrative Data: Second Edition.* New Jersey, USA: John Wiley & Sons, Ltd., 2014, pp. 25-28.
13. Arip M, *RT Administrative Boundary Base Map of Sebubus Forest Area Developed from Village Sketch Maps of Indonesia Population Census 2010*. Mendeley Data v1, 2019.