Semi-Automated Land Parcel Plotting: a Machine Learning Approach Based on Geospatial Data Matching

MUHAMMAD GHALY KURNIAWAN
June 2024

SUPERVISORS:
Dr. A.M. Pinto Soares Madureira
Dr. D. Todorovski

PROCEDURAL ADVISOR:
Dr. M.C. Chipofya

Semi-Automated Land Parcel Plotting: a Machine Learning Approach Based on Geospatial Data Matching

MUHAMMAD GHALY KURNIAWAN
Enschede, The Netherlands, June 2024

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformation Management for Land Administration

SUPERVISORS:
Dr. A.M. Pinto Soares Madureira
Dr. D. Todorovski

PROCEDURAL ADVISOR:
Dr. M.C. Chipofya

THESIS ASSESSMENT BOARD:
Dr. M.N. Koeva (Chair)
Prof. Ir. Trias Aditya K.M., S.T., M.Sc., Ph.D., IPU (External Examiner, Universitas Gadjah Mada)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

The Indonesian cadastral quality improvement process, established in 2018, faces huge challenges due to its huge land area, diverse natural and social conditions, and past mapping practices. With 15 million land parcels yet to be plotted accurately, the manual methods prove insufficient which may be risking land conflicts. This research proposed a solution based on machine learning to semi-automatically search the location for plotting land parcels using geospatial data matching.

The research identified eight causes of unplotted parcels, the issues of being tied to the local control points and incomplete information on the available documents appear as the main causes. The machine learning model was built based on the manual plotting method, focused on the studio process. Five geometric matching variables were optimized using the RCGA optimization algorithm. By combining those with the several strategies of textual attribute matching, the candidate location to plot the parcels will be identified.

The model's performance was evaluated using two datasets: The Sample Data and the KKP Database. In the first test, the model achieved a precision value of 98.75% and a recall value of 88.27%. The second test, involving the real condition Indonesian cadaster data which is more complex, generates a lower precision value of 91% and a recall value of 57%, due to issues like the homogeneous shape of candidate locations and overlapping rights. To enhance the model performance, textual matching was used with the best result of using the unique attribute of land parcels such as registered areas. It resulted in the improvement of recall value to 91% and the consistent results of the precision value of 92%.

The machine learning model based on geospatial data matching to find the location of unplotted land parcels is novel due to the limited use of the methodology in certain fields. This approach can accelerate the current process of cadastral quality improvement in Indonesia, especially in the city that has been declared as the "Kota Lengkap". For future research, it is recommended to test the model's performance in different cities to evaluate its accuracy across varied cadastral data conditions. Combining several automation techniques in converting analog to digital data could also enhance the model performance. Creating a fully automated model for plotting the unplotted land parcels with limited position information.

KEYWORDS: Land Administration, Cadastral Quality Improvement, Machine learning, Geospatial Data Matching, RCGA Optimization Algorithm, Unplotted Land Parcels.

ACKNOWLEDGMENTS

Firstly, I would like to say thanks to the LPDP and the government of the Republic of Indonesia that giving me opportunities to continue studying abroad and giving support during my academic journey in the Netherlands.

I would like to express my heartfelt gratitude to my supervisors, Dr. M.C. Chipofya, Dr. D. Todorovski, and Dr. A.M. Pinto Soares Madureira for their support and guidance throughout my thesis work. Your input and encouragement helped me to complete this research. I would like to express my sincere gratitude to Prof. J. Zevenbergen for chairing my proposal and mid-term presentation. Your insightful feedback contributed to improving the quality of my thesis. Additionally, I would like to sincerely thank Dr. M.N. Koeva for becoming my chair during the final defense, your feedback was invaluable.

I would like to extend my sincere thanks to all the GIMLA lecturers who gave me knowledge on land administration. Your teachings have shaped me to think critically about a new solution to solve the current issue in the land administration.

I also express my gratitude to the Ministry of ATR/BPN, especially the Directorate of Cadastral Survey and Mapping who helped me a lot during my research in Indonesia. Additionally, I am thankful to Prof. Trias and his team from Universitas Gadjah Mada for their support during my fieldwork in Yogyakarta and to the Surabaya II Land Office for their support in Surabaya.

To the class of 2022 thank you for the stimulating discussion and help during our time at GIMLA. Thank you to all my Indonesian Plus Family: Andi, Clava, Nasir, Rifqi, Wibi, Amy, Sry, Salsa, and Mas Ganda for the laughter, food, and warming our house to make it like home. I would like to thank the board of PPI Enschede 2023 who have worked together in managing the organization. To my Indonesian neighbors in Enschede for their kindness and support to my family during our stay.

I sincerely thank my parents and my brother for their prayers and support in my life. Lastly, but most importantly, I would like to say thank you to my wife, Aprodita Fiska, and our little hero Qael Kurniawan for accompanying me during my study in Enschede. Fiska, your hard work and motivation have helped me through the hard times, and for Qael, I hope to inspire you to dream big and achieve the seemingly impossible.

من صبر ظفر
He who endures, triumphs.

من جدّ وجد
He who strives, succeeds.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY OF TERMS

Buku Tanah/Land Book : A document in the form of a list that contains the juridical and physical data of a certified land parcel.

Flying Parcel : The other term unplotted land parcels refers to parcels that are floated somewhere else because doesn't have the position information

Gambar Ukur/Title Plan : A document that contains the maps of a parcel with its surrounding features from a certain process of measurement.

IP4T : Activities to collect data on control, ownership, use, and utilization of the parcels, which is processed using the GIS technology to produce maps regarding the land ownership by the applicant.

KKP : Komputerisasi Kantor Pertanahan, a name for the Indonesian Cadaster Database developed by the Ministry of ATR/BPN.

Kementerian ATR/BPN : Ministry of Agrarian Affairs and Spatial Planning / National Land Agency, an official institution under the president that is responsible for managing the land administration in Indonesia.

K4 Land Parcels : The other terms of unplotted land parcels in Indonesia.

Kota Lengkap : A city that has declared its completion of the land registration process.

Measurement Letter/Surat Ukur : A document that contains the physical data of a land parcel in the form of maps and descriptions.

NIB (Nomor Induk Bidang) : The parcel's Identification Number used in the Indonesian Cadastral Database contains information about the province, municipality, district, and village of the land parcels.

PTSL : Pendaftaran Tanah Sistematik Lengkap is a program from the government to systematically map the land parcels toward the complete land registration of Indonesia.

1. INTRODUCTION

1.1. Background

Land is an important aspect of human life. It is not only the place to build a house for a living but also has a wider impact on the sustenance of life and livelihoods. To administrate land properly, certain rules are used as guidance known as the land management paradigm. The land management paradigm consists of three interrelated aspects: land policy, land information structures, and land administration infrastructures (Enemark, 2005). The cadastral map is a part of land information structures that provides essential information about the location and area of a land parcel. These parcels are connected to the legal right of landowners which differentiates them from the other geospatial data. This means the land parcels cannot overlap, as an overlapping area can also be interpreted as overlapping ownership rights. The situation potentially leads to land disputes, which may create economic loss and national instability that contributes to increasing poverty (Hutabarat, 2011).

Various survey and mapping methods have been used over time to obtain geospatial information on a land parcel, resulting in a different map quality that may cause boundary overlapping. To ensure that land parcels don't overlap each other, geoinformation science plays an important role in it. The cadastral map is created using the survey and mapping techniques and stored in a Geo Information System (GIS) database. To enable the process that involves the GIS database, the old land parcel with its specific coordinates system which is mainly local needs to be plotted in the current cadaster database. The process of digitizing old land parcel data and remapping it with certain accuracy standards in a digital cadastral map is also called cadastral quality improvement (Grant et al., 2018).

The land registration process always goes along with the process of cadastral quality improvement which is done by many countries around the world. Spain and Turkey use information from old cadastral maps to improve the spatial quality of the current cadastral database (Femenia-Ribera et al., 2022; Yildiz & Erden, 2020). South Korea spent money and time on this process by doing a remeasurement of all land parcels in the country and remapping it with present accuracy standards to avoid overlap and ensure tenure security (Joo & Kim, 2014).

Remapping and plotting these land parcels is related to a sequence of activities that takes time, money, and human resources. Meanwhile, the longer time to complete the process will create a bigger potential for land disputes that may be changed into land conflicts (Agegnehu et al., 2021). In some countries like Indonesia, this process took longer time to complete due to the large number of parcels and the complexity of the problem. The issues that emerge from the process are related to the large number of land parcels that don't come with enough spatial information which creates difficulties when replotting is carried out to the cadaster database.

To cut the time spent on cadastral quality improvements, the use of technology also has been researched and proven to have a crucial impact in creating an efficient process of land administration (FAO et al., 2022). One of the technology implementations in land administration is using machine learning to accelerate the manual process. Most of the current implementations are mainly focused on automatically digitizing land boundaries from a satellite image (Crommelinck et al., 2019; Jong et al., 2022; Wudye Tareke, 2022; Zhang et al., 2021). Meanwhile, the ability of machine learning to replicate human abilities to make decisions in solving problems (Mirshekarian & Sormaz, 2018) opens up an opportunity to implement these methods in land administration, especially in the cadastral quality improvement process which is still unexplored.

The goal of this research is to implement machine learning in land administration, especially in the cadastral quality improvement process. By using the ability to learn from the training data, this research will automate the current manual method to create the right decision for a certain problem which is also called a heuristic process. The automation of heuristic processes will simplify the calculations needed by the computer to find the fastest way to properly solve a real problem (Steed & Williams, 2020). The inspiration comes from puzzle games which use the shape of puzzle pieces to find the best location for it in a puzzle board. This research will learn several prompts from the manually mapped land parcels such as parcel shapes and textual attributes. The expected output will be a model that can automatically spot the possible location in the Cadaster database to plot the land parcels that have limited spatial information.

1.2. Problem Statement

Indonesia is one of the countries that is also working on cadastral quality improvements. The process of cadastral quality improvement in Indonesia started in 2018 in the city of Surakarta and continues to other major cities such as Jakarta, Batam, Pontianak, Bali, and Surabaya. The large land area and the diversity of its nature and social structure bring challenges in the replotting process of old cadaster maps in Indonesia. There are approximately 126 million land parcels in Indonesia, and 100 million of them are already mapped in the GIS database. From that number of parcels, there are approximately 21 million parcels that are not yet plotted in the correct location (Ministry of ATR/BPN, 2023). This may have happened because of the sporadic adjudication, incomplete documents, and old survey and mapping methods that bring the challenge to the present progress of registration and need to be solved to create a complete land registration (Aditya, Santosa, et al., 2021). To plot that huge number of parcels, using a manual searching method is not fast enough. Meanwhile, delaying the time of plotting can create a bigger potential for land conflict. This land conflict can escalate into a violent conflict if not well managed and anticipated (Alston et al., 2000).

The old land parcels are very important to be plotted correctly in the current cadastral database for some reason. One of them which is related to this research is to create a complete cadaster database. The main idea to create a complete cadaster appeared in 1996 when the meeting between BPN, FIG, and the UN was held (FIG, 1996). The declaration was followed by some programs from the government such as PRONA and PTSL to achieve the target of complete cadaster in Indonesia. Those programs give the lesson learned to administrators that the complete cadaster cannot be achieved if the old paper-form certificate is not plotted yet to the current cadaster database. Those types of parcels are giving uncertainty to the current progress of registration and also endanger the security of tenure for the landowners (Martono et al., 2022). Moreover, the owners of old certificates mostly do not live in the parcel's location so they are not aware of the land parcel's correct position. This situation creates a problem when the new owners of the land come to claim the parcels and make the certificate. This problem can escalate into the court processes which takes more time to finish and delay the whole process of land registration.

Several challenges are already mentioned at the beginning of this paragraph but in general, they can be categorized into two: technical-related and legal-related. The legal-related challenges come as the result of the previous regulation which allows the issuing of certificates without the maps on them. Republic Indonesia Government (1961) stated that if the Surat Ukur (measurement letter) cannot be produced for some reason, a replacement certificate can be issued that has the same function as the certificate. It means that those types of certificates have the same legal force as the normal ones and also need to be plotted in the current cadastral database. The big question mark that appears then is how to plot a certificate that has no maps or any spatial data on it.

The other challenge that is related to the technical aspect is the use of a local coordinate system in the old land parcel mapping. The condition happened because of the limited availability of the national control point (Titik Dasar Teknik) at that time and there were limitations in surveying technology (Handono et al., 2020). To map the land parcel, natural features such as roads or rivers are very common to map the land parcels. The problem appears when those features are changed in the present time so the location cannot be backtracked and the land parcels become "flying" somewhere else. The situation creates a common term within ATR/BPN for this type of parcel as a "flying parcel".

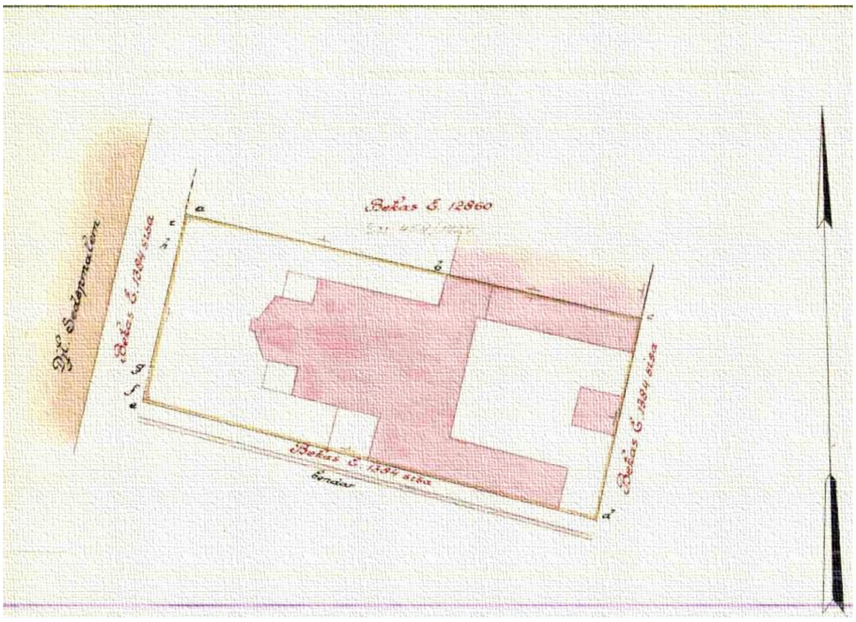


Figure 1. Parcel map (Surat Ukur) tied to natural features

Based on the challenge of plotting the land parcel without adequate spatial information, this research will adopt a machine learning algorithm to find the most possible location to plot the parcels. By using the measurable components from the manual process as heuristics, the geospatial matching between the unplotted land parcels and available empty locations will be calculated. The model then will be generalized and used to find the most possible location in a cadastral database to semi-automatically plot the land parcel.

The outcome of this research can help the surveyor accelerate the manual process of finding the most possible location from a set of cadastral databases. With automation, it is possible to plot more land parcels compared with the manual method. The more land parcels to be plotted, the faster of whole Cadastral Quality Improvement process in general and prevent the tenure insecurity caused by unreliable cadastral data.

1.3. Objectives and Research Questions

1.3.1. Main Objective

The main objective of this research is to automate the manual heuristic process of finding the best-fitting location for the unplotted land parcel based on geospatial data matching.

1.3.2. Sub-Objectives and Questions

SO 1: To identify the causes of the land parcel becoming unplotted and identify the heuristic process to plot that type of land parcel.

1. What are the causes of the land parcel becoming unplotted?
2. What heuristic process has been used to plot the land parcel manually to a current cadastral database?

SO 2: To adopt a machine learning model in identifying the matching between two geospatial data to automatically find the best-fitting location for the unplotted land parcel.

3. What are the matching components from the heuristic process that can be used to identify the matching between an unplotted land parcel and the available locations in the parcel database?
4. How to optimize the components to correctly identify matching between an unplotted land parcel and the available locations in the parcel database?

SO 3: Evaluate the model's performance in identifying the matching between two geospatial data to automatically find the best-fitting location for the unplotted land parcel.

5. How many correct matches does the model get for finding the best-fitting location for the unplotted land parcel?
6. How does each component influence the model's performance of finding the best-fitting location for the unplotted land parcel?
7. What are the factors that contribute to the model performance of finding the best-fitting location for the unplotted land parcel?

1.4. Conceptual Framework

This research aimed to adapt a machine learning model to automate the process of manual search for the locations to plot the unplotted land parcels. Predictions were made based on the geospatial data matching between the unplotted land parcel and the available empty location in a set of cadaster data. Results from this model will filter the candidate for the land parcel to be checked in the next step and enable the accelerator to save time for plotting the land parcel without position information.

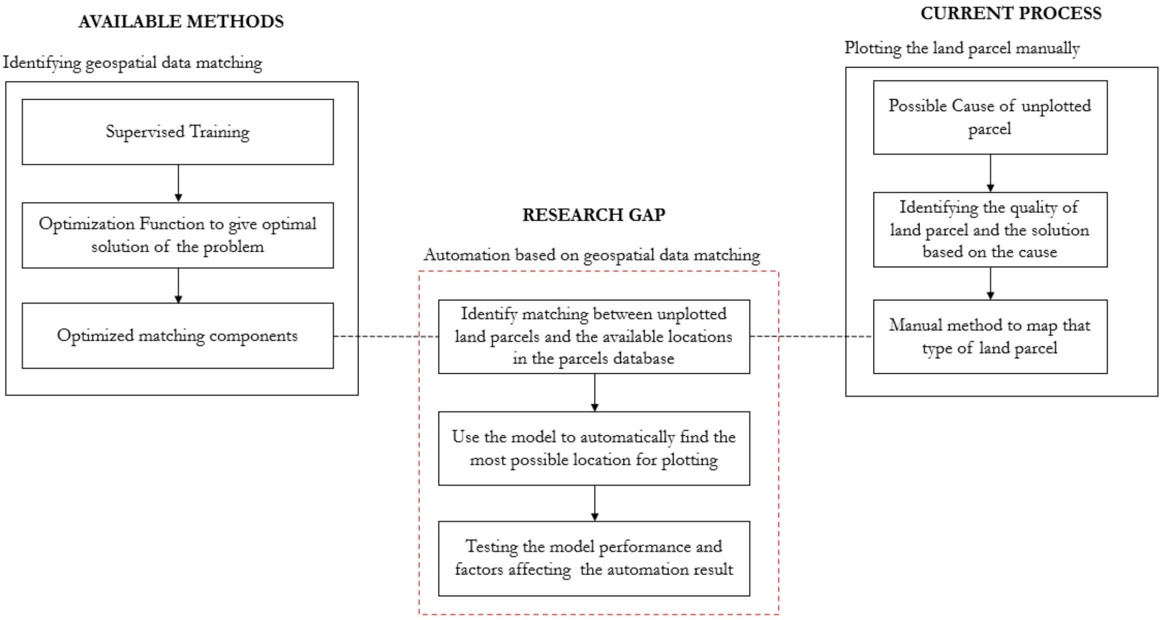

AVAILABLE METHODS
Identifying geospatial data matching
Supervised Training
Optimization Function to give optimal solution of the problem
Optimized matching components
RESEARCH GAP
Automation based on geospatial data matching
Identify matching between unplotted land parcels and the available locations in the parcels database
Use the model to automatically find the most possible location for plotting
Testing the model performance and factors affecting the automation result
CURRENT PROCESS
Plotting the land parcel manually
Possible Cause of unplotted parcel
Identifying the quality of land parcel and the solution based on the cause
Manual method to map that type of land parcel

Figure 2. The Conceptual Framework

1.5. Structures of The Thesis

This thesis consists of six chapters, each providing a structured explanation of the research questions and their corresponding answers:

Chapter 1: Introduction

This chapter introduces the background and the problem that inspired the research. It also elaborates on the objectives and questions which will be solved using the described conceptual framework.

Chapter 2: Literature Review

This chapter presents the result of previous work related to the cadaster and the cadaster quality improvement process both in best practices and in the Indonesian case. The use of machine learning in land administration and specifically in the process of cadastral quality improvement. The related works on the use of optimization algorithms in machine learning to find the optimum solution to certain problems.

Chapter 3: Research Design and Methods

This chapter elaborates on the research design used in this research, together with the complete method used to answer the research question. The methods were divided into qualitative methods: open-ended interviews and quantitative data analysis using machine learning.

Chapter 4: Result

This chapter presents the result of the interview process done in the fieldwork in the graphs and tables. The results of the geometric matching process to find the location were described in this chapter as well.

Chapter 5: Analysis and Discussion

This chapter analyses the results from the previous chapter on land administration and the cadaster data condition in Indonesia. It also presents the result of enhancing model performance using textual matching and elaborates on the factors related to model performance, which are important to future implementation.

Chapter 6: Conclusion and Recommendation

This chapter concludes the results and analysis from previous sections, addressing their relevance to the research questions. This chapter also provides recommendations for the institution and suggestions for future research.

2. LITERATURE REVIEW

2.1. Cadaster as Part of Land Administration

Land Administration is not a new field of knowledge, it has already been used in previous times as the process of managing the land to get benefit from it for instance for collecting tax or trading. The formal definition itself settled in 1996 as the *"process of determining, recording, and disseminating information related to ownership, value, and use of land when implementing land management policies"* (UNECE, 1996).

As part of the land administration, cadaster is defined by many literatures in a different way. FIG (1995) defines a working definition *"A Cadaster is normally a parcel-based system which consists of boundaries that are marked and uniquely identified"*. Cadaster is also seen as a process of providing spatial and attribute information for the process of registering, valuing, and managing the land (UNECE, 2005). The implementations of cadaster in many countries also vary as explained by Duncan & Rahman (2013) and Rajabifard et al (2007). To conclude all of the definitions, the cadastre is seen as a system rather than a single subject that connects the identification, registration, valuation, and taxation which is also related to a concept called multi-purpose cadaster (D. Grant et al., 2020).

The function of cadaster as a system is explained by Enemark (2005) as a junction between land tenure, land taxation, and land development. The system covers the recognition of the land parcels using mapping technology and their legal aspect of land ownership which can be held formally and informally based on the country's condition (Adam et al., 2019). It also provides valuable information in the process of land valuation and taxation to correctly determine a proper value for a land depending on its spatial conditions. The cadastral system is also required to create a sustainable plan for land use development and resource management using spatial analysis capability. Those three pillars will be working collaboratively to achieve the ultimate goals of land management such as social stability, sustainable economic growth, and security of tenure.

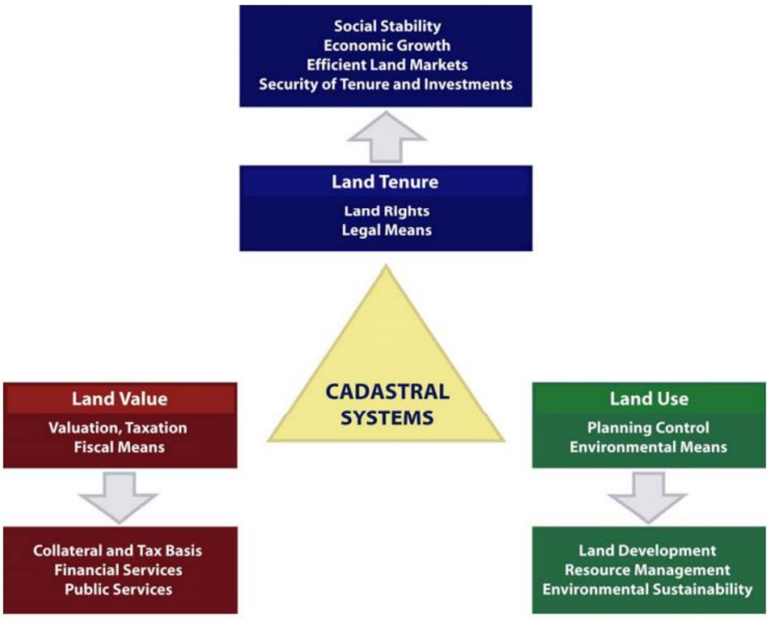


Figure 3. The Concept of Cadastral Systems (Enemark, 2005)

2.2. The Evolution of Cadaster Data Acquisition

The history of the acquisition of cadaster data goes along with the history of survey and mapping itself. It starts with the simple use of a magnetic compass by the Egyptians and Chinese around 3.000 BC (Holsen & Lsen, 1984). With limited accuracy, those technologies have been used in several applications such as road construction and land division. The more advanced technique was invented by Willebrord Snell, a Dutch cartographer who used the triangulation concept to measure the angle using two reference lines (Murdin, 2009). Instead of the old invention date, the concept of triangulation is still used in the present time in many advanced applications.

During the time, the use of more accurate devices such as Theodolite was common in the 1870s and made it a milestone to use precision measuring instruments in surveying activity (Avram et al., 2016). But still, the result from the theodolite doesn't give the location information about the land parcel and prompts a human error because the angle calculation is still manually done by the surveyor. The need to accurately define the coordinates of a location in a universal coordinate system brings the use of Geodetic GNSS in cadaster data acquisition. It also gives the accurate position of the points as the smallest representation of a land parcel in the millimeter fraction (Khomsin et al., 2019).

The mapping results from those measuring devices as mentioned above also differ. The early period of theodolite only gave the paper maps that needed to be digitized and rectified into the current digital system. Meanwhile, the current sophisticated method such as Geodetic GNSS gives digital output with the high accuracy of point and mapped in the universal projection system. The situation resulting a different level of cadaster data quality based on the accuracy of measurement and becomes a challenge in the present land registration process (Aditya, Santosa, et al., 2021).

2.3. Cadastral Quality Improvement

The different methods of measurement in the past have resulted in the different quality of cadastral data. There are several methods used by previous researchers to improve the quality of old cadaster maps. Spain digitized their old cadastral map and announced it in the online system to clear up the dispute on the cadastral boundary (Femenia-Ribera et al., 2022). Turkey also uses their old cadastral data to categorize it into several classes, those classes will be categorized and solved differently based on the source of the problem (Yildiz & Erden, 2020).

The improvement process is not only limited to the digitization of old parcel maps but also touches the geometric aspects of the land parcel. It is important to have an accurate land parcel map with minimum error and not overlapping with each other. Malaysia used a certain mathematical model to reduce systematic and gross errors in their National Digital Cadastral Database (Hashim et al., 2013). The research was also done in Israel by using simple mathematic principles to do a block adjustment of a separated land parcel block (Klebanov & Doytsher, 2009).

The existence of documentation in the cadastral map and the physical boundaries of land parcels in the field is very important to support the process of improving cadastral quality. Both of these components need to be aligned to develop and maintain spatial boundaries (D. Grant et al., 2020). In the research that was done in Australia and New Zealand, Grant et al. (2018) defined seven levels of cadastral quality improvement based on its form of documentation and the level of uncertainty. Level 0 shows the cadastral map in a graphical paper which contains high positional uncertainty. The highest level of this categorization shows the cadaster map with a legal coordinate, giving more certainty to the parcel's position.

2.4. Cadastral Quality Improvement in Indonesia

Cadastral quality improvement has also been done in Indonesia as a result of unreliable cadaster data and the high amount of overlapping parcels that brought more disputes and cases in the court. The research from Sabekti (2010) gives the first step of the Indonesian cadastral maps quality improvement by making the strategy to convert old-paper data to digital data that is acceptable with the current database. The activity of "forensic cadaster" by using a mobile application to gather information about parcels' position on the field was also done in the city of Denpasar. It resulted in good progress of 5970 certificates being validated in only three months (Aditya, Sucaya, et al., 2021).

From experience in the cadaster quality improvement process, Indonesia divides its data into six classes based on the availability of land records and maps of each parcel. This process then creates a different method of quality improvement based on each parcel class. The highest quality of land parcel is Quality 1 (KW 1) which is already plotted in the cadastral map, has complete documentation, and shows a consistent spatial quality between paper, document, and electronic records. The land parcel that became the focus of this research is KW 4, KW 5, and KW 6 land parcels which are not yet plotted in the Cadastral Maps. For this type of land parcel, the suitable treatment is map redrawing or spatial adjustment. (Aditya, Santosa, et al., 2021).

2.5. The Use of Machine Learning in Land Administration

Machine Learning is a general term that is used to describe the process of imitating the human's ability to understand semantic meanings or detect patterns from a dataset (Nichols et al., 2019). Its ability to automate the manual process by humans using certain algorithms leads to a broad application of machine learning in many fields.

To cover the activity stated in the land administration definition, there are several applications of machine learning to support the process related to land ownership, land value, and land use. The capability of automation offered by machine learning is used in a new tool called Smart Sketch Map which converts handwritten sketches from the community to information which helps accelerate the ownership recognition process. To enable the positioning, it uses a relative position on the sketch to be compared with the real features in a geocoordinates map (Chipofya et al., 2017).

For the other main process in land administration, machine learning has proven to accelerate the process of making a valuation for property. Mayer et al. (2022) made a prediction of land and structure in Miami and Switzerland using STAR Models and Deep Learning. The research was also done in Germany and Los Angeles to predict the value of real estate using textual information about the property. The research uses several predictive models based on random forest, gradient boosting, and regression calculation models (Baur et al., 2023).

The last aspect of land administration which is also the most researched is land use and land management. Most of the research is trying to make an automatic prediction of land use and land cover from satellite or aerial photogrammetry images (Alem & Kumar, 2020; Chaturvedi & de Vries, 2021; Yuh et al., 2023). The model mainly used a machine learning classification based on the nearest neighbor, the support vector machine, and the random forest algorithm.

2.6. Machine Learning for Cadastral Quality Improvement

Machine learning is also used in the process of cadastral quality improvement in helping to automate the current manual process to make it faster. The most application is in the process of media conversion of the certificate from analog to digital. Automatic boundary delineation is the most common use to accelerate the process of manual digitization.

Some research has used the technology to accelerate the process of delineating visual boundaries that can be detected from satellite images using deep learning models such as CNN and FCN (Gafurov, 2023; Wudye Tareke, 2022; Zhang et al., 2021). Other research has also been done to improve the quality of automatic delineation using ResUNet (Jong et al., 2022) but still, the application is only applicable in the visible cadastral boundaries such as ricefield or irrigation.

The process of automatically digitizing cadastral boundaries has also been done in The Netherlands, different from previous research this process utilizes the old cadastral maps and integrates the result of a digitized map into the new cadastral database (Franken & Florijn, 2021). The research was also done by Wouters et al. (2010) to automatically read the information written in the certificate using text and optical character recognition.

The recent research on improving the quality of the current cadastral database by combining the results from other maps was done in the city of Tehran, Iran. This research uses parcel matching based on the center points from polygons to be matched with polygons from the municipality database. By using several machine learning optimization algorithms such as Random Forest and Genetic Algorithms. The result from the matching will prompt a change and will be the object of further parcel map enrichment (Hajiheidari et al., 2024).

2.7. Identify Geospatial Data Matching Using Machine Learning

The use of machine learning for detecting matching between two geospatial data has already been researched with many different methods and applied in many fields. The similarity between two features in a polygon format can be determined by comparing the similarity of the polygon's skeleton (Mortara & Spagnuolo, 2001). The feature matching measurement using the turning function adapted from Arkin et al. (1991) was applied to assess the quality of the building polygon in the Open Street Map (Fan et al., 2014). The other research summarizes all variables that are possible to identify geospatial data matching including geometric features, topological, attribute, context, and semantics which also explains the method to analyze the performance of each method used (Xavier et al., 2016).

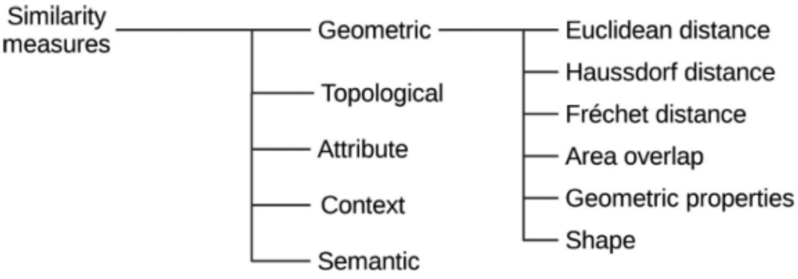


Figure 4. The similarity variables for geospatial data (Xavier et al., 2016)

The use of machine learning has also been researched to detect similarities between two databases that contain a polygon of building. The method used in this research is to optimize each geometrical variable obtained from previous research and use the result to detect matching between those two databases. (Ruiz-Lendínez et al., 2017).

2.8. Machine Learning for Optimization Process

The relationship between a mathematical model and problem-solving using several consecutive processes Noble (1982) has inspired the use of machine learning to solve problems in various fields of knowledge. Machine learning can imitate all the processes humans do in solving a problem using a model. One of the processes that could be done using machine learning is the optimization process. Optimization is a term for the process of searching for the optimal solution to a problem by maximizing or minimizing an objective function (Sun et al., 2019).

Several algorithms were used to search for an optimal solution to a problem such as Model-Free Algorithms, Gradient-Based Optimization, Bayer Optimization, and Metaheuristic Algorithms which can be selected based on the complexity of the data and the problem itself. The research from Yang & Shami (2022) gives a comparison in terms of calculation time, accuracy, strength, and limitations of each available algorithm. It is important to select proper algorithms that can handle the randomness of the objective function and the complexity of searching space to minimize the time needed for calculation and optimize the solution accuracy (Claesen & De Moor, 2015).

2.9. Genetic Algorithm for Optimization

Genetic algorithms are one example of metaheuristic optimization which are inspired by a natural process of selection and mutation. The concept behind this algorithm is to find the top combination from successive generations that gives the optimal solution to the problem (Holland, 1992). The robustness of this algorithm and the ability to search within a complex search space within a reasonable processing time (Herrera et al., 1998) has made a wide application of this algorithm for optimization. One of the applications that commonly uses the principle of genetic algorithm is calculating similarity.

There are several research on similarity calculation using this algorithm, it is not limited to numerical problems but also solves textual-related problems. The research to automate the process of summarizing a text in the Hindi language is done using the genetic algorithm by transforming the text into mathematical variables, those variables will be used in the optimization process to find the maximum value (Jain et al., 2022).

The other applications related to the geoinformation problems were also researched for example the use of genetic algorithms in geotechnics problems (Simpson & Priest, 1993), spatial analysis based on geographic information systems (X. Li et al., 2005), and optimizing land use (Ding et al., 2021). Not only for maximizing the variable to find the fittest chromosome, the genetic algorithm is also applicable to minimize the objective function in the variable weighting process. Despite the infrequent use of genetic algorithms in this field (Hamarat & Kilic, 2010), the use of the genetic algorithm to weigh the attribute in the specific field is proven to be reliable with several modifications to the original code (D. Li et al., 2016; Varpa et al., 2014).

2.10. Calculating the Geometric Variable of a Polygon

The geometric aspect of geospatial data is one of the prompts that can be used to identify matching between two spatial objects. Several variables from the geometric aspect can be used such as the Hausdroff distance, area overlap, and geometrics and shape (Xavier et al., 2016). The research used five geometric variables based on similar research and the information gathered from the interview. The first variable was the polygon area, denoted with A. The polygon area is calculated using Gauss's area formula to accommodate the polygon's shape irregularity, the formula used to calculate the polygon area is described in equation (1).

$$A = \frac{1}{2}\Sigma_{i=0}^{n-1}(x_i . y_{i+1} - x_{i+1} . y_i) \qquad (1)$$

where n is the number of vertices of the polygon;
(x_i, y_i) are the coordinates of the vertices

The next variable was the number of vertices represented as n, there is no specific formula to calculate this variable, the number of vertices was determined by counting the coordinates of the outer boundary (exterior) of the polygon.

The third variable was the polygon perimeter represented as P, which is calculated by summing the lengths of all line segments forming the polygon's exterior. The formula used to calculate the polygon perimeter is described in equation (2).

$$P = \Sigma_{i=0}^{n-1}\sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \qquad (2)$$

where n is the number of vertices of the polygon;
(x_i, y_i) are the coordinates of the vertices

The fourth variable used in the calculation was the Minimum Bounding Rectangle (MBR) which is defined as the smallest rectangle that entirely encloses the polygon (Caldwell, 2005). To better understand the concept of MBR figure 5 shows the illustration of MBR.

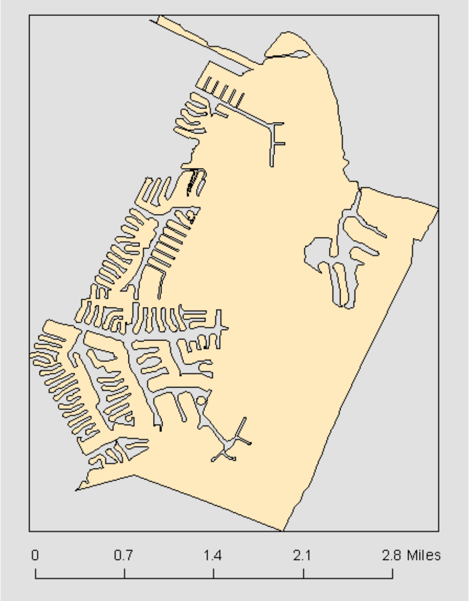


Figure 5. The Illustration of The MBR of a Polygon (Caldwell, 2005)

The area of MBR (m^2) needs to be determined to include this variable in the matching calculation by using equation (3).

$$MBR_{Area} = \text{Width x Height} \qquad (3)$$

The last variable that was used in this research was the Arkin Graph Angel (AGA). AGA is defined as an area under the turning function of polygon 1 (θ_1). The turning function is a function that represents the angle that measures the anticlockwise tangent at every point along the boundary of the polygon (E. M. Arkin et al., 1991).

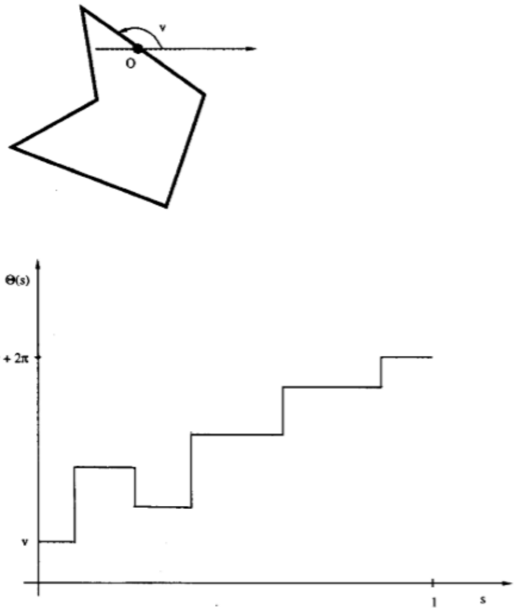


Figure 6. Defining the turning function of Polygon 1 (θ_1) (E. M. Arkin et al., 1991)

The AGA was calculated using equation (4) by performing an integral calculation of the turning function for each vertice in the polygon. The vertices plotted in the x-axis of the turning function were normalized to make a value from 0 to 1.

$$AGA = \int_0^1 \theta_1 \; (s)ds \qquad (4)$$

where θ_1 is the turning function of polygon 1;
s is the arc-length parameter of polygon 1

2.11. Chapter Summary

This chapter reviewed the existing research on several topics that develop the research framework, including the basic definition of cadaster in land administration, the cadaster quality improvement process, the application of machine learning for land administration, and specific methods in machine learning that will be applicable for the research. The next chapter will explain methods that are used to answer all of the objectives and research questions.

3. RESEARCH DESIGN AND METHODS

3.1. Study Area

The study area for this research is Indonesia with the scope of analysis within the city of Surabaya. Surabaya is the second largest city with an area of 326,81 km2 and has a population density of 8.633 people per square kilometer (Badan Pusat Statistik Kota Surabaya, 2023).

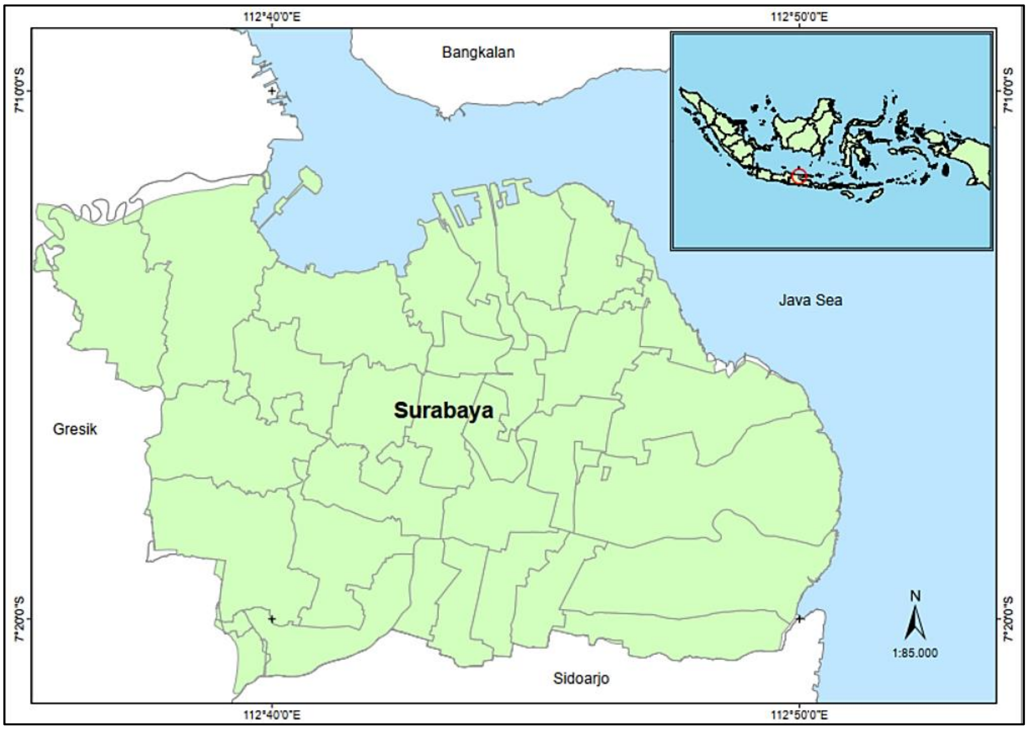


Figure 7. Map of Surabaya

The city of Surabaya has been chosen as the study area because in 2022 there was research from the University of Gadjah Mada in cooperation with the local land office to increase the quality of registered land parcels including the process of landed unmapped parcels manually. Based on the report, there were 532.007 land parcels in Surabaya and 7% of the total is still unmapped (Tim Peneliti Teknik Geodesi UGM, 2022). Those land parcels were then successfully remapped into the correct location on the ground using several methods like document tracing, name searching using tax maps, and participative mapping by the owners.

The huge amount of data that has been successfully landed, will provide this research valuable input of data to be trained in the model. The model hopefully captures the condition of the parcel and creates weight for each variable that can be used generally in other locations in Indonesia. This huge data available also enables this research to capture the variation of the condition of land parcels in an urban area which is not only in a heterogonous form (i.e. industrial and agricultural area) but also includes some homogenous form parcels (housing complex).

3.2. Research Design

The research will be divided into three main stages: semi-structured interview, optimizing the geospatial data matching components, and performance evaluation. The semi-structured interview will be done with the practitioner that experienced in plotting the land parcel manually. The second stage is related to using an optimization function in machine learning to get a better combination of components from previous steps in detecting a match between two geospatial data. The performance of the optimization model in detecting matches and the influence of each component will be evaluated. Figure 8 shows the complete workflow for the research.

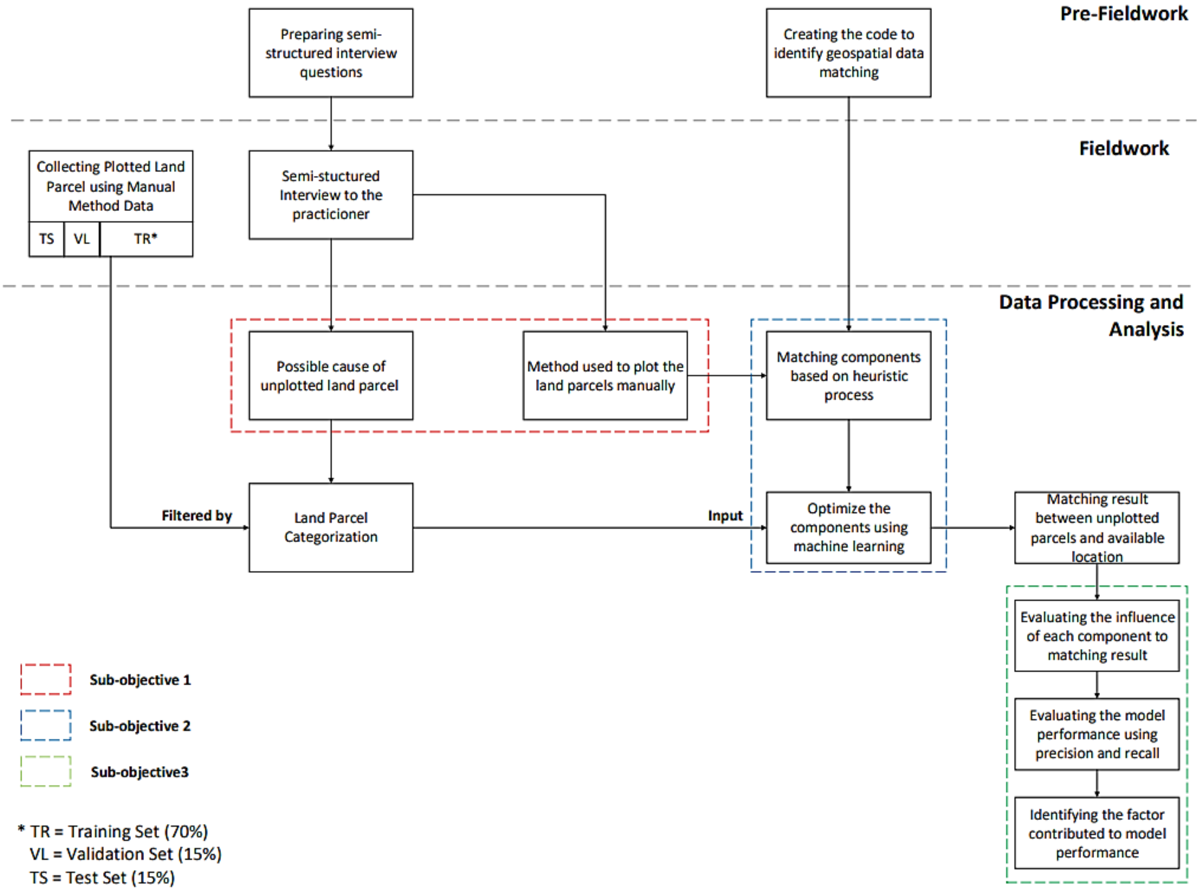

Pre-Fieldwork
Preparing semi-structured interview questions
Creating the code to identify geospatial data matching
Fieldwork
Collecting Plotted Land Parcel using Manual Method Data
TS
VL
TR*
Semi-stuctured Interview to the practicioner
Data Processing and Analysis
Possible cause of unplotted land parcel
Method used to plot the land parcels manually
Matching components based on heuristic process
Filtered by
Land Parcel Categorization
Input
Optimize the components using machine learning
Matching result between unplotted parcels and available location
Evaluating the influence of each component to matching result
Evaluating the model performance using precision and recall
Identifying the factor contributed to model performance
Sub-objective 1
Sub-objective 2
Sub-objective3
* TR = Training Set (70%)
VL = Validation Set (15%)
TS = Test Set (15%)

Figure 8. The Research Design

3.3. Research Methods

3.3.1. Pre-Fieldwork

The research inspiration came from the real problem that is still in demand for a solution. However, to build the proper research question and objective several literature reviews were conducted during the pre-fieldwork process. During this phase, the initial code for the geospatial data matching was built in the Phyton-based programming environment. Some previous research related to this work was used as guidance to develop an initial model to detect the similarity between two polygons.

In preparation for the semi-structured interview, several iterations were made to develop a proper question list. The questions were categorized into two main themes to answer the research questions. Several prompts were also prepared in this phase to help create follow-up questions in the fieldwork phase.

3.3.2. Fieldwork

During the fieldwork phase, there were two main activities: the semi-structured interview and secondary data collection.

3.3.2.1. Semi-structured Interview

The semi-structured interview is a type of qualitative data analysis to understand one unknown concept using a set of open-ended questions with additional questions that arise during the interview process (DiCicco-Bloom & Crabtree, 2006). This method was constructed by a guided set of questions but also enables the interviewer to go deeper if more explanation is needed.

The interview was held in 4 cities in Indonesia: Jakarta, Yogyakarta, Surabaya, and Serang, involving 11 respondents who have been involved in the process of cadastral quality improvement. Figure 9 shows the process of in-depth interviews that was done during the fieldwork phase.

Figure 9. Semi-structured Interview

The choice of the city and respondent profile is to cover the information not only from the people who directly did the cadastral quality improvement process but also from the person who made the regulation and the research team who were researching the process using the manual method. The profile of the respondent is explained in Table 1.

Table 1: Profile of The Respondent

Institution	Status	Location	Number of Respondent
Directorate of Cadastral Survey and Mapping	Regulator	Jakarta	5
	Project Coordinator	Jakarta	1
Universitas Gadjah Mada	Researcher	Yogyakarta	1
	Surveyor	Yogyakarta	1
Surabaya II Land Office	Project Coordinator	Surabaya	1
	Surveyor	Surabaya	1
Serang Land Office	Project Coordinator	Serang	1

3.3.2.2. Secondary Data Collection

In the fieldwork phase, secondary data was collected at Surabaya City from the cadastral data quality improvement process that was done in 2022. Due to its large area, two Land Office are serving Surabaya City: Surabaya I and Surabaya II. For this research, all of the land parcels were used only from the Surabaya II Land Office which covers 15 districts from the total of 31 districts in Surabaya (Indonesian National Land Agency, 2010). The data consists of land parcels in shapefile format, textual documents related to the process, and photogrammetry images. The list and complete description of the data collected are explained in Table 2.

Table 2: Description of the Data Collected

Data	Format	Amount of Data	Period of Acquisition
Manually Plotted Land Parcels	Shapefile (.shp) Shapefile Database (.dbf)	7915 parcels	2022
Surabaya II Cadastral Data	Shapefile (.shp) Shapefile Database (.dbf)	265504 parcels	2024
Surabaya Photogramettry Image	Enhance Compression Wavelet (.ecw)	79 images	2017
List of plotted land parcel	Microsoft Excel Spreadsheet (.xlsx)	103 files	2022

3.3.3. Data Processing and Analysis

The data processing and analysis phase was done after the fieldwork so it also can be referred to as the post-fieldwork phase. This phase is related to four sequential activities: qualitative data analysis, dataset preparation, optimization algorithm, and evaluation.

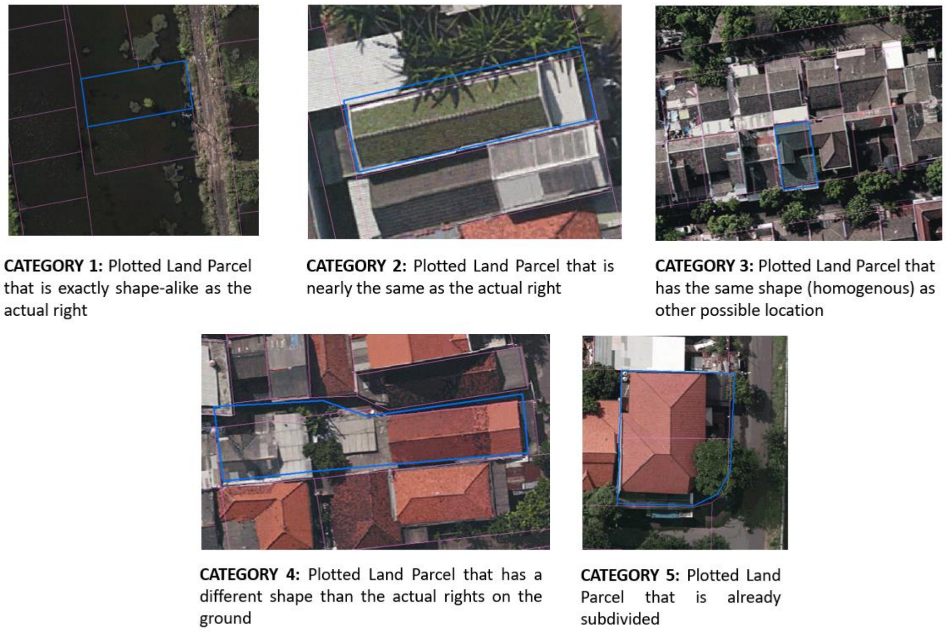
3.3.3.1. Qualitative Data Analysis

To extract important information from the interview a qualitative data analysis was done to the results. It was transcripted and analyzed iteratively until no new themes appeared, also identified as theme saturation (DiCicco-Bloom & Crabtree, 2006). The analysis was done using a combination of inductive and deductive methods. This method was collecting the information from the interview and developing it with the guidance of existing regulations and concepts (Brooks et al., 2019).

The transcript was uploaded into Atlas.ti and divided into parts based on its general theme. Those parts were divided into quotations and from that the labels were given to a similar theme. The process of labeling similar information from a quotation is also called a coding process (Smit, 2002). The quotation was quantified and presented as a graph to support the analysis. The table of code used in this research is available in the annex.

3.3.3.2. Dataset Preparation

The machine learning model used the manually plotted land parcels and the actual parcel from the database as the training data. Based on the geometric correlation of both datasets, there are 5 categories of land parcels used in this research. The proportion of this categorization was maintained in the training, validation, and testing set to ensure the representativeness and generalization of the model. Figure 10 shows the categorization with parcel examples to make it clearer.



CATEGORY 1: Plotted Land Parcel that is exactly shape-alike as the actual right

CATEGORY 2: Plotted Land Parcel that is nearly the same as the actual right

CATEGORY 3: Plotted Land Parcel that has the same shape (homogenous) as other possible location

CATEGORY 4: Plotted Land Parcel that has a different shape than the actual rights on the ground

CATEGORY 5: Plotted Land Parcel that is already subdivided

Figure 10. Categorization of Training Data

The dataset gathered from the fieldwork is a separate polygon group that needs to be processed. To prepare the original data into a training-ready dataset, several steps need to be done that is described in Figure 11. From the total number of land parcels available for training, this research chose 1199 parcels to optimize the calculation time while keeping the Parcel Category proportion the same as the previous.

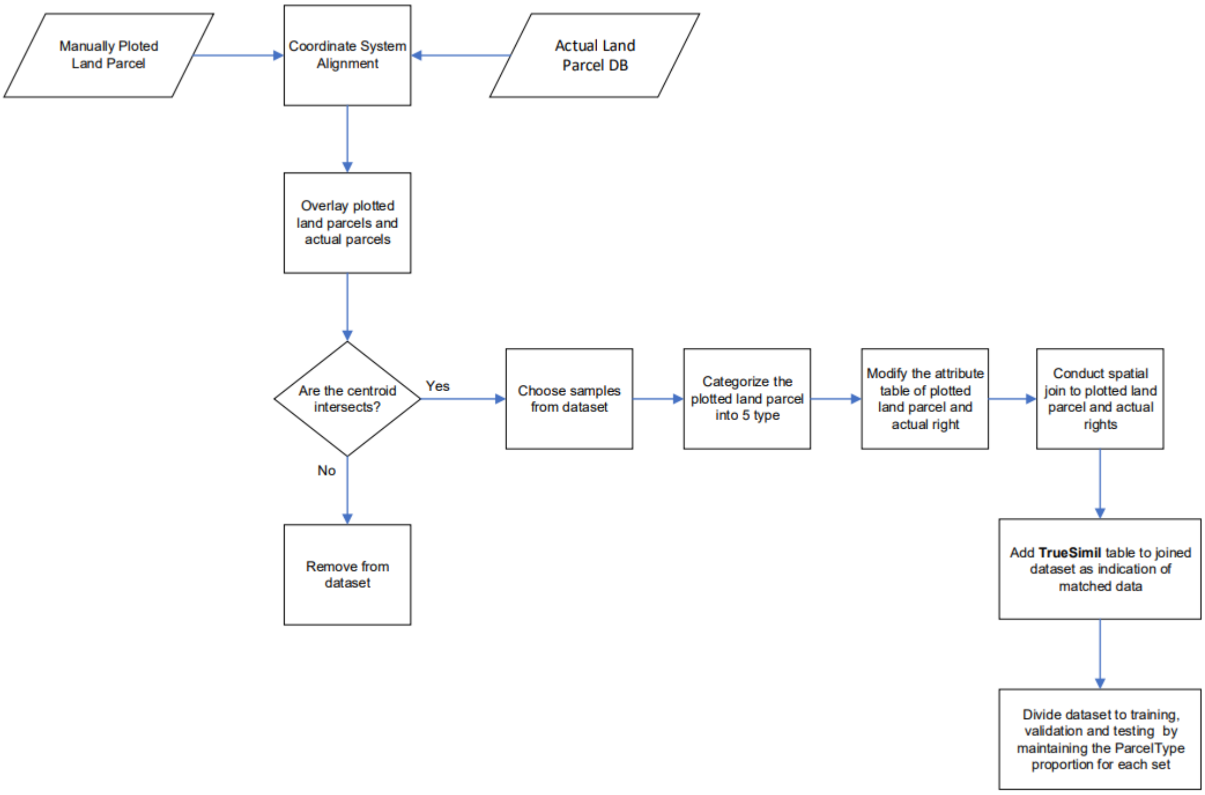

Manually Ploted Land Parcel
Coordinate System Alignment
Actual Land Parcel DB
Overlay plotted land parcels and actual parcels
Are the centroid intersects?
Yes
No
Remove from dataset
Choose samples from dataset
Categorize the plotted land parcel into 5 type
Modify the attribute table of plotted land parcel and actual right
Conduct spatial join to plotted land parcel and actual rights
Add **TrueSimil** table to joined dataset as indication of matched data
Divide dataset to training, validation and testing by maintaining the ParcelType proportion for each set

Figure 11. Dataset Preparation Workflow

3.3.3.3. Geospatial Data Matching Using Geometric Variable

The geospatial data matching between unplotted land parcels and the location from the parcel database were calculated by giving the weight to each variable from the geometric aspect. The weight was used as the parameter to be adjusted in the optimization algorithm. The geometric matching for each variable needs to be calculated before the weighting process occurs. Equation (5), (6), (7), (8), and (9) shows the formula to perform calculations.

$$A_{match} = 1 - \frac{|A_1 - A_2|}{A_1 + A_2} \quad (5)$$

$$n_{match} = 1 - \frac{|n_1 - n_2|}{n_1 + n_2} \quad (6)$$

$$P_{match} = 1 - \frac{|P_1 - P_2|}{P_1 + P_2} \quad (7)$$

$$MBR_{match} = \frac{MBR_1 \cap MBR_2}{MBR_1 \cup MBR_2} \quad (8)$$

$$AGA_{match} = \frac{\min(AGA_1, AGA_2)}{\max(AGA_1, AGA_2)} \quad (9)$$

The notations A_1, n_1, P_1, MBR_1, and AGA_1 refer to the components belonging to the polygon of the land parcels from the manual plotting process. In contrast, the notation A_2, n_2, P_2, MBR_2, and AGA_2 refers to the component of the polygon from the parcel database. The matching value for each variable spreads between 0 and 1, where 0 indicates no match and 1 indicates an identical match. The overall matching (OM) of two land parcels is calculated by accumulating the score from each variable multiplied by each weight. Equation (10) shows the formula for the calculation.

$$OM = \omega_1 . A_{match} + \omega_2 . n_{match} + \omega_3 . P_{match} + \omega_4 . MBR_{match} + \omega_5 . AGA_{match} \quad (10)$$

3.3.3.4. Define the Objective Function

The objective function is a mathematical function that defines the goal of an optimization algorithm, quantifying how well the model's prediction matches the true values. Specifically, it calculates the mean squared error (MSE) between the true similarities of parcels (initially set to 1, denoted as **TrueSimil**) and the predicted matches (denoted as **PredMatch**). The predicted matches are calculated iteratively over parcels in the training set, to find the set of weights that minimizes the MSE. These weights were then used in the exhaustive search to find the best-fitting location for plotting the parcels based on geometric variables. Equation (11) shows the formula to calculate the objective function for the parcel's pair ι based on the weight ω.

$$\text{Objective Function} = \frac{1}{n}\sum_{i=1}^{n}\left(TrueSimil_i - \text{PredMatch}_i(\omega)\right)^2 \quad (11)$$

where n is the number of parcels in the training/validation/testing set;

TrueSimil_i is the true similarity for parcel pair ι, initially set to 1;

$\text{PredMatch}_i(\omega)$ is the predicted similarity for parcel pair ι, calculated using the weights ω

3.3.3.5. Weight Optimization of Geometric Variable using RCGA

RCGA stands for Real Code Genetic Algorithm, another improvement of binary Genetic Algorithm optimization function. The benefit of using RCGA can solve the continuous problem in a set of vector values which the binary GA couldn't (Katoch et al., 2021).

The objective function that was previously defined was optimized using the RCGA algorithm. The goal is to minimize the MSE error between true similarities and the predicted similarities which will improve the accuracy of the model's prediction. This method uses a randomly generated population to be optimized iteratively to find the best solution using several crossovers and mutation operators. Crossover is an operator that combines information from two wellspring solutions to create a new ancestor solution. The mutation is an operator that changes the value of the variables in an individual solution, it is important to prevent the suboptimal solution caused by early convergence (Katoch et al., 2021). The optimized solution from the iterative calculation was chosen as the selected weight that will be used in the matching calculation. Figure 12 explains the process involved in the optimization function modified from the previous research.

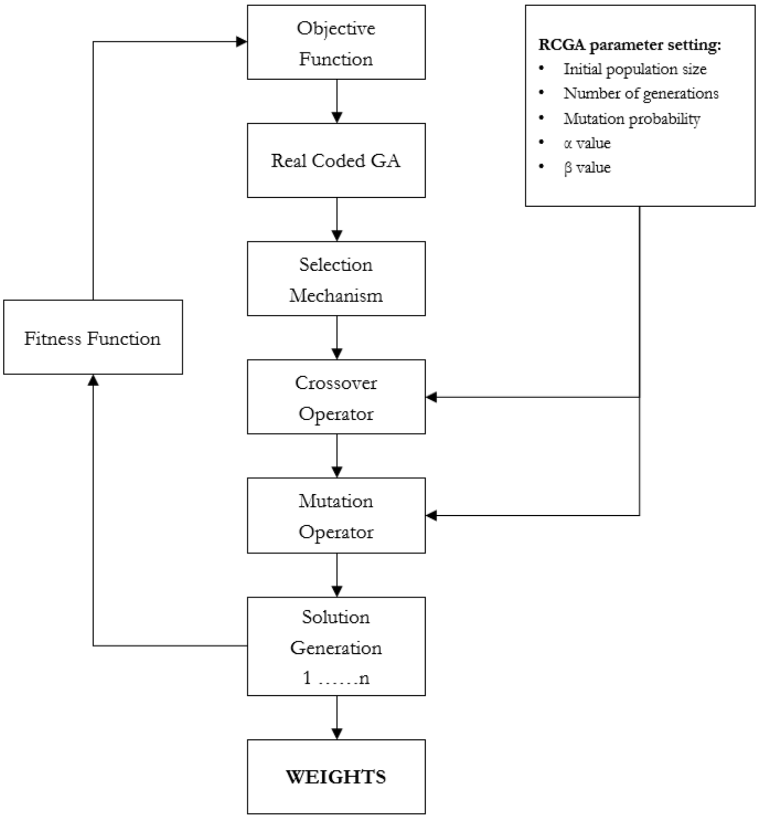


Figure 12. RCGA Optimization Process (modified from Ruiz-Lendínez et al. (2017))

3.3.3.6. RCGA Parameter Adjustment Strategy

As mentioned in the previous part, the five parameters of RCGA calculations need to be adjusted to achieve an optimal solution. The first parameter adjusted is the initial population size which represents how much data are taken randomly to be involved in the calculation (Rodriguez-Maya et al., 2016). The smaller size will give a faster calculation time but tends to have less data diversity and leads to early convergence meanwhile, the larger size resulting the opposite. This parameter is adjusted depending on the number of generations with a maintained ratio of 1/5 (Ruiz-Lendínez et al., 2017).

The next parameter is the number of generations which represents how many iterations are done to each population of solutions to find the optimum result (Hemanth Sai Kumar, 2023). The strategy to optimize this parameter was done by evaluating the fitness value of each generation in the training set.

The crossover operator and the mutation operator for the calculation were represented with the alpha value (α) and the beta value (β) respectively. It was the parameter that ensured the process of crossover and mutation was set to a desired level of exploration and exploitation. For this research, the α value was set to 0.5 to ensure the same chance for exploration and exploitation and to prevent early convergence (Herrera et al., 1998). The last parameter was mutation probability which represents the frequency of mutations to happen in one process calculation. This parameter will be tested together with the mutation operator (β) in the validation set using the same strategy as the number of generations.

3.3.3.7. Model Testing Strategy

The optimized weight from previous steps was used to calculate the geospatial matching between the unplotted land parcel and the locations available in the parcel database. Those parcels then will be tested to two parcel databases: the sample data and the Indonesian cadaster database (The KKP Database). Due to the huge number of land parcels to be tested in the KKP database (±118.000 parcels), only 30% of the test set will be selected by maintaining the ratio of the category to ensure equal representativeness of training data. For the parcels that are tested in the sample data, several variable combinations will be used and their precision-recall value will be evaluated together with the calculation time to find the optimum variable combination and understand the influence of each variable on the model's performance. Figure 13 shows the complete steps of performance testing done in this research.

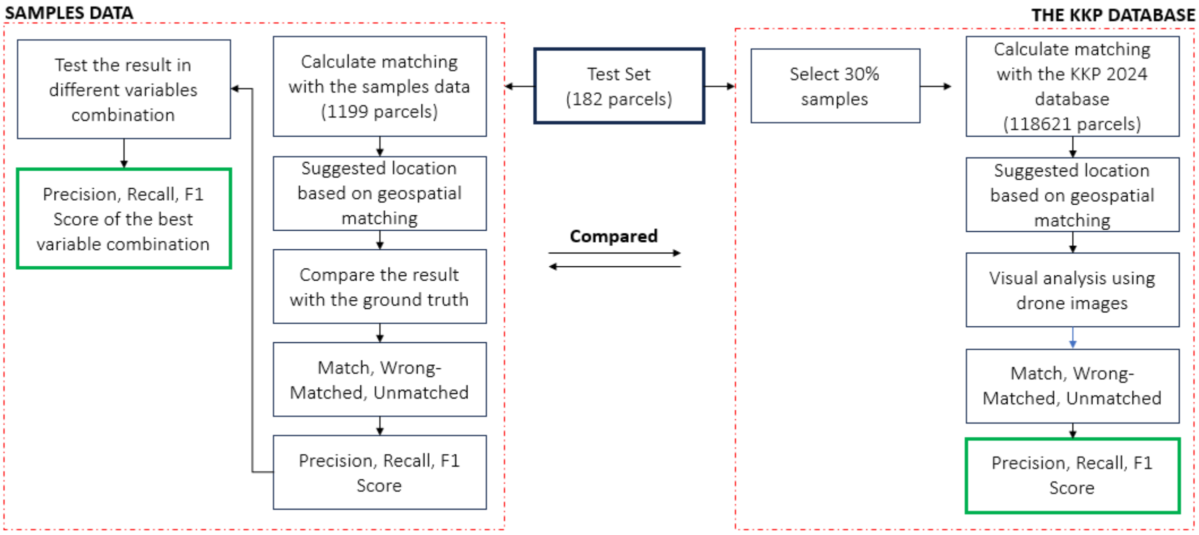

SAMPLES DATA
Test the result in different variables combination
Precision, Recall, F1 Score of the best variable combination
Calculate matching with the samples data (1199 parcels)
Suggested location based on geospatial matching
Compare the result with the ground truth
Match, Wrong-Matched, Unmatched
Precision, Recall, F1 Score
Test Set (182 parcels)
Compared
THE KKP DATABASE
Select 30% samples
Calculate matching with the KKP 2024 database (118621 parcels)
Suggested location based on geospatial matching
Visual analysis using drone images
Match, Wrong-Matched, Unmatched
Precision, Recall, F1 Score

Figure 13. Model Testing Strategy

3.3.3.8. Geospatial Data Matching Using Textual Attribute

To enhance the accuracy of the testing, textual information from the neighboring parcel's attribute was used as an additional prompt to find the location. The attribute matching was done to the top 3 polygons with has highest matching score from the previous calculation. The strategy was taken to minimize the calculation time which may increase if the search was done to the whole set. This strategy was also inspired by the manual process of plotting land parcels.

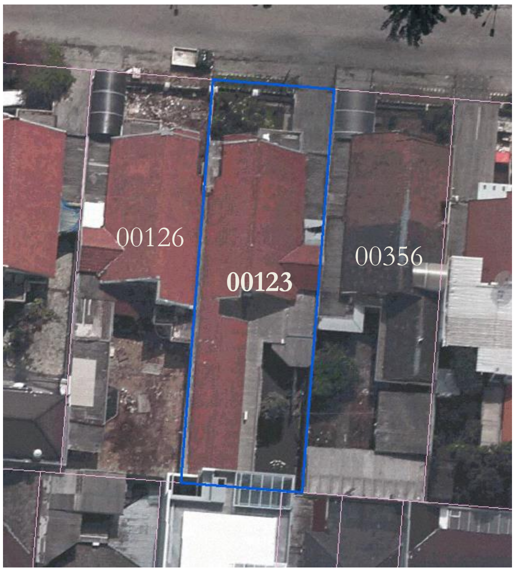


Figure 14. Textual Matching using Neighboring Parcel's ID

Figure 14 illustrates how the identification number of a parcel can be used to detect its relative location. For the parcels tested in the Cadaster database, the textual information used was the Parcel Identification Number (NIB, village name information, and the registered area information. To protect the private information of the parcel owners, the original parcel identification number will be replaced by a made-up number with a similar format.

3.3.3.9. Evaluation Method

The model's performance was analyzed using precision-recall analysis from the test data concerning the reference data. The concept of precision and recall is to minimize the incorrect matches and unmatch and maximize the correct matches from the result (Xavier et al., 2016). The result of geospatial data matching can be considered the correct one if the location selected is the same as the ground truth meanwhile, the wrong-matched result is the opposite. The unmatched result is a condition where the model cannot differentiate between several chosen locations on the ground. The situation will create more than one location suggestion.

The concept was taken from the precision and recall analysis using true positive, true negative, false positive, and false negative. The correct matches will shown as true positives, the false positive value was obtained from wrong matches and the false negative value is the result of unmatched. Thus, the formula for calculating Precision and Recall for matching features is defined in these equations:

$$Precision = \frac{corect}{correct + wrong} \quad (12)$$

$$Recall \quad = \frac{corect}{correct + unmatched} \quad (13)$$

$$F1 \quad = \frac{2\,(precision * recall)}{precision + recall} \quad (14)$$

3.4. Ethical Considerations and Risk

Ethical concerns that may appear during this research have been identified and mitigated to prevent the risks that might happen to the human participants of this research. Including the process of the semi-structured interview and the training data which may contain sensitive information about parcel ownership.

In the interview process, there might be a risk of feeling unease and fear of being judged negatively by the workplace. To mitigate this during the interview, the respondent's identity was not revealed and the questions in the interview did not enlist any personal information. The main purpose of the interview and how the data was handled to ensure respondent security were also explained before the interview. Related to training data management, there was a concern about the disclosure of private information related to land parcels. To mitigate this, the land parcels used in the research did not contain any information about the owner, including the name, ID, and address. The land parcels were identified using a different identifier to prevent tracking from their real parcel numbers.

Both the interview and the training data were stored in a personal cloud drive provided by ITC and only can be accessed using a two-step authorization. Only the researchers could access the original data. The data was also backed up in the personal drive of the researchers equipped with a password to prevent unintended access.

3.5. Chapter Summary

This chapter explained the reason for choosing the study area and the methodology used in the research for every phase that has been passed. The pre-fieldwork phase involves the process of initial code building and building the question to be asked in the next phase. In the fieldwork phase, primary data collections were done using an open-ended interview together with the secondary data collection from the local land office. The last phase was the data processing included the process of qualitative data analysis, preparing the dataset, calculating weight using RCGA algorithms, testing different options to optimize the parameters, and evaluation strategies to obtain the model's performance. The next chapter will present the result of every step done in the methodology section.

4. RESULTS

4.1. Cause of the Unplotted Land Parcels

The cause of the unpotted land parcels was analyzed through the coded interview result that was strengthened by the statistical number from the Surabaya City cadastral quality improvement process result.

4.1.1. General Cause of the Unplotted Land Parcels

The cause of unplotted land parcels that were gathered from the open-ended interview are coded based on it's the saturated theme from 11 respondents. Figure 15 shows the number of quotations from the interview on what are the main problems that create a condition of unplotted land parcels. Land Parcels that were tied to a local control point or to natural features that already changing becoming the main cause of this situation together with insufficient information on the available documents.

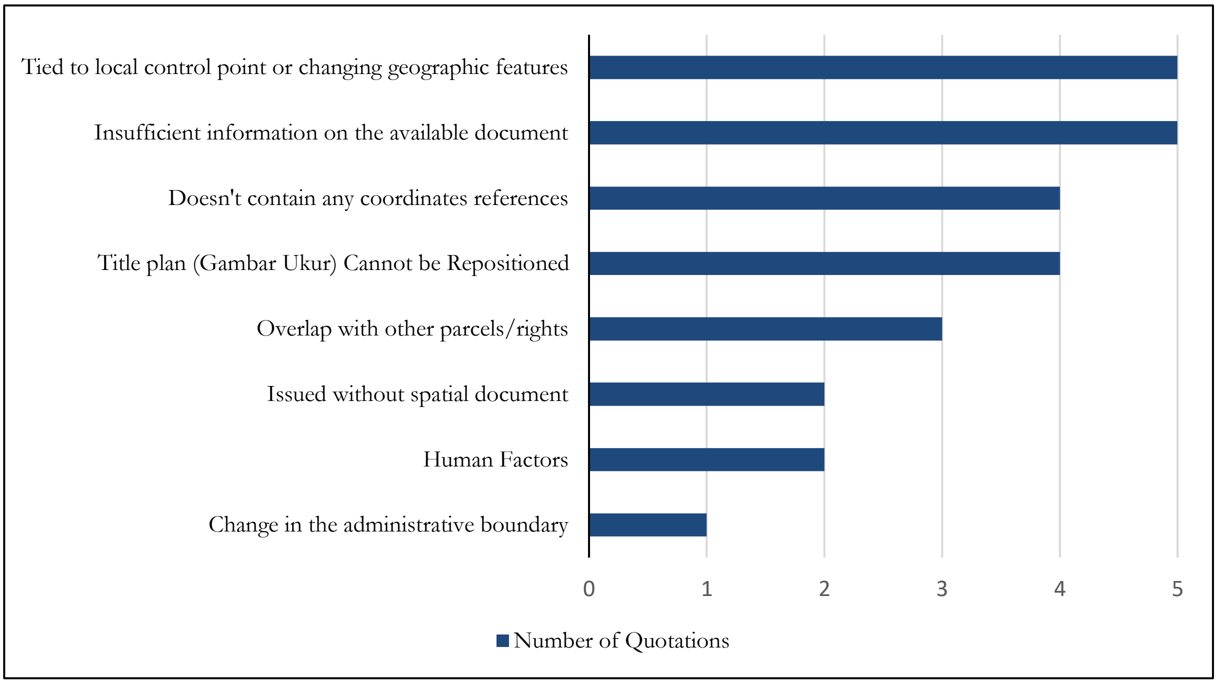


Figure 15. The Cause of the Unplotted Land Parcels

The first condition was explained by one of the respondents who has experience in leading the project *"This might happen because in the past measurement that tied to a local coordinate system was allowed due to the limitation of national control points."* The other respondents who are responsible for creating the regulations also explained: *"The Surveyors in the past, were using geographic features such as river and made an approximation of the distance between land parcels and that geographical features without attaching it into proper coordinate systems."*

The other main cause of unplotted land parcels that significantly arose from the respondents was insufficient information on the available documents. The documents could be spatial documents such as the Title Plan (*Gambar Ukur*) and the Measurement Letter (*Surat Ukur*) and also the textual documents such as the Land Book (*Buku Tanah*). The information that is lacking may vary from the unavailability of scale information or the coordinate system until the complete attribute that should be provided by the Surveyors in the Measurement Letter as explained by a respondent that is experienced in doing the improvements process *"What consumes the most times is to land a parcel that has incomplete information such as textual attributes."*

4.1.2. Condition of The Unplotted Land Parcel in Surabaya City

The result from the interview was analyzed using the result of the cadastral quality improvement process in Surabaya City. Based on the result, from 7915 land parcels 53% of them only have at least one spatial document either was title plan (*Gambar Ukur*) or measurement letter (*Surat Ukur*). From that category, the main cause of unplotted land parcels is insufficient information on the measurement letter with 45% of parcels. There are two causes of unplotted land parcels that emerged during the interview process which percentage was not available in the results.

The other unique condition that emerged in the interview was the land rights that were issued without any spatial document. This takes 47% proportion of the manually plotted land parcels or equal to the amount of 3576 parcels. The category can be divided into two more types, land rights that only have textual documents on them and a small number of land rights that were not equipped with any spatial or textual document. Figure 16 shows the visualization of the document condition and its relation with the cause of the unplotted land parcel in the city of Surabaya.

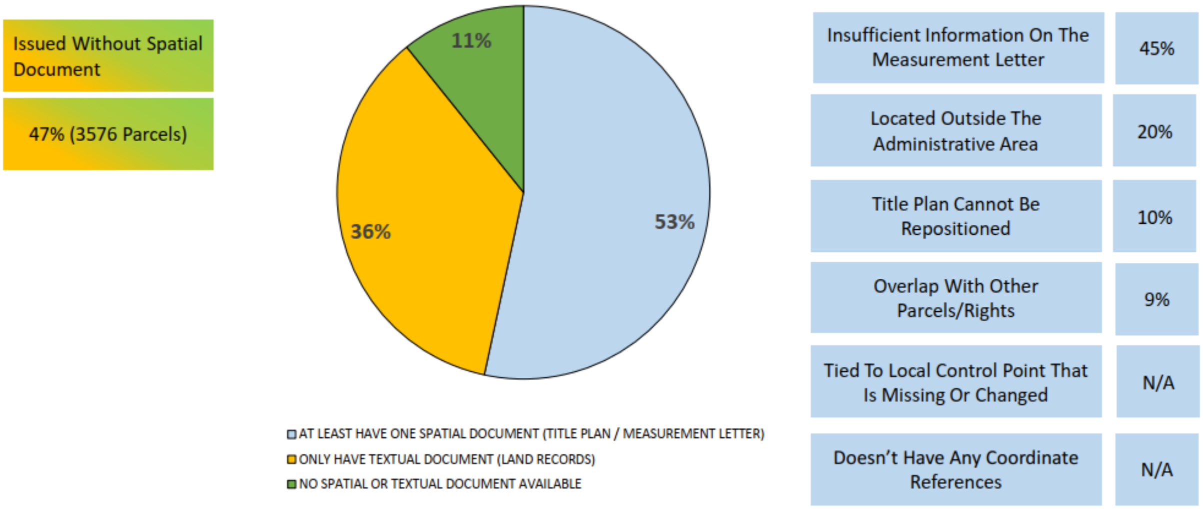


Figure 16. The Data Condition of The Unplotted Land Parcels in Surabaya City

4.2. Current Method to Plot the Land Parcel Manually

In general, two methods were used in the cadastral quality improvement process to plot the unplotted land parcels. The first one was the studio process which was related to a process sequence that was done in the office. The studio process only relies on the available data of the land parcels with the help of some supporting documents such as the tax map. If the studio process is not enough to plot the land parcels, the second method was chosen which is called the field verification process. Despite the process involving the activity that happened on the field, no re-measurement was done to the land parcels. This process involves the external party such as the owner of unplotted land parcels, or the local government who in some areas could be represented by the local elders. The information gathered from both sides was used to plot the estimated land parcel location on an aerial map. The result was plotted in the Cadaster database and used as a prompt to other neighboring unplotted land parcels.

Those methods, meanwhile didn’t ensure that all of the unplotted land parcels were successfully landed on the Cadaster database. That’s why at the end of the process, there will be a list of the unvalid land parcels and the unsolved land parcels to accommodate the special case as told before. Figure 17 visualizes the complete process of manual plotting for the unplotted land parcels.

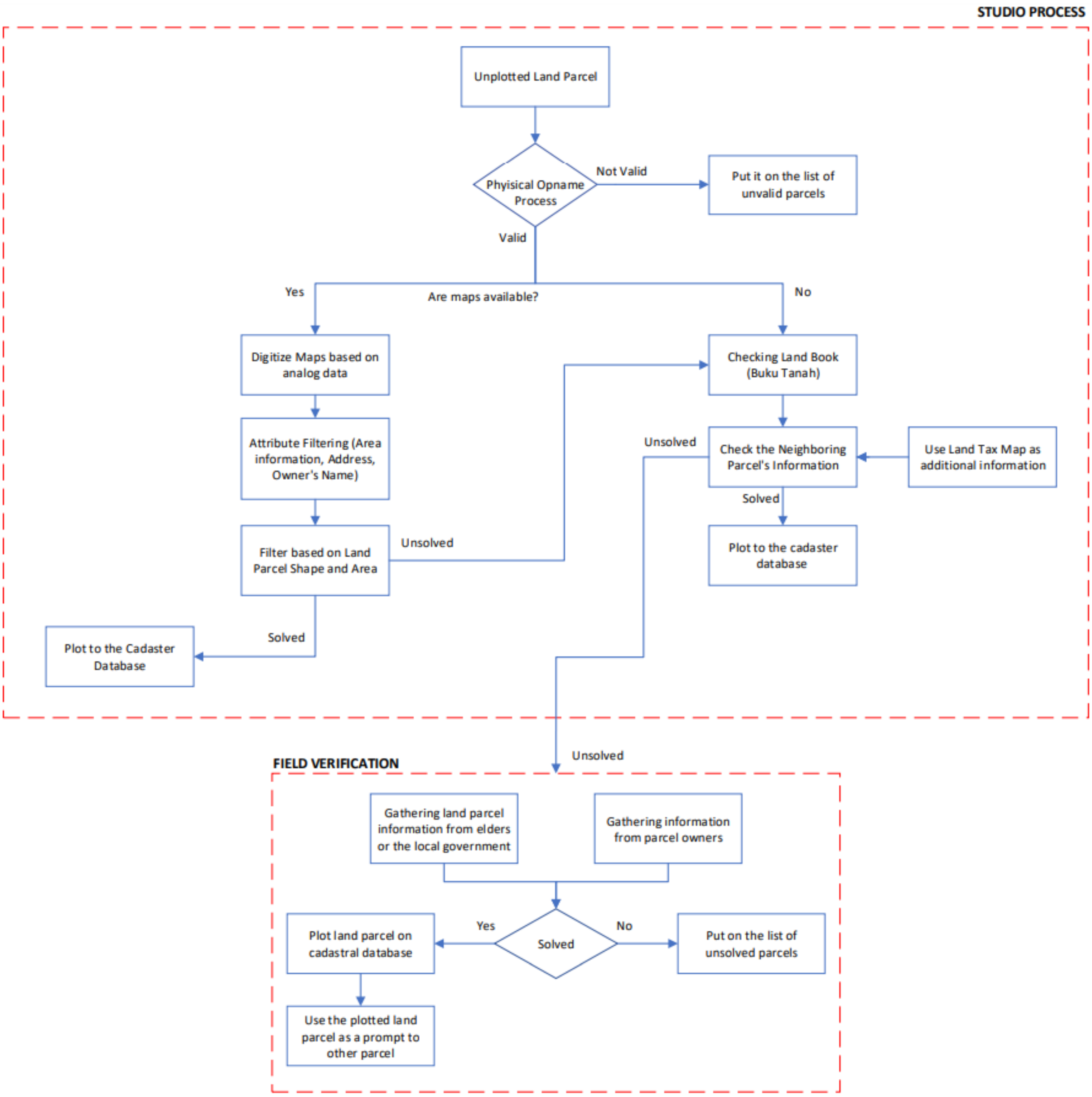

STUDIO PROCESS
Unplotted Land Parcel
Phyisical Opname Process
Not Valid
Put it on the list of unvalid parcels
Valid
Are maps available?
Yes
No
Digitize Maps based on analog data
Attribute Filtering (Area information, Address, Owner's Name)
Filter based on Land Parcel Shape and Area
Unsolved
Solved
Plot to the Cadaster Database
Checking Land Book (Buku Tanah)
Check the Neighboring Parcel's Information
Use Land Tax Map as additional information
Unsolved
Solved
Plot to the cadaster database
Unsolved
FIELD VERIFICATION
Gathering land parcel information from elders or the local government
Gathering information from parcel owners
Yes
Solved
No
Plot land parcel on cadastral database
Put on the list of unsolved parcels
Use the plotted land parcel as a prompt to other parcel

Figure 17. Method to Plot the Land Parcels Manually

This research will develop a machine-learning model to automate the manual process of plotting land parcels when the map archive is available. This model involves the process of attribute and shape matching which will be explained in the next part.

4.3. Design of the Machine Learning Model to Plot the Land Parcel

Based on the interview process about the heuristic process of finding the location for land parcel plotting, a model based on machine learning was designed to semi-automate the current process. This automation process involves the optimation calculation using the RCGA algorithm to weigh the geometric variables. The use of attribute matching from the land parcel was used after the geometric matching to enhance the model's performance in finding the location. Figure 18 illustrates the model's architecture in a graphic.

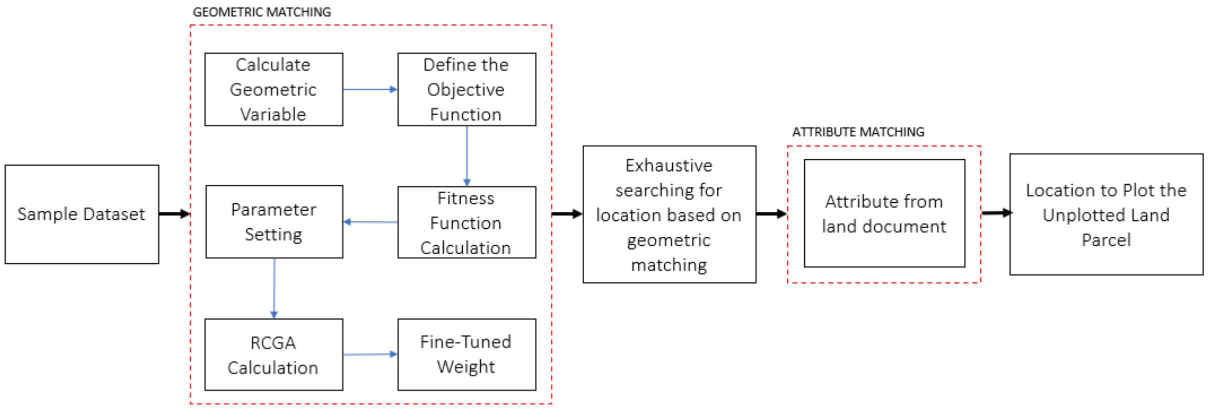

GEOMETRIC MATCHING
Sample Dataset
Calculate Geometric Variable
Define the Objective Function
Fitness Function Calculation
Parameter Setting
RCGA Calculation
Fine-Tuned Weight
Exhaustive searching for location based on geometric matching
ATTRIBUTE MATCHING
Attribute from land document
Location to Plot the Unplotted Land Parcel

Figure 18. Model's Design to Find the Plotting Location

Five matching variables were chosen in the geometric matching process: perimeter, number of vertices, MBR, the Arkin Graph Angle, and polygon's area. These variables were selected based on two prompts elaborated by the interview respondents. The first prompt was used of the parcel's shape, highlighted by respondents who mentioned, *"We are doing a one-by-one checking based on the visual shape of the land parcels."* and "*We detect the position through visual representation in the aerial photography maps.*" The human's ability to detect an object's shape was translated into mathematical variables that a computer could understand such as perimeter, number of vertices, MBR, and the Arkin Graph Angle (AGA). The process of defining and calculating those variables was explained in the methodology chapter. The second prompt gathered from the interview was the use of Parcel's Area. A respondent who worked as a surveyor noted, *"We do a filtering based on the area of the parcels and compared it with the available maps such as tax map."* The parcel's area was calculated using the formula of polygon's area, as the parcels in the database are polygonal in shape.

Using the formula of calculating a geometric match between two geospatial data, an exhaustive search was done in the object database to find the three locations with the highest geometric matching score. Subsequently, attribute matching was done to those selected locations to find the best-fitting location for plotting. The method of attribute matching was repeatedly mentioned by the respondents, *"We are looking into the legal documents such as the land book to find the information about the registered area, the owners, and the parcels's identification numbers."* And *"We are using the address information from the measurement letter to find the approximate location on the tax maps."* The attribute checking was done not only to the unplotted land parcel but also to the surrounding parcel document, if available. As one respondent noted, *"We are also checking the information from the neighboring parcels such as the parcel identification number and the documents number."* Based on the interview, three textual attribute matching were selected for testing: village name, neighboring parcel's NIB (parcel identification number), and the registered area information.

4.4. Fine-Tuned Weight of the Geometric Variables

The weight of each geometric variable was determined using the RCGA optimization algorithm. Before getting to the final result, several steps affect the result such as dividing the dataset, tuning the RCGA parameter, and finally the result of fine-tuned weight. The result from each step is explained separately in different parts below:

4.4.1. Sample Dataset

The results of the categorization of unplotted land parcels based on their condition compared with actual rights on the ground are shown in Table 3., Category 2 represents the condition of unplotted land parcels that have almost similar shapes to the actual rights had the highest percentage with 44.9% which was equal to 538 land parcels. Meanwhile, Category 5 which represents the condition of the unplotted land parcel that has been subdivided on the actual rights became the category with the smallest proportion with 5 land parcels.

Table 3: The Number of Parcels for Each Category

Category	Number of parcels	Percentage
Category 1	457	38.1%
Category 2	538	44.9%
Category 3	136	11.3%
Category 4	63	5.3%
Category 5	5	0.4%

From that categorization, the sample datasets were divided into three sets: test set, validation set, and test set. The division by keeping the constant proportion for each category to ensure representativeness, training sets have the highest percentage with 70% followed by the validation and test sets with a percentage of 15% each. Table 4 lists the full description of each set.

Table 4: The Dataset to Build the Model

Dataset	Number of parcels	Percentage
Training Set	837	70%
Validation Set	180	15%
Test Set	182	15%

4.4.2. RCGA Parameter Tuning

Before doing the training for the model, several parameters need to be tuned using the strategy as explained in the methodology section. Figure 19 shows the result of tuning the num_generations parameters and its zoomed version to refine the optimum value. The test was done to the model until the 400 generations, at the first 50 generations it shows that the fitness value drastically changed from 0.45 to 0.70 which shows the better solutions for the problems. After 50 generations it fluctuated until reached the highest fitness value within the range of 250 – 300 generations. If it is zoomed the model reaches its best performance on 240 generations. After 240 generations, the fitness value fluctuated again but never touched the same peak by sacrificing a longer time to do the iterations. Based on the model's efficiency, 240 was chosen as the optimum num_generations to be used in the model training.

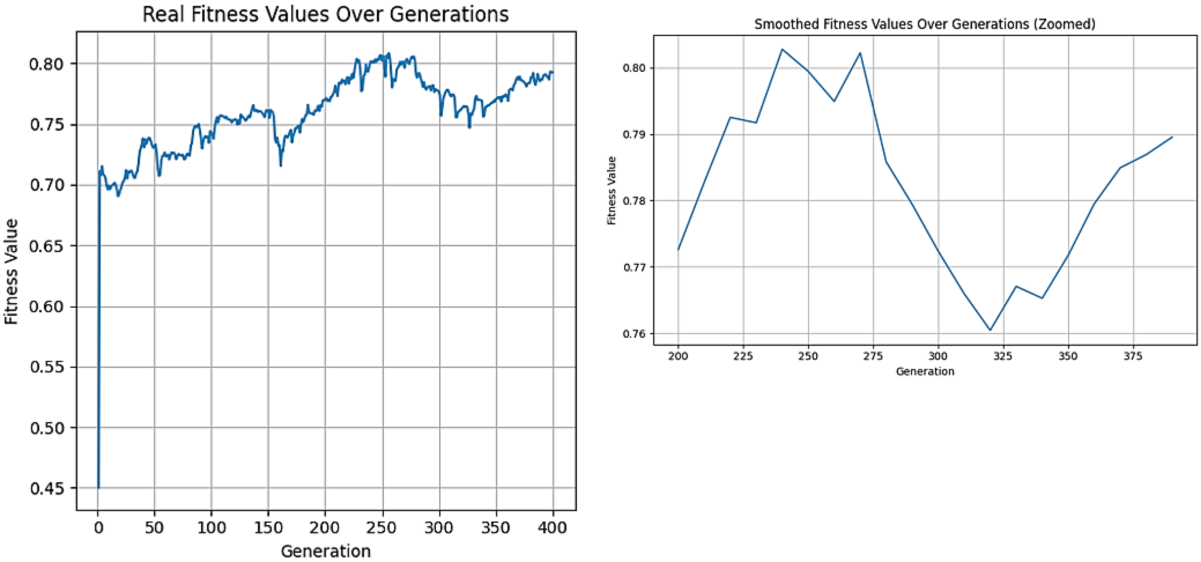


Figure 19. Model's Fitness Value Over Generations

The parameter tuning was also done to the mutation-related parameters namely mutation rate (β) and the mutation probability (MP). The test was done firstly to every possible combination for both parameters then the top 4 combinations that give the highest average fitness value were selected. Figure 20 shows the result of those 4 combinations over the generations.

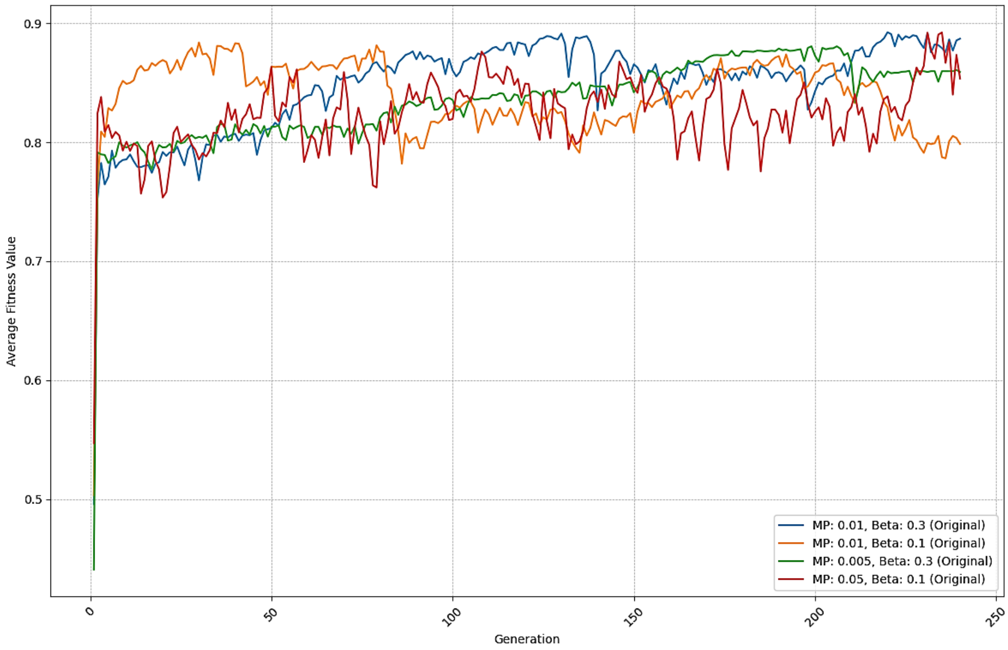


Figure 20. The Top 4 Combinations of Mutation Parameter

From the graph, it can be obtained that the blue line was higher on average compared with the other lines. In the optimum generations of 240, the blue line also has the highest average fitness value amongst others. That line represents the combination of MP equal to 0.01 and β with the value of 0.3. The orange line which represents the combination of MP = 0.01 and β = 0.1 shows better results in the early stages of calculation but if the generations continued to the optimum value of 240, the fitness value decreased significantly to the lowest amongst the four combinations. The use of MP = 0.05 and β = 0.1 shows the extreme fluctuation over the generations which might cause an instability in the model's performance.

Combined with other parameters that have been set to a certain value as mentioned in the methodology section, Table 5 lists the optimum parameters that will be used in the model training for calculating the optimum weight for every geometric matching variable.

Table 5: The Optimum RCGA Parameter Value

RCGA Parameter	Optimum Value
Initial Population Size	48
Number of Generations	240
Chromosome Length	5
Mutation Probability	0.01
α value	0.5
β value	0.3

4.4.3. Fine-Tuned Weight for Geometric Variable

The fine-tuned weight for every geometric variable was calculated using the RCGA optimization function with the architecture as explained in the previous section. The results shown in Table 6 were obtained after 240 successful generations of an offspring solution to find the optimum fitness value for the optimization problem.

Table 6: Fine-Tuned Weight for Geometric Variable

Geometric Variable	Weight
Polygon's Area (A)	0.139
Number of Vertices (n)	0.304
Perimeter (P)	0.197
Minimum Bounding Rectangle (MBR)	0.222
Arkin Graph Angle (AGA)	0.138

From the table, it can be obtained that the number of vertices (n) variable had the highest weight among all indicating an important role in geometric matching. The Arkin Graph Angle (AGA) which is the variable that ensures the scalability of the polygon becomes the lowest affecting variable together with the polygon's area with the respective weight of 0.138 and 0.139. The discussion about the correlation of this result with the data condition will be explained in more detail in the next section.

4.5. Testing the Model Performance

This part will explain the result of the model performance testing divided into two parts, the first one is the data tested in the samples data and the second is the data tested in the KKP Database.

4.5.1. Testing the Model with the Samples Data

In this testing phase, 182 parcels in the test dataset are matched against the 1199 land parcel data in the sample dataset. The combination of variables was also tested by considering its weight from the previous part. Variables with the highest weight were put first on the combination followed by other lower-weight variables. Table 7 shows the complete result of testing the model with the sample data.

Table 7: The Testing Result with the Samples Dataset

Variables Combination	Correct	Wrong	Unmatched
2 variables (n, MBR)	12	0	169
3 variables (n, MBR, P)	134	1	46
4 variables (n, MBR, P, A)	155	3	23
5 variables (n, MBR, P, A, AGA)	158	2	21

As listed in the result, using only 2 variables is not sufficient for the model to accurately match the shape of unplotted land parcels and the available parcels on the dataset. It is shown by the high number of unmatched parcels. By adding more geometric variables, the result of unmatched parcels was stepped down and the result of the correct matching was increased. To choose the suitable combination of variables the parameter of F1 score compared with the cost of calculation was used, Table 8 shows the result of precision, recall, F1-score, and the average calculation time for each combination.

Table 8: Precision-Recall Analysis with The Average Calculation Time

Variables Combination	Precision	Recall	F1-Score	Average Calculation Time (s)
2 variables (n, MBR)	100%	6.63%	0.12	1.59
3 variables (n, MBR, P)	99.26%	74.44%	0.85	1.60
4 variables (n, MBR, P, A)	98.10%	87.08%	0.92	1.62
5 variables (n, MBR, P, A, AGA)	98.75%	88.27%	0.93	1.98

The results show that using the combination of 4 variables doesn't affect the model's performance compared with the use of 5 variables with a difference of 0.01 in the F1-score. Meanwhile, in the average calculation time, there was a significant rise between those two combinations with a 0.36-second difference in processing time for one land parcel. To delve more into the analysis, the next chapter will show the result of testing the model in the KKP Database.

4.5.2. Testing the Model with the KKP Database

In this testing strategy, 57 land parcels were chosen from the test dataset to be matched against 118621 parcels in the KKP Database. The results will be compared with the previous result on sample data to obtain the best combination of variables, Table 9 lists the testing result of the KKP Database together with its Precision, Recall, and F1 Score using two combinations of variables.

Table 9: The Testing Result with the KKP Database

Tested on	Correct	Wrong	Unmatched	Precision	Recall	F1-Score
Sample Data (4 variables)	155	3	23	98.10%	87.08%	0.92
Sample Data (5 variables)	158	2	21	98.75%	88.27%	0.93
The KKP Database (4 variables)	27	2	28	93%	49%	0.64
The KKP Database (5 variables)	31	3	23	91%	57%	0.70

Different from the previous testing using sample data, the difference between using 4 variables and 5 variables combination was higher in the KKP Database. The combination of 4 variables generated the F1-Score of 0.64 meanwhile, the addition of 1 variable resulting a 0.06 higher result. The results of higher F1-Score contributed by the slight increase in the Recall Value to 57% indicating more unmatched land parcels that can be identified its similarities with the available location. For the precision value, there was not much difference between 4 and 5 variables with a slight decrease of 2% while running the model with 5 variables.

4.6. Chapter Summary

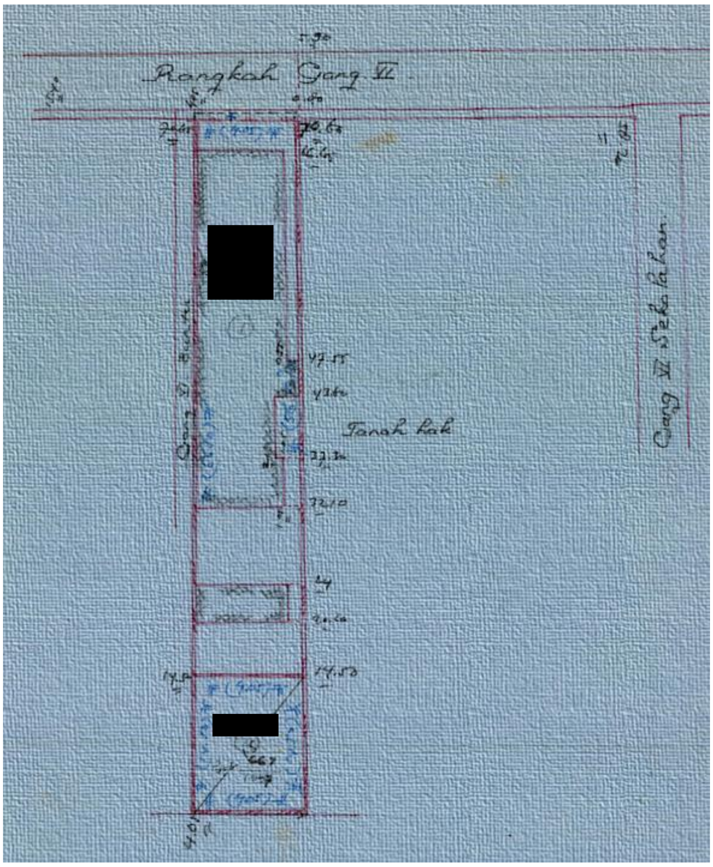
This chapter presents the result of the fieldwork, qualitative analysis, model building, and experiment on the model that has been done in the research. It focused more on geometric matching to find the location to plot the land parcels. The result provided in this chapter was based on the training in the Sample Data and has been tested by two datasets, the Sample Data, and The KKP Database. The next chapter will present the discussion and analysis of the geometric matching result regarding its relation to the land administration aspect, especially the cadaster data condition, and use the other elements of geospatial matching to increase the model's performance. The next chapter will also present the factors hindering the implementation of this model related to the data condition.

5. ANALYSIS AND DISCUSSION

5.1. Lesson Learnt From The Past: The Previous Condition and Impact on Current Cadastral Data

The semi-structured interview with the expert gave an important insight on what is the cause of the unplotted land parcel especially in a land parcel that was equipped with spatial data but the original position remained unknown. There are two causes that most of the respondents appealed to in the interview: tied to a local reference point or a changing geographic feature and insufficient information on the available document. This part will analyze the cause more deeply based on the actual document conditions.

For the first cause of being tied to a local reference, it was stated in the old regulation that the measurement of the base point in local coordinates is allowed due to incapability (that may be caused by many reasons), but the local base point should be transformed into the national control point in the future time (The Minister of Agrarian/Head of National Land Agency, 1996). In principle, this regulation had already implemented the concept of Fit-For-Purpose Land Administration (FFPLA), especially in the idea of doing a step-by-step refinement of the data (Enemark et al., 2021). But in practice, many land parcels remained in local references or didn't contain any coordinate references due to some limitations (Ary Sucaya, 2009), creating the condition known in the present time as the unplotted land parcels. Despite the regulation having been changed with several revisions in the following year, this type of parcel still appeared and became the top cause of unplotted land parcels. Figure 21 gives an example of the title plan (Gambar Ukur) of land parcels that were mapped in local references.



[redacted]

[redacted]

Figure 21. Title Plan (Gambar Ukur) of Land Parcels in Local References

From the images, it could be obtained that most of the unplotted land parcels lacked of position information that was needed for re-plotting. The example showed the parcels that only tied into neighborhood parcels which may change in shape or ownership and only had two street information which may change also in the name or location compared with the present condition. This situation brought a big challenge to the Surveyors to find the place to plot those analog maps into the current digital cadaster database as also mentioned by one respondent *"There was a change in the boundary so we cannot find the relationship with adjacent boundaries and creating difficulties for the plotting process."*

The second cause of the unplotted land parcels was insufficient information on the available documents, especially the information related to the position of land parcels. This situation can be explained by an example from Figure 22 which represents the title plan (Gambar Ukur) that was mapped without sufficient location information.

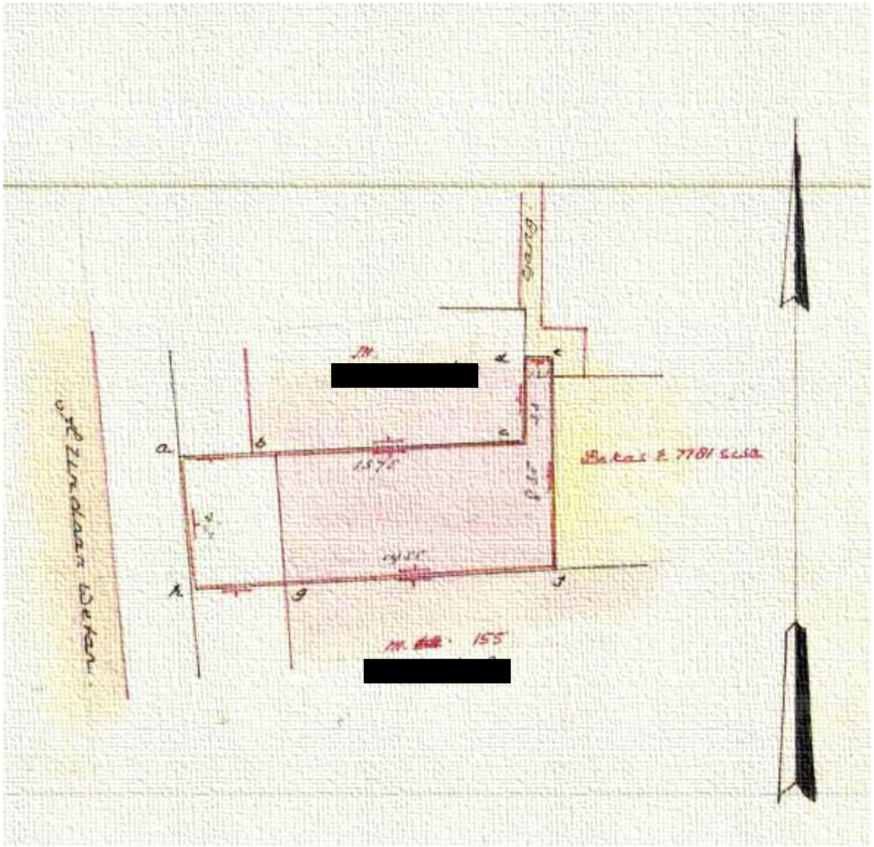


Figure 22. Title Plan (Gambar Ukur) of Land Parcels without Sufficient Position Information

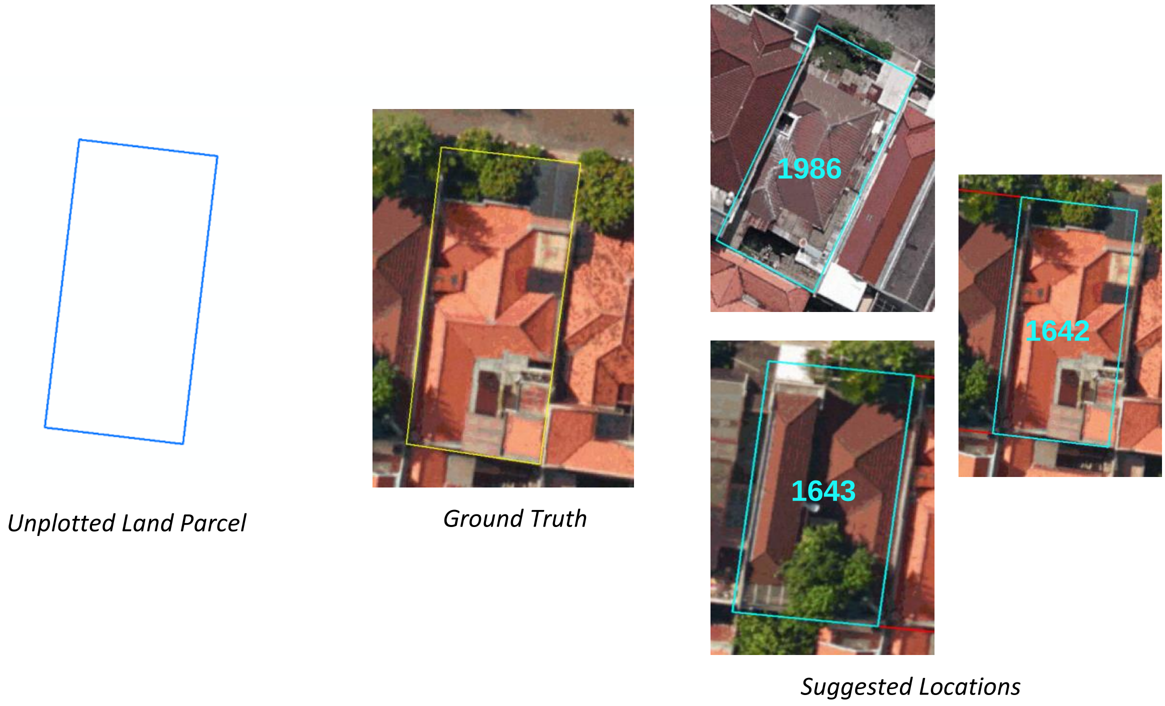
The image shows the example of the maps without any required spatial information such as coordinates grid, coordinate system information, measurement value to tie the parcels with the nearest control point, and the relative location information of the land parcels. The example shows the extreme condition where all of the required elements were missing, in the real condition those four missing attributes did not always come together in one map but still created difficulties for the surveyors in finding the correct location in the current database. A respondent who is experienced in doing the cadastral improvement process says *"It was taking most of the time to check the documents manually one by one to find the location of the unplotted land parcels."*

From both examples on the possible cause of the unplotted land parcel, the only prompts that can be used were the parcel shape and its relative position to the neighboring parcels or available geographic features such as roads. That was the reason for the importance of having the position information of neighboring parcels in the cadaster quality improvement process because it can help the surrounding parcels to find their location. It is also important to have information on the relative position of the land parcels with available geographic features because it can help the plotting process as the neighboring parcels did. The effect of including those two important attributes in the Cadaster database was discussed more in the recommendation chapter.

5.2. Learning from the Present Data: The Correlation of Model's Result with The Cadastral Data Condition

5.2.1. Matching Results based on Geometric Matching

The model's performance was satisfying when tested with the sample dataset. From 182 parcels that were tested using the combination of five variables, 158 of those were successfully matched while only 2 were mismatched from the right locations on the ground and showed a precision value of 98.75%. Despite the good result, the number of parcels that cannot be differentiated in this combination was still high with 21 parcels. It contributes to the results of 88.27% recall value. Going deeper into the condition of unmatched land parcels, Category 3 which represents the condition that many available locations on the ground had similar shapes to the unplotted land parcels' shape has the largest proportion of the unmatched result. The matching technique that only relies on the geometric features cannot differentiate those conditions on the ground and may decrease the model's performance. Figure 23 shows the example of the unplotted land parcel with category 3 that resulted in an unmatched result.



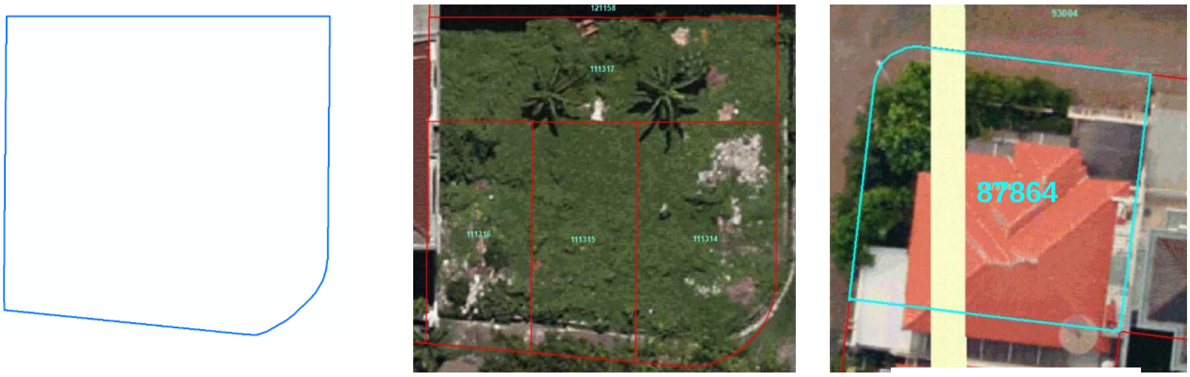
Unplotted Land Parcel

Ground Truth

Suggested Locations

Figure 23. Unmatched Prediction on the Similar-Shape Parcels

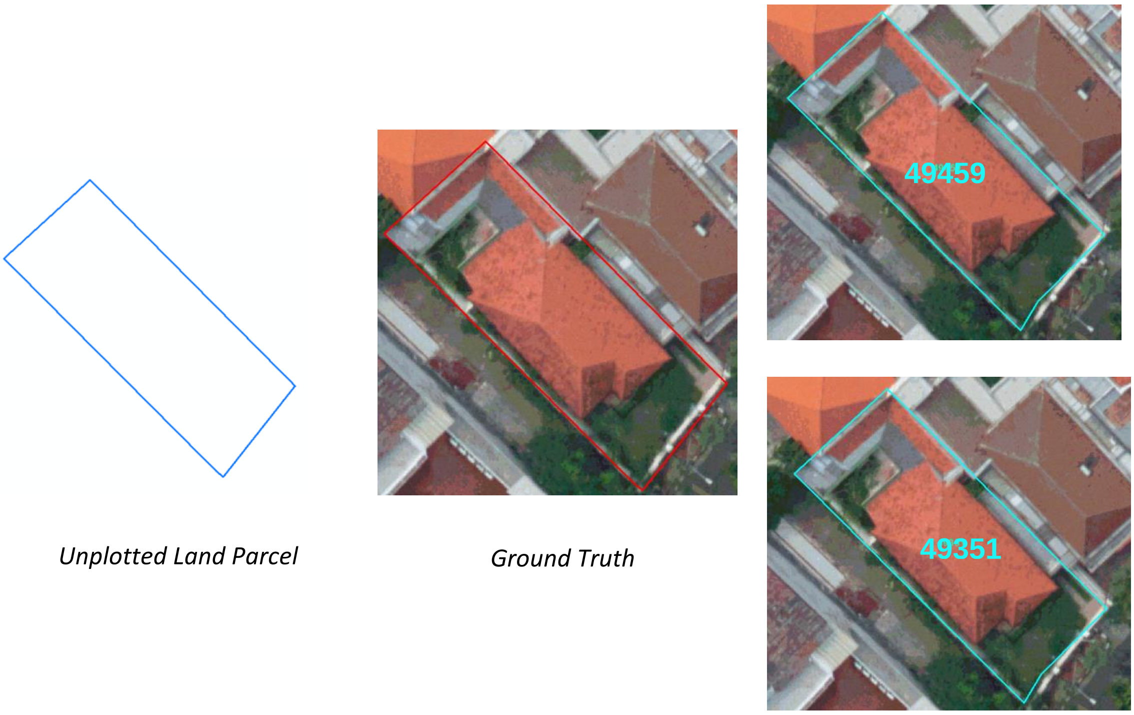
The same situation appeared when the parcels were being tested on the KKP Database, 31 parcels were successfully matched into their locations on the ground while only 3 parcels were incorrectly matched and generated a 91% precision value. All parcels that the locations failed to identify were the parcels that have been subdivided in the present database. By only relying on the geometric condition, it is impossible even for humans to retrace the original locations of unplotted land parcels. Figure 24 shows an example of how the subdivision affects the model's performance.



Unplotted Land Parcel *Ground Truth* *Suggested Location*

Figure 24. Wrong Prediction on a Subdivided Land Parcel

Despite the satisfying precision result, the percentage of land parcels that were unmatched was higher when tested in the KKP Database. There were 23 out of 57 parcels that couldn't be considered its actual location on the ground which resulting a 57% recall value. The complexity of the parcel's condition and the large number of possible candidates to be matched might be affecting the model performance and creating a low recall value. There are two conditions related to the data: overlapped rights on the ground and the similar shape of the candidate location. The second condition was not delved into because it showed the same pattern as the previous testing, Figure 25 is more focused on showing the condition of two overlapped suggested locations from the testing on the KKP Database.



Unplotted Land Parcel *Ground Truth* *Suggested Locations*

Figure 25. Unmatched Prediction on the Overlapped Rights

It needs to be understood that the Sample Data was a prepared data that was not only used for data testing but also for the data training, and parameter validation. So, it made sense that the condition of overlapped locations was not found in the testing phase using The Sample Data. Meanwhile, the KKP Database is an actual database that was downloaded directly from the systems. It makes this dataset not overlap-free and redundant-free which might represent the real data condition in Indonesian Cadaster. The ability of the model to determine the locations of overlapped suggested locations was quite satisfying, but because the matching was done based on the Parcel's NIB it naturally created an unmatched result when there were two or more overlapped features on the ground. The model based on only geometric matching cannot determine which NIB fits best the unplotted land parcels., there is a need to use another matching technique to differentiate these conditions.

5.2.2. Enhance the Model Performance Using Attribute Matching

The element of geospatial matching not only consists of geometric matching but also includes other matching elements such as attribute matching (Xavier et al., 2016) opens an opportunity to increase the accuracy of selecting the right location for the unplotted land parcels. The model used not only geometric matching to find the location but also involves the textual data from the land parcel's attribute to enhance the model performance. The attribute matching process was selected from the top three locations suggested by geometric matching to find the location for plotting the unplotted land parcels. The model was tested using three different attributes: registered area, village name, and the NIB of neighboring parcels. The result of textual attribute matching is presented in Table 10.

Table 10: The Model's Performance after Attribute Matching

Matching Strategy	Correct	Wrong	Unmatched	Precision	Recall	F1-Score
Before Attribute Matching	31	3	23	91%	57%	0.70
"Village Name" Matching	43	3	11	93%	80%	0.86
"The Neighboring NIB" Matching	47	3	7	94%	87%	0.90
"Registered Area" Matching	48	4	5	92%	91%	0.91

The addition of textual attributes of land parcels in the matching process has proven to increase the model's performance, especially in differentiating the unmatched land parcels. By adding village-name matching, the land parcels with the same shape but located in different villages can be filtered. It was proven to significantly decrease the unmatched parcels by up to half from 23 to 11 parcels and increase the recall value to 80%.

The next attribute used was The Neighboring NIB, this strategy was used to give a proximity analysis of the plotting position. In this research, the neighboring NIB matching was done using the intersection function of a polygon. The land parcels are considered matched if intersect with a certain neighboring polygon. By using this strategy, showed better results than using the previous one by decreasing the unmatched parcels to 7 parcels and increasing the correct matches to 47. It was resulting a precision value of 94% and a recall value of 87% which was 30% higher than the result before textual matching.

The last attribute was the Registered Area of a Land Parcel, in the Indonesia Land Administration there were two acknowledged area information: the spatial area and the registered area. That two area information should have the same value in a normal condition, but there is a condition that the registered area did not match with the spatial area due to some reasons such as mistype. The use of registered area was to replace the owner's name of the land parcels as a unique attribute of parcels which was not available in this research due to data privacy.

By using the unique attribute of registered area, it was proven to decrease the unmatched land parcel up until its lowest amount of 5 parcels and generated the highest recall value from all experiments with 91%. But in the aspect of precision, it shows a slight decrease of precision value to 92% due to the nature of the attribute. As explained previously, the registered area is manually typed area information that has the potential of mistake due to mistyping, the mistake captured on the model increased the wrong matched parcels to 4 and decreased the precision value compared to other matching strategies. The use of a unique attribute such as registered area, owner's information, or tax number has proven to be effective in eliminating the model to predict two or more overlapped rights. This attribute could eliminate the other candidates because uniqueness of land parcels compared with others. Future improvement on how to maximize the potential use of unique attributes will be delivered in the recommendations part.

5.3. Towards the Future: Possibility of Implementation and Factors Related to the Model Performance

5.3.1. Selected geometric matching variables

In the result phase, the model was tested using several combinations of geometric variables to find the best performance with the fastest calculation time. From the combinations, it can be obtained that the combination of 4 variables and 5 variables gives the best performance for both testing on the Sample Data and the KKP Data. For this reason, this chapter will delve into both combinations to select which compound will be chosen for further model implementations. Table 11. Present the comparison of the results from both combinations tested on two datasets.

Table 11: The Comparison of Variables Combination

Considered Factor	**4 Variables Combination** (n, MBR, P, A)	**5 Variables Combination** (n, MBR, P, A, AGA)
F1-Score on the Sample Data	0.92	0.93 (+1%)
F1-Score on the KKP Database	0.64	0.70 (+9%)
Unmatched Result on the Sample Data (parcels)	23	21 (-8%)
Unmatched Result on the KKP Database (parcels)	28	23 (-17%)
Average Processing Time (s)	1.62 (-18%)	1.98

Despite having faster processing time, the 4 variables combination resulting lower results in the rest of the considered factors. When tested in the Sample Data, the gap between 4 variables and 5 variables was not too obvious with only a 1% difference in F1-Score and an 8% difference in the matching result. But when tested in more complex data with more parcel shape variety, the F1-Score gaps increased by 9% and the performance to differentiate unmatched results increased by 17%.

This performance improvement was influenced by the addition of one variable named the Arkin Graph Angel (AGA). This variable in the background process, took longer times to calculate because involving two steps of calculation: calculating the tangent angles for each vertices and calculating the area below turning functions. Considering the angles for each vertice, made the matching calculation free of the different scale effects which may appear between the unplotted land parcels and the available locations on the ground. That's why adding this variable in the Sample Data didn't affect the performance because both data were set into the same scale at the beginning. Making use of this variable together with other variables is important when implementing the model in the real cadaster database, despite the higher cost of calculation time. Based on the analysis, the 5 variables combination (n, MBR, P, A, and AGA) were recommended to be used in the Indonesian data condition to eliminate the possible mistake due to scale difference.

Nevertheless, the choice of variable to be used in the geometric matching phase depends on the data conditions and the number of parcels to be tested. If the spatial data was already well organized and didn't have any scale difference, the choice of 4 variables would give faster results with not-so-different accuracy results. This is also applied to the application that demands more parcels to be tested so that the small differences in the model's accuracy can be ignored.

5.3.2. Factors Related to The Model Performance

Based on the experiments of the model that have been done on two different test datasets which had different characteristics of data, several factors need to be taken into consideration to get the best result from the model. Those factors are important to directly implement the location searching by using the given parameter and variable weight as mentioned in the Result Part without doing further training which may take time depending on the size of the dataset.

The first factor was related to the input needed by the model which is a digitized unplotted land parcel. This model was run based on the vector analysis to calculate geometric matching for every variable that will be considering the best location by combining it with the attribute matching. That's why, the process of parcel digitization was important in the process of geometric matching. The nature of the unplotted land parcels which the data was only available in analog format, needed an additional step of map digitizing. It is important to have a good quality scanner that can prevent the analog map from shrinking or expanding during the scanning process. The digitization of the scanned maps was also critical to maintaining the original shape of the land parcels. Nowadays, many other implementations of Machine Learning can help to increase the accuracy of analog map digitization.

It is important also to assign correct coordinate references to the digitized maps based on the candidate's location. In the land administration database of Indonesia, there was a division of 16 zones based on the Transverse Mercator System to accommodate the large mapping area. Although this research did not delve into how this mistake affects the result, in theory, it will affect the result calculation for the polygon's area (A) variables and affect the performance of the whole model.

The next factor that is important to be taken care of is the condition of the candidate's location. An experiment was done on another spatial database to understand the effect of location candidates on the performance of the model in searching the candidate location. At least two conditions will affect the model: no available candidate location on the ground and the shape of the candidate location doesn't reflect the real condition.

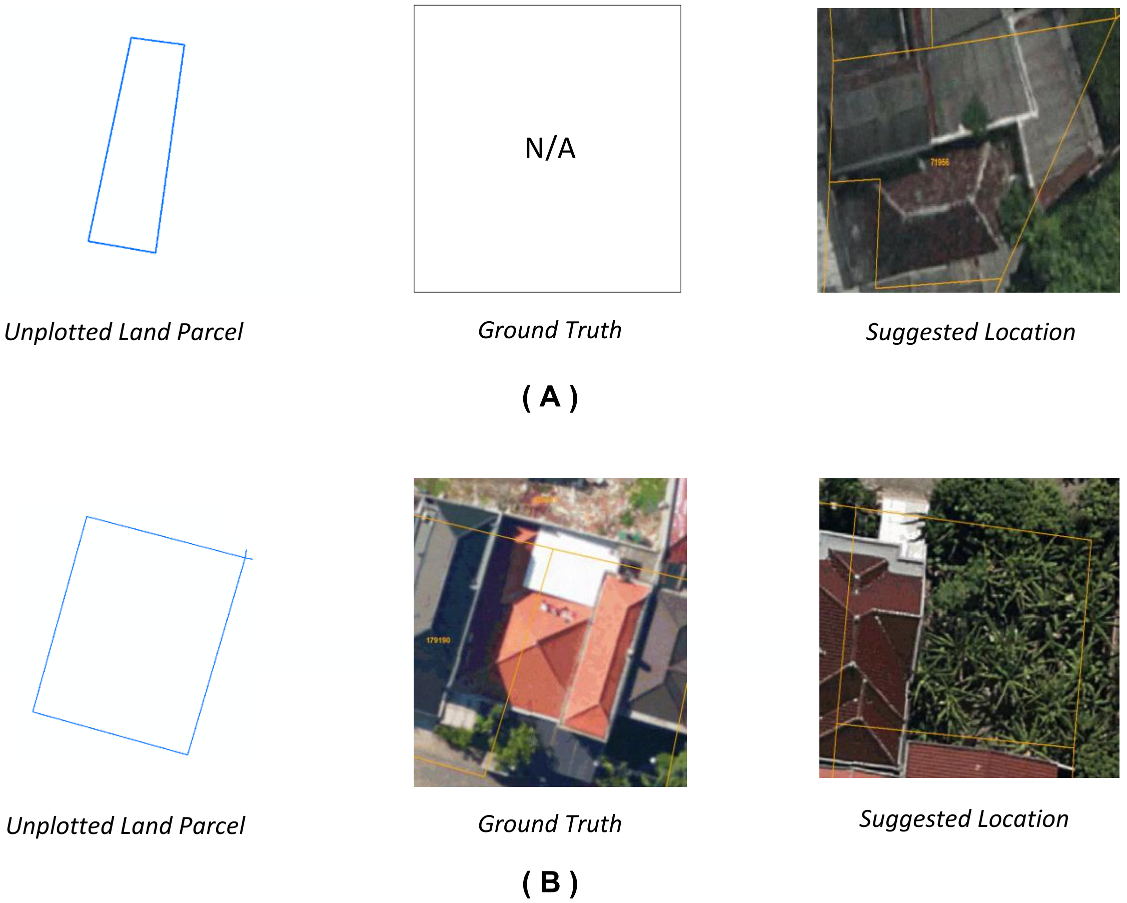


Figure 26. The Effect of Unavailable Candidate Location (A) and The Candidate Location that Doesn't Reflect Real Condition (B) on The Model Performance

Figure 26 (A) shows an example of how the unavailability of candidate locations. Please remember that the model selected the location based on the geometric matching of polygon and attribute matching. It involved two polygons: the unplotted land parcel and the available locations. If the second polygon was not available in the database, the first polygon automatically searched for the other polygon with the highest geometric matching. This could have led to a wrong prediction of the location of the land parcel. It is also reflected in Figure 26 (B) where the available location on the ground didn't reflect the real condition based on the satellite image. The model continued to search and calculate the other polygon with the highest similarity, resulting in a wrong prediction.

The next factor was related to the cadaster data condition of the desired location. The model was trained using the plotted land parcels done in the city of Surabaya. Surabaya is the second largest city in Indonesia, which provided the richness of the data condition. It consists of the urban area, agriculture area, industrial-scale fish pond, industrial area, and some large buildings that have their uniqueness. The choice of the study area with its heterogeneity was expected to represent all of the locations in Indonesia. But still, some parts of the country remain homogenous in land use for example the areas that mainly consist of plantations or farming areas with their unique conditions. Based on this, it can be concluded that this model will be suitable for the big cities in Indonesia with heterogonous land use until the research on the use of this pre-trained model to the other data conditions is done.

The condition of overlapping rights in a location and involving the invalid rights to the available location will affect the model's performance as well. If the invalid rights were not filtered from the beginning, they would have the same opportunity as the valid ones to be chosen as the location of the plotting. That makes the "physical-opname" process before training the model important to filter those types of invalid rights. The overlapping rights also affect the model performance by increasing the number of unmatched parcels as mentioned in the previous part. An accurate land parcel attribute will help to overcome the condition.

The last factor that needs to be considered is the attribute of land parcels. As mentioned earlier about the importance of attribute matching to the model, if the textual attribute that the parcels had doesn't reflect the actual data it will also create a wrong location prediction. It also includes the typos that may happen when adding the textual attribute that will affect the model's accuracy.

Besides the technical performance that is related to the accuracy of the model in predicting the location for plotting the land parcel, several conditions affect the model calculation time which might be as important as the model's accuracy. The thing that must be kept in mind is the calculation time is not only related to the size of the data, there are many other things such as the computational ability of the computer. The bigger dataset if being computed with a higher performance computer will finish faster compared with the smaller one with the lower computation ability. This may be one of several gaps in implementing this to the Indonesia Land Administration System which will be presented in detail in the next chapter.

5.3.3. The gaps in implementation

The future of machine learning in the Indonesian Land Administration shows a bright path ahead. From the interview with the respondents, at least there were already two implementations of machine learning for the Indonesian Land Administration: To extract building boundaries automatically and to detect possible gaps and overlaps in the cadaster database. Despite the implementation being still in the pilot project stage, the large data amount and the high target to quickly achieve a complete land registration brought the use of every possible way to accelerate the process were open including the use of machine learning. To implement this innovation in the current system, several gaps need to be covered that could be divided into administration gaps and technical gaps.

There are at least two administration gaps highlighted by the respondents in the interview. The first one is the unavailability of regulations that allow the use of machine learning for land administration. The idea was mentioned by one respondent from the head office *"There should be a regulation either in the technical guidance or in a higher law to regulate in which area condition this machine learning can be applied."* It was also mentioned in the interview by several respondents that it can be implemented in an area that has at least 80% of land registration coverage. It is to ensure that there were available locations to be used in the testing and can enhance the prediction's accuracy.

The next concern related to the administration was the non-uniform structure of the cadastral data. This idea was related to the big data involved in the machine learning implementation which may have created a misperception when the data itself was not yet structured. One of the respondents said, *"There is an importance to make a standard of the data that can be involved in the machine learning analysis to make it more usable and efficient."*

On the technical side, two gaps were mentioned by the respondents and came from the observations during the fieldwork process. The first one is the concern of prediction accuracy given by machine learning. One of the respondents who is experienced in leading the surveyor team mentioned that *"There was a doubt about the results, which in my opinion still needed to be verified by an experienced surveyor."* This notion was understandable concerning the status of the certificate in the Indonesian Land Administration which was not only roled as a technical product but also a legal product that may generate a legal problem in the future time if not well managed. The other interview in a local land office also added the requirement to add some field checking to the result to convince the administrator when issuing the product and minimize the imminent risk.

The next gap was the computational ability of the current computer that is available in the office. The machine learning model involved an exhaustive search of the location for plotting a land parcel from out of maybe a hundred thousand or more available locations on the database. The calculation process required a powerful system to prevent the computational complexity error. During the process of research, a virtual computer that provides high computational capacity was used. This solution might be useful for further implementation in Indonesia to hire a high-speed virtual computer that is accessible everywhere and at any time rather than using a personal computer with limited capability.

Despite there being technical and administration gaps in the implementation of machine learning in the current land administration process, optimism has appeared during the interview process regarding the success of implementation. One of the respondents mentioned that *"The machine learning innovation could be a solution to immediately plotted the K4 Parcels (The Unplotted Land Parcel) which already reached 15 million in total."* The other also delivers an optimism in the interview process *"The possibility of using machine learning in our land administration is big as long as the terms and conditions regarding the data condition and accuracy were applied."*

5.3.4. Contribution of the Research in the Land Administration

The research on utilizing the geospatial matching of land parcels based on machine learning had an impact on both land administration in general and on the Indonesian Land Administration in particular. This chapter will review several contributions to both fields.

The first contribution is to encourage the application of machine learning, especially in the land administration sector. The use of geospatial matching based on machine learning to find the possible location for plotting the land parcels is new. The previous implementation of the polygon's matching technique was done to enhance the cadaster base map in the city of Tehran, Iran (Hajiheidari et al., 2024). This research used only the geometric aspect of parcels and recommended to use of descriptive information about the land parcels. Another related study focused on matching between two different databases of building footprints in Spain (Ruiz-Lendínez et al., 2017). This research recommended performing a one-to-many matching technique, to improve a one-to-one strategy initially chosen. Building on these insights, this research gives a wider application of the matching algorithm in land administration and incorporates the suggested improvement. The model, built based on the experience of the current manual plotting process, gives a distinction by including the information gathered from the interview into a machine learning model.

Additionally, The machine learning model to find the location for plotting the unplotted land parcel can also be adapted in other countries that face the same problem of plotting the land parcel with limited spatial information. Figure 27 illustrates the proposed workflow for implementing location searching using geospatial data matching. As seen in the figure, the model needs at least one spatial data to find the location in the database. If there is no spatial information available, it is suggested to do the parcel remeasurement process. It also involved a ground checking of a group of unplotted land parcels as indicated by the interview respondents, ensuring the accuracy of the prediction.

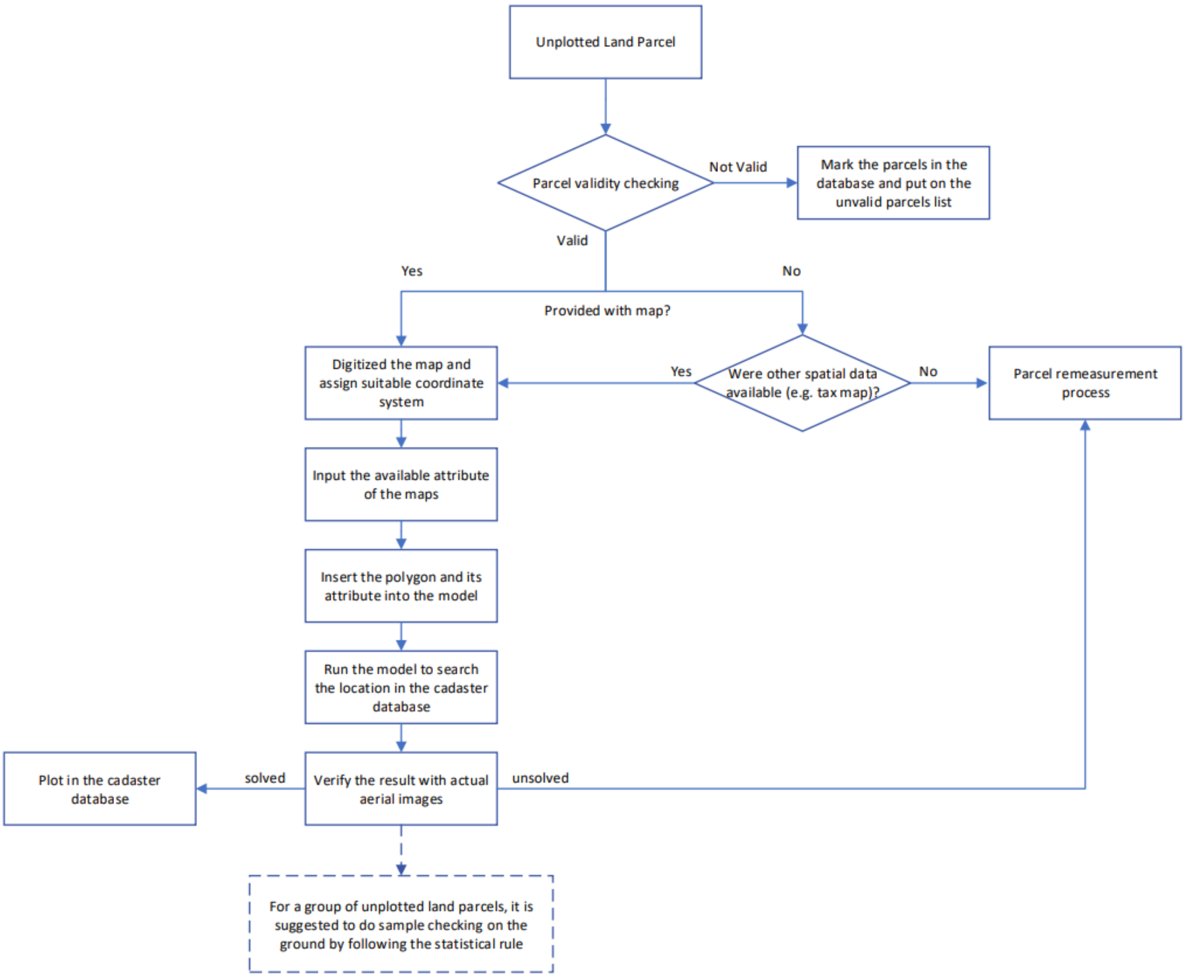

Unplotted Land Parcel
Parcel validity checking
Not Valid
Mark the parcels in the database and put on the unvalid parcels list
Valid
Yes
No
Provided with map?
Digitized the map and assign suitable coordinate system
Yes
Were other spatial data available (e.g. tax map)?
No
Parcel remeasurement process
Input the available attribute of the maps
Insert the polygon and its attribute into the model
Run the model to search the location in the cadaster database
Plot in the cadaster database
solved
Verify the result with actual aerial images
unsolved
For a group of unplotted land parcels, it is suggested to do sample checking on the ground by following the statistical rule

Figure 27. Proposed Workflow for Implementation

For the Indonesian Land Administration, the pre-trained model learning hopefully can contribute to accelerating the process of cadastral quality improvement process to plot the KW4 Land Parcels (Unplotted Land Parcels) which is estimated at 15 million parcels. The implementation of this model could be tested in several pilot-project cities that had the similarity with Surabaya Condition or could also done in the first 15 "Kota Lengkap" (cities that completed their land registration). The capability of the model to find a location with a similar shape can be applied also in other cities in Indonesia in the process of Cadastral Remeasurement. Cadastral Remeasurement was a common process in the Land Office that is applied by the landowners to find the location of their certificate, the problem is common in Indonesia, often occurring because landowners purchase certificates without maps or no longer reside at the certificate locations and forget about the boundary and location of their certificate. By using this model, certificates could be checked before the remeasurement activity occurs, preventing overlapping with other available rights and potentially sparing surveyors from court disputes.

5.4. Chapter Summary

This chapter presents an analysis and discussion of the result based on the interview and additional information gathered from the fieldwork. The analysis was divided into three phases, learning from the past, analyzing present data, and strategy to implement the model in the future. The first phase was reviewing the lessons learned from the previous data related to the regulation and the effect on the current cadaster data. The second phase analyzed the effect of the current cadastral data on the matching result of the model and how to enhance the model performance using geospatial matching. The last phase talked more about the possibility of the implementation and the factors hindering the application of machine learning in the Indonesian Land Administration. This chapter completed the previous on answering the research questions, especially in sub-objective 3 which focused on enhancing model performance and understanding the factors related to the model performance.

6. CONCLUSION AND RECOMMENDATIONS

6.1. Conclusion

The machine learning model based on geospatial data matching was built based on two aspects of geospatial data matching: geometric matching and attribute matching. The combination of these techniques has proven to generate good accuracy in finding the location of the unplotted land parcels. It also provides a broader application of the related techniques to the land registration process. This part will conclude the results and discussions based on the sub-objectives and questions they satisfied.

6.1.1. Sub-Objective 1: To identify the causes of the land parcel becoming unplotted and identify the heuristic process to plot that type of land parcel.

Q1: What are the causes of the land parcel becoming unplotted?
Eight causes of the unplotted land parcels in Indonesia were identified from the open-ended interview. The two most frequently mentioned causes were issues tied to local control points/geographic features and insufficient information on available documents. For the first main cause, one triggering factor was allowing the use of local coordinates for mapping in the past due to some limitations. However, transforming these coordinates from the local to the national system was not done in the future. The second cause was mainly related to missing the attributes on the title plan (Gambar Ukur) such as coordinates grid or measured value from nearest tied points, that could help identify the location of land parcels

Q2: What heuristic process has been used to plot the land parcel manually to a current cadastral database?
There are two processes, in general, to plot the land parcel manually: the studio process and the field verification. The studio process involves a series of works done in the office which only relies on the available documents with some help of supporting documents such as tax maps. If the studio process was not able to plot the land parcels, the field verification was done by directly asking the location of parcels to the local leader or the owner of the parcel itself and using the aerial image to plot the land parcels.

6.1.2. Sub-Objective 2: To adopt a machine learning model in identifying the matching between two geospatial data to automatically find the best-fitting location for the unplotted land parcel.

Q3: What are the matching components from the heuristic process that can be used to identify the matching between an unplotted land parcel and the available locations in the parcel database?
There are five components used to identify the geometric matching between the unplotted land parcels and the available location. Those five components are polygon's area, polygon's perimeter, number of vertices, minimum bounding rectangle, and arkin graph angle. To enhance the model performance an attribute matching was added: neighboring parcel's NIB, village name, and the registered area information. The combination of geometric matching and textual attribute matching creates a model to find the best location to plot the land parcels based on the geospatial data matching.

Q4: How to optimize the components to correctly identify matching between an unplotted land parcel and the available locations in the parcel database?
To optimize the geometric component, an optimization algorithm called the RCGA was used to find the optimum weight of each geometric component. The weight will then be used in the exhaustive calculation to find the location of the land parcel based on its overall matching score. The top three locations with the highest matching scores were filtered using attribute matching to find the best location for plotting the land parcels in a cadastral database.

6.1.3. Sub-Objective 3: Evaluate the model's performance in identifying the matching between two geospatial data to automatically find the best-fitting location for the unplotted land parcel.

Q5: How many correct matches does the model get for finding the best-fitting location for the unplotted land parcel?

The model was tested by using two databases: The Sample Data and The KKP Database. For the first test, the model had the best result of 158 matched locations, 2 wrong-matched results, and 21 results that the model could not identify as matching (unmatched results). From those statistics, the precision value of the first test was 98.75% and the recall value was 88.27%. In the second test using the more complex database, the precision and recall value decreased to 91% precision value and 57% recall value due to more parcels that could not be identified for some reason such as homogenous candidate locations and overlapping right on the database. To enhance the model performance, an attribute matching was done to the result from the KKP Database which resulting the best result of 48 correct matches, 4 incorrect matches, and 5 unmatched results. The results improve the precision value to 92% and the recall value to 91%.

Q6: How does each component influence the model's performance of finding the best-fitting location for the unplotted land parcel?

The geospatial matching component was categorized into two: geometric and textual attributes. For the geometric matching, an experiment was conducted to understand the effect of every matching variable on the model's result. It was concluded that the combination of five variables (A, n, P, MBR, and AGA) was the best choice for identifying matches in a database with complex conditions. Despite having longer calculation times, using all complete variables could eliminate the possible mistake due to scale differences, resulting in higher performance of the model. This approach also allowed for the identification of parcels with complex shapes which resulted in a better performance on the model.

In the aspect of textual matching, three attributes from land parcels were tested on the geometric matching results. It was concluded that the use of unique identifiers of land parcels such as registered area could eliminate unmatched results caused by overlapping rights in the database. Additionally, the neighboring NIB attribute was effective in eliminating the possible mistake due to the similar-shaped candidate locations. The last attribute that did not affect much the model's performance was the village name, which could filter the parcels with similar matching scores located in different villages.

Q7: What are the factors that contribute to the model performance of finding the best-fitting location for the unplotted land parcel?

The factors that contribute to model performances can be divided into internal and external factors. The internal factors include the condition of digitized unplotted parcels and the conditions of the candidate locations. An experiment was conducted on the model to find the location in a database that had no matched candidate locations, which resulted in wrong location predictions. The next affecting factor was an external factor related to the database condition and the accuracy of the land parcels attribute. The model was built using the cadastral data from a large city in Indonesia with a complete combination of land use, making its performance in other city conditions unknown. The model also depends on the attributes of the land parcel which are manually inputted into the database and may contain errors.

In addition to these two factors related to the model accuracy, there is another factor that may affect the calculation time of the model: the computational ability of the computer. It is important to consider this as an affecting factor to prevent computational error due to limited computational capacity.

6.2. Limitations of The Research

The research faced some limitations, specifically in the scope of work and the data handling capabilities. Firstly, it only calculates the similarity between the unplotted land parcel and the available locations stored in vector format. This required a manual intervention to convert the available analog data of land parcels into the digitized vector format, which can be time-consuming and error-prone. Additionally, the attribute matching done in this research relies solely on the data inputted by the officers into the KKP Database. This process is also prone to human error while inputting the information, and the model does not delve into the automation of inserting the attribute into the database.

The further limitations related to the scope of the research work which limited to data from Surabaya City due to the limited time. They resulted in the model being built only applicable in the city with similar data conditions and leaving the applicability of the result in other cities in Indonesia uncertain. These constraints highlight the need for enhancements in the automation of the data processing and the wider implementation of the model to improve the model's usability and generalizability.

6.3. Recommendations

The recommendations will be divided into two parts: the recommendation for future research and the recommendation for the institution. Firstly, it is recommended to continue this research to analyze the use of the current model in other cities to evaluate its performance and accuracy across different cadaster data conditions and city classifications. Secondly, it is recommended to integrate the model with the available models to automatically translate the analog version of spatial and textual data into the digital format to minimize possible errors during the process. Additionally, it is recommended to continue the research by implementing the proposed workflow in Figure 27 and review the improvement in process time and the accuracy of search results compared with the manual process.

The next recommendation for the institution is to emphasize the importance of having position information in the official maps (Gambar Ukur or Surat Ukur) in the survey and mapping official guidelines. This practice is important to be used as the backup if the digital data is lost. Additionally, it is important to include the geographical features' names in the Cadaster database. This innovation accommodates the available information in the Gambar Ukur or Surat Ukur that often ties the flying parcel to geographical features, thereby simplifying locating these unplotted land parcels. Lastly, it is recommended to continue programs like the IP4T and the Declaration of Kota Lengkap, which gave indicative parcel boundaries to establish a reliable base for plotting the land parcels and detecting the land parcels plotted in the wrong location. Combining those two programs with the current PTSL program will increase the number of registered land parcels in Indonesia without leaving the land parcels from the past unresolved.

LIST OF REFERENCES

Adam, A. G., Cikara, A. M., Kayuza, H., Wabineno, L. M., Potel, J., Wayumba, R. N., Turimubumwe, P., & Zevenbergen, J. A. (2019). *Land Governance Arrangements in Eastern Africa: Description and Comparison*. 1–18. https://www.uneca.org/clpa2019

Aditya, T., Santosa, P. B., Yulaikhah, Y., Widjajanti, N., Atunggal, D., & Sulistyawati, M. (2021). Title Validation and collaborative mapping to accelerate quality assurance of land registration. *Land Use Policy*, *109*, 105689. https://doi.org/10.1016/j.landusepol.2021.105689

Aditya, T., Sucaya, I. K. G. A., & Nugroho Adi, F. (2021). LADM-compliant field data collector for cadastral surveyors. *Land Use Policy*, *104*, 105356. https://doi.org/10.1016/J.LANDUSEPOL.2021.105356

Agegnehu, S. K., Dires, T., Nega, W., & Mansberger, R. (2021). Land tenure disputes and resolution mechanisms: Evidence from peri-urban and nearby rural kebeles of debre markos town, ethiopia. *Land*, *10*(10). https://doi.org/10.3390/land10101071

Alem, A., & Kumar, S. (2020). Deep Learning Methods for Land Cover and Land Use Classification in Remote Sensing: A Review. *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, 903–908. https://doi.org/10.1109/ICRITO48877.2020.9197824

Alston, L. J., Libecap, G. D., & Mueller, B. (2000). Land Reform Policies, the Sources of Violent Conflict, and Implications for Deforestation in the Brazilian Amazon. *Journal of Environmental Economics and Management*, *39*, 162–188. https://doi.org/10.1006/jeem.1999.1103

Arkin, E., Chew, L., Huttenlocher, D., Kedem, K., & Mitchell, J. (1991). An Efficiently Computable Metric for Comparing Polygonal Shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, *13*, 209–216. https://doi.org/10.1145/320176.320190

Arkin, E. M., Chew, L. P., Huttenlocher, D. P., & Kedem, K. (1991). An Efficiently Computable Metric for Comparing Polygonal Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(3), 209–216. https://doi.org/10.1109/34.75509

Ary Sucaya, I. K. G. (2009). *Application and validation the Land Administration Domain Model in a real life situation (A case study in Indonesia)*. University of Twente.

Avram, D., Bratosin, I., Ilie, D., Coordinator, S., & Eng Mariana CALIN, L. (2016). Surveying Theodolite Between Past and Future. *Journal of Young Scientist*, *IV*.

Badan Pusat Statistik Kota Surabaya. (2023). *Kota Surabaya dalam angka 2023*.

Baur, K., Rosenfelder, M., & Lutz, B. (2023). Automated real estate valuation with machine learning models using property descriptions. *Expert Systems with Applications*, *213*, 119147. https://doi.org/10.1016/J.ESWA.2022.119147

Brooks, H., Bee, P., & Rogers, A. (2019). Introduction to qualitative data analysis. *A Research Handbook for Patient and Public Involvement Researchers*. https://doi.org/10.7765/9781526136527.00013

Caldwell, D. R. (2005, January 22). *Unlocking the Mysteries of the Bounding Box*. https://archive.ph/20120721210709/http://purl.oclc.org/coordinates/a2.htm

Chaturvedi, V., & de Vries, W. T. (2021). Machine Learning Algorithms for Urban Land Use Planning: A Review. *Urban Science*, *5*(3). https://doi.org/10.3390/URBANSCI5030068

Chipofya, M., Jan, S., Schultz, C., & Schwering, A. (2017). Towards Smart Sketch Maps for Community-driven Land Tenure Recording Activities. *The 20th AGILE International Conference on Geographic Information Science*.

Claesen, M., & De Moor, B. (2015). Hyperparameter Search in Machine Learning. *MIC 2015: The XI Metaheuristics International Conference*. https://www.codalab.org/competitions/2321.

Crommelinck, S., Koeva, M., Yang, M. Y., & Vosselman, G. (2019). Application of Deep Learning for Delineation of Visible Cadastral Boundaries from Remote Sensing Imagery. *Remote Sensing 2019, Vol. 11, Page 2505*, *11*(21), 2505. https://doi.org/10.3390/RS11212505

DiCicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical Education*, *40*(4), 314–321. https://doi.org/10.1111/J.1365-2929.2006.02418.X

Ding, X., Zheng, M., & Zheng, X. (2021). The Application of Genetic Algorithm in Land Use Optimization Research: A Review. *Land 2021, Vol. 10, Page 526*, *10*(5), 526. https://doi.org/10.3390/LAND10050526

Duncan, E. E., & Rahman, A. A. (2013). A multipurpose cadastral framework for developing countries-concepts. *Electronic Journal of Information Systems in Developing Countries*, *58*(1), 1–16. https://doi.org/10.1002/J.1681-4835.2013.TB00411.X

Enemark, S. (2005). Understanding the land management paradigm. *FIG COM 7 Symposium On Innovate Technologies for Land Administration*. https://www.researchgate.net/publication/228342504

Enemark, S., McLaren, R., & Lemmen, C. (2021). Fit-for-purpose land administration—providing secure land rights at scale. *Land*, *10*(9). https://doi.org/10.3390/LAND10090972

Fan, H., Zipf, A., Fu, Q., & Neis, P. (2014). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, *28*(4), 700–719. https://doi.org/10.1080/13658816.2013.867495

FAO, UNECE, & FIG. (2022). Digital transformation and land administration - Sustainable practices from the UNECE region and beyond. In *Digital transformation and land administration*. FAO; UNECE (United Nations Economic Commission for Europe); https://doi.org/10.4060/cc1908en

Femenia-Ribera, C., Mora-Navarro, G., & Santos Pérez, L. J. (2022). Evaluating the use of old cadastral maps. *Land Use Policy*, *114*, 105984. https://doi.org/10.1016/j.landusepol.2022.105984

FIG. (1995). *FIG Statement on the Cadastre*. Publication 11. https://www.fig.net/resources/publications/figpub/pub11/figpub11.asp#4

Franken, J., & Florijn, W. (2021). Rebuilding the Cadastral Map of The Netherlands: The Artificial Intelligence Solution. *FIG*.

Gafurov, A. (2023). Automated Mapping of Cropland Boundaries Using Deep Neural Networks. *AgriEngineering 2023, Vol. 5, Pages 1568-1580*, *5*(3), 1568–1580. https://doi.org/10.3390/AGRIENGINEERING5030097

Government Regulation 10/1961 (1961). https://peraturan.bpk.go.id/Details/72692/pp-no-10-tahun-1961

Grant, D. B., Mccamley, G., Mitchell, D., Enemark, S., & Zevenbergen, J. (2018). *Upgrading Spatial Cadastres in Australia and New Zealand: Functions, Benefits & Optimal Spatial Uncertainty*.

Grant, D., Enemark, S., Zevenbergen, J., Mitchell, D., & McCamley, G. (2020). The Cadastral triangular model. *Land Use Policy*, *97*, 104758. https://doi.org/10.1016/J.LANDUSEPOL.2020.104758

Hajiheidari, A., Delavar, M. R., & Rajabifard, A. (2024). Smart Urban Cadastral Map Enrichment—A Machine Learning Method. *ISPRS International Journal of Geo-Information 2024, Vol. 13, Page 80*, *13*(3), 80. https://doi.org/10.3390/IJGI13030080

Hamarat, C., & Kilic, K. (2010). A genetic algorithm based feature weighting methodology. *40th International Conference on Computers and Industrial Engineering: Soft Computing Techniques for Advanced Manufacturing and Service Systems, CIE40 2010*. https://doi.org/10.1109/ICCIE.2010.5668297

Handono, A. B., Suhattanto, Muh. A., & Nugroho, A. (2020). Strategi Percepatan Peningkatan Kualitas Data Pertanahan di Kantor Pertanahan Kabupaten Karanganyar. *Jurnal Tunas Agraria*, *3*(3).

Hashim, N. M., Omar, A. H., Ramli, S. N. M., Omar, K. M., & Din, N. (2013). Cadastral database positional accuracy improvement. *International Archives of the Photogrammetry, Remote Sensing and Spatial*

Information Sciences - ISPRS Archives, *42*(4W5), 91–96. https://doi.org/10.5194/isprs-archives-XLII-4-W5-91-2017

Hemanth Sai Kumar, V. (2023). *Selection of Real-Coded Genetic Algorithm parameters in solving simulation-optimization problems for the design of water distribution networks*. https://doi.org/10.2166/ws.2023.301

Herrera, F., Lozano, M., & Verdegay, J. L. (1998). Tackling Real-Coded Genetic Algorithms: Operators and Tools for Behavioural Analysis. *Artificial Intelligence Review*, *12*, 265.

Holland, J. H. (1992). Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. *Adaptation in Natural and Artificial Systems*. https://doi.org/10.7551/MITPRESS/1090.001.0001

Holsen, J., & Lsen, H. O. (1984). The Development of Survey Instruments. *International Hydrographic Review*, *LXI*(1).

Hutabarat, S. M. S. (2011). *Land Dispute Resolution Mechanisms in the Perspective of Good Governance: the Case study in Indonesia*. University of Twente.

Jain, A., Arora, A., Morato, J., Yadav, D., & Kumar, K. V. (2022). Automatic Text Summarization for Hindi Using Real Coded Genetic Algorithm. *Applied Sciences (Switzerland)*, *12*(13). https://doi.org/10.3390/APP12136584

Jong, M., Guan, K., Wang, S., Huang, Y., & Peng, B. (2022). Improving field boundary delineation in ResUNets via adversarial deep learning. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102877. https://doi.org/10.1016/J.JAG.2022.102877

Joo, Y. J., & Kim, D. H. (2014). The big data analytics regarding the cadastral resurvey news articles. *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, *32*(6), 651–659. https://doi.org/10.7848/KSGPC.2014.32.6.651

Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, *80*(5), 8091. https://doi.org/10.1007/S11042-020-10139-6

Khomsin, K., Anjasmara, I., Pratomo, D., & Ristanto, W. (2019). Accuracy Analysis of GNSS (GPS, GLONASS and BEIDOU) Obsevation For Positioning. *E3S Web of Conferences*, *94*, 01019. https://doi.org/10.1051/e3sconf/20199401019

Klebanov, M., & Doytsher, Y. (2009). *Cadastral Triangulation: A Block Adjustment Approach for Joining Numerous Cadastral Blocks*.

Li, D., Luo, L., Zhang, W., Liu, F., & Luo, F. (2016). A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. *BMC Bioinformatics*, *17*(1), 1–11. https://doi.org/10.1186/S12859-016-1206-3/TABLES/9

Li, X., Anthony, &, Yeh, G.-O., Li{, X., & Gar-On Yeh{, A. (2005). Integration of genetic algorithms and GIS for optimal location search. *International Journal of Geographical Information Science*, *19*(5), 581–601. https://doi.org/10.1080/13658810500032388

Martono, D. B., Aditya, T., Subaryono, S., & Nugroho, P. (2022). Cadastre Typology as a Baseline for Incremental Improvement of Spatial Cadastre in Jakarta: Towards a Complete Cadastre. *Land 2022, Vol. 11, Page 1732*, *11*(10), 1732. https://doi.org/10.3390/LAND11101732

Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2022). Machine Learning Applications to Land and Structure Valuation. *Journal of Risk and Financial Management 2022, Vol. 15, Page 193*, *15*(5), 193. https://doi.org/10.3390/JRFM15050193

Ministry of ATR/BPN. (2023). *Rapat Kerja Nasional*.

Mirshekarian, S., & Sormaz, D. (2018). Machine Learning Approaches to Learning Heuristics for Combinatorial Optimization Problems. *Procedia Manufacturing*, *17*, 102–109. https://doi.org/10.1016/J.PROMFG.2018.10.019

Mortara, M., & Spagnuolo, M. (2001). Similarity measures for blending polygonal shapes. *Computers & Graphics*, *25*(1), 13–27. https://doi.org/10.1016/S0097-8493(00)00104-7

Murdin, P. (2009). *Full meridian of glory : perilous adventures in the competition to measure the earth*. Copernicus Books/Springer.

Nichols, J. A., Herbert Chan, H. W., & Baker, M. A. B. (2019). Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews*, *11*(1), 111. https://doi.org/10.1007/S12551-018-0449-9

Noble, R. D. (1982). MATHEMATICAL MODELLING IN THE CONTEXT OF PROBLEM SOLVING. *Mathematical Modelling*, *3*, 215–219.

Rajabifard, A., Williamson, I., Steudler, D., Binns, A., & King, M. (2007). Assessing the worldwide comparison of cadastral systems. *Land Use Policy*, *24*(1), 275–288. https://doi.org/10.1016/J.LANDUSEPOL.2005.11.005

Rodriguez-Maya, N., Flores, J. J., & Graff, M. (2016). Predicting the rcga performance for the university course timetabling problem. *Communications in Computer and Information Science*, *597*, 31–45. https://doi.org/10.1007/978-3-319-30447-2_3

Ruiz-Lendínez, J. J., Ureña-Cámara, M. A., & Ariza-López, F. J. (2017). *Geo-Information A Polygon and Point-Based Approach to Matching Geospatial Features*. https://doi.org/10.3390/ijgi6120399

Sabekti, W. S. (2010). *A Conversion Strategy to Improve the Quality of Cadastral Map and to Support the Registration Process: Indonesian Case*.

Simpson, A. R., & Priest, S. D. (1993). The application of genetic algorithms to optimisation problems in geotechnics. *Computers and Geotechnics*, *15*(1), 1–19. https://doi.org/10.1016/0266-352X(93)90014-X

Smit, B. (2002). Atlas.ti for qualitative data analysis. *Perspectives in Education*, *20*(3).

Steed, R., & Williams, B. (2020). *Heuristic-Based Weak Learning for Automated Decision-Making*. https://github.com/

Sun, S., Cao, Z., Zhu, H., & Zhao, J. (2019). *A Survey of Optimization Methods from a Machine Learning Perspective*.

FIG. (1996). *THE BOGOR DECLARATION*. https://fig.net/organisation/comm/7/library/reports/events/sing97/sing974.htm

The Head of National Land Agency Regulation 16/2010 (2010). https://landregulations.files.wordpress.com/2017/02/th-2010_perkbpn_-16-tahun-2010-tentang-perubahan-atas-perkbpn-ri-no-9-tahun-2009-tentang-pemekaran-kantah-kota-surabaya.pdf

The Minister of Agrarian/Head of National Land Agency Regulations 2/1996 (1996). https://www.ndaru.net/wp-content/peraturan/pmna/pmna_1996_02.pdf

Tim Peneliti Teknik Geodesi UGM. (2022). *Laporan Akhir Kegiatan Peningkatan Kualitas Bidang Tanah Terdaftar (Pbt K4 Non Sistematis) Menuju Kota/Kabupaten Lengkap di Kota Surabaya, Provinsi Jawa Timur*.

UNECE. (1996). *Land administration guidelines : with special reference to countries in transition*. United Nations.

UNECE. (2005). *LAND ADMINISTRATION IN THE UNECE REGION Development trends and main principles*. http://www.unece.org

Varpa, K., Iltanen, K., & Juhola, M. (2014). Genetic Algorithm Based Approach in Attribute Weighting for a Medical Data Set. *Journal of Computational Medicine*, *2014*, 1–11. https://doi.org/10.1155/2014/526801

Wouters, R., Meijerink, G., Vaandrager, R., & Zavrel, J. (2010). Extracting Information from Deeds by Optical Character Recognition (OCR) and Text Interpretation. *FIG Congress*.

Wudye Tareke, B. (2022). *Visible Cadastral Boundary Extraction Using VHR Remote Sensing Images: A Deep Learning Approach*. University of Twente.

Xavier, E. M. A., Ariza-L, F. J., Opez, ´, Ure˜na, M. A., Ariza-López, F. J., & Ureña, M. A. (2016). A Survey of Measures and Methods for Matching Geospatial Vector Datasets. *ACM Comput. Surv*, *49*. https://doi.org/10.1145/2963147

Yang, L., & Shami, A. (2022). *On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice*. https://github.com/

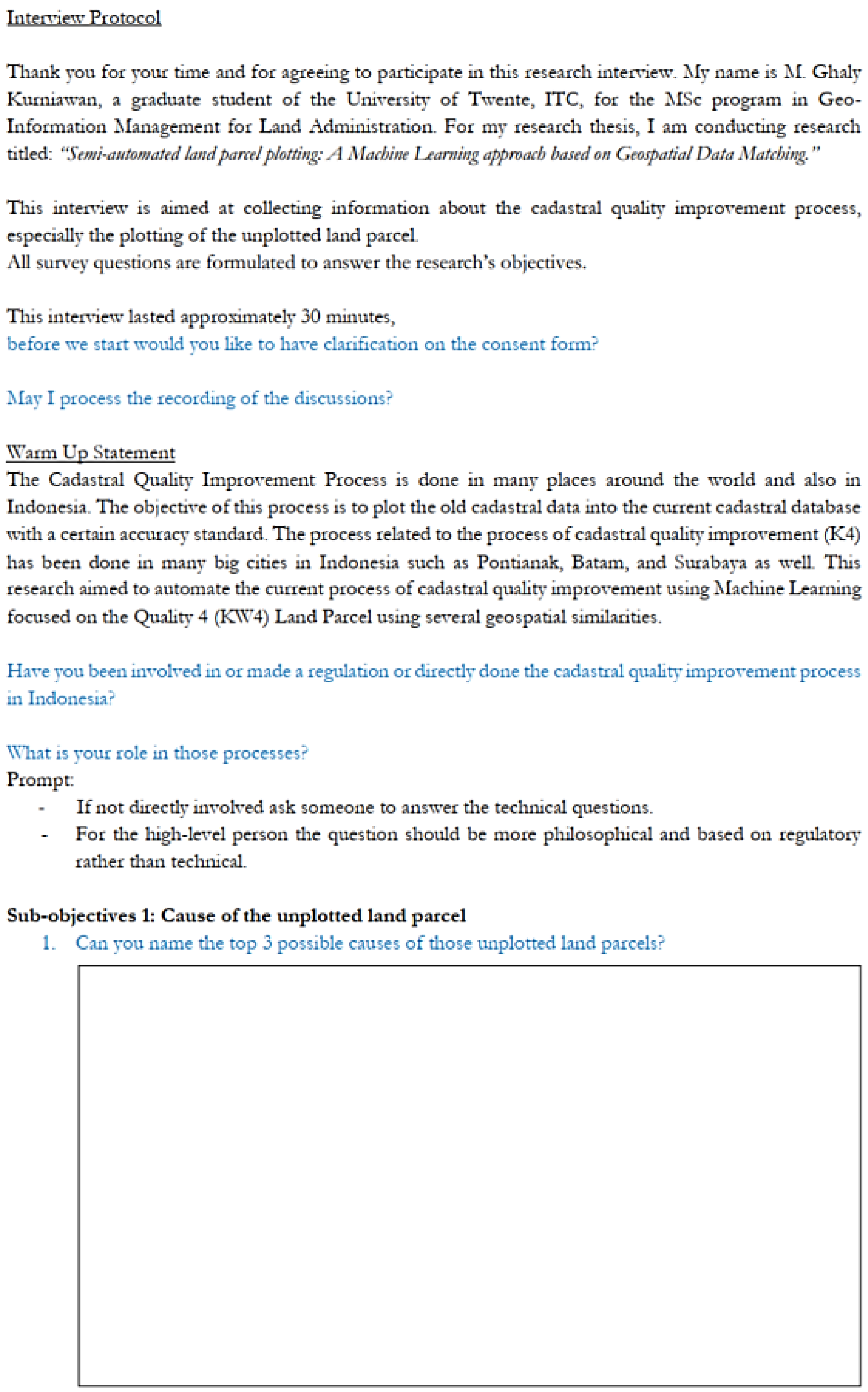
Yildiz, O., & Erden, Ç. (2020). Cadastral updating: the case of Turkey. *Survey Review*, *53*(379), 335–348. https://doi.org/10.1080/00396265.2020.1759982

Yuh, Y. G., Tracz, W., Damon Matthews, H., & Turner, S. E. (2023). Application of machine learning approaches for land cover monitoring in northern Cameroon. *Ecological Informatics*, *74*, 101955. https://doi.org/10.1016/j.ecoinf.2022.101955

Zhang, H., Liu, M., Wang, Y., Shang, J., Liu, X., Li, B., Song, A., & Li, Q. (2021). Automated delineation of agricultural field boundaries from Sentinel-2 images using recurrent residual U-Net. *International Journal of Applied Earth Observation and Geoinformation*, *105*, 102557. https://doi.org/10.1016/J.JAG.2021.102557

APPENDICES

APPENDIX 1: OPEN-ENDED INTERVIEW QUESTION LISTS



Interview Protocol

Thank you for your time and for agreeing to participate in this research interview. My name is M. Ghaly Kurniawan, a graduate student of the University of Twente, ITC, for the MSc program in Geo-Information Management for Land Administration. For my research thesis, I am conducting research titled: *"Semi-automated land parcel plotting: A Machine Learning approach based on Geospatial Data Matching."*

This interview is aimed at collecting information about the cadastral quality improvement process, especially the plotting of the unplotted land parcel.
All survey questions are formulated to answer the research's objectives.

This interview lasted approximately 30 minutes,
before we start would you like to have clarification on the consent form?

May I process the recording of the discussions?

Warm Up Statement
The Cadastral Quality Improvement Process is done in many places around the world and also in Indonesia. The objective of this process is to plot the old cadastral data into the current cadastral database with a certain accuracy standard. The process related to the process of cadastral quality improvement (K4) has been done in many big cities in Indonesia such as Pontianak, Batam, and Surabaya as well. This research aimed to automate the current process of cadastral quality improvement using Machine Learning focused on the Quality 4 (KW4) Land Parcel using several geospatial similarities.

Have you been involved in or made a regulation or directly done the cadastral quality improvement process in Indonesia?

What is your role in those processes?
Prompt:

- If not directly involved ask someone to answer the technical questions.
- For the high-level person the question should be more philosophical and based on regulatory rather than technical.

Sub-objectives 1: Cause of the unplotted land parcel

1. Can you name the top 3 possible causes of those unplotted land parcels?

UNIVERSITY OF TWENTE.

Prompt :

- Is it's related to the land parcel that plotted in local coordinates?
- Or is it tied to the natural features that have already been changed?

2. If the causes are related to the mapped local coordinates or tied to natural features, how often it is happening in one place?

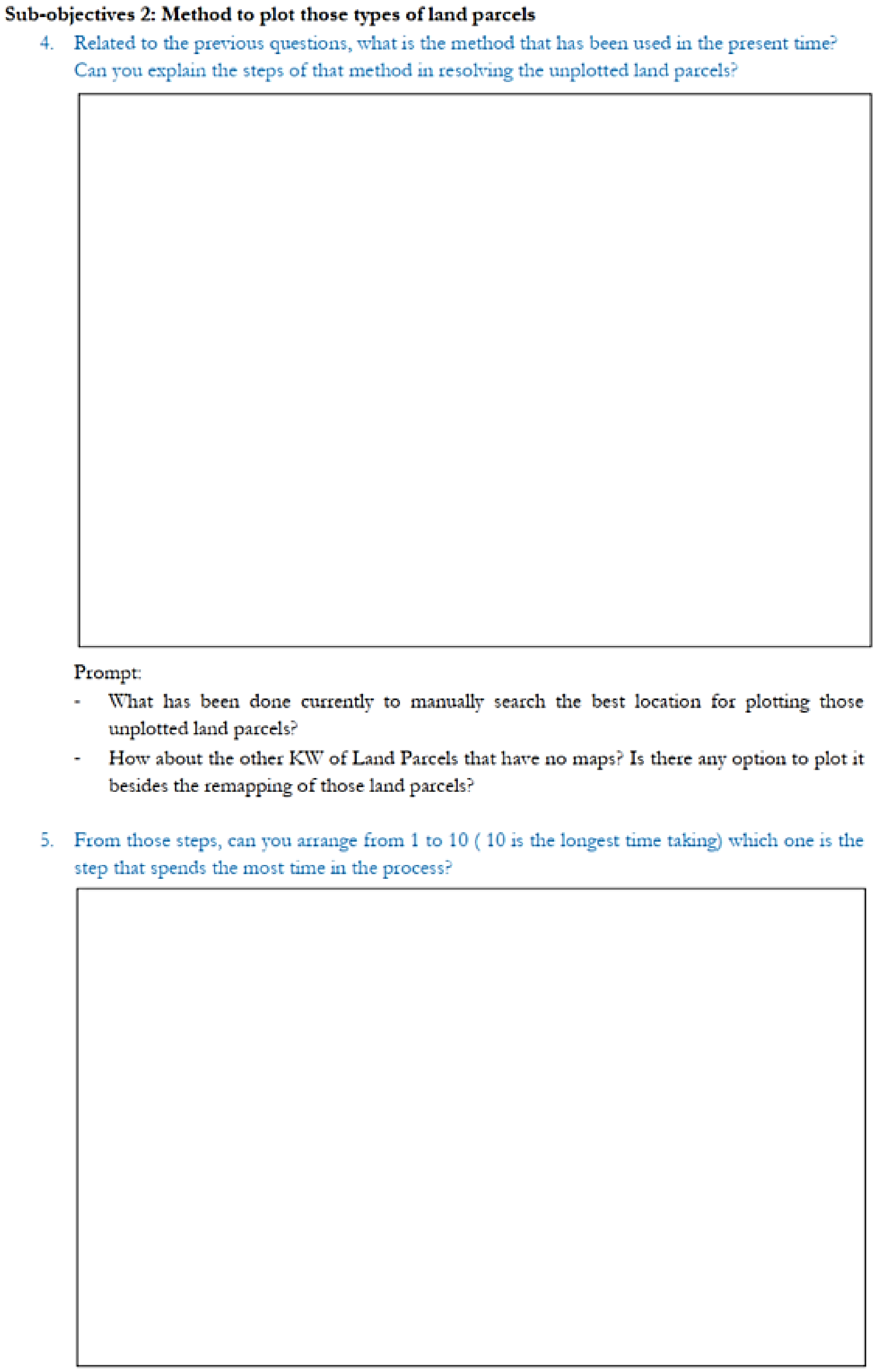
Prompt:

- Related to KW4 Land Parcels
- Related to the type of boundary based on the features tied to it (e.g.: river, mountain, hill). Is there any special treatment based on the type of natural features or not?

3. How that caused difficulties in mapping in the current cadaster database?

Prompt:

- Dig down to the cause related to KW4 parcels.



Sub-objectives 2: Method to plot those types of land parcels

4. Related to the previous questions, what is the method that has been used in the present time? Can you explain the steps of that method in resolving the unplotted land parcels?

Prompt:

- What has been done currently to manually search the best location for plotting those unplotted land parcels?
- How about the other KW of Land Parcels that have no maps? Is there any option to plot it besides the remapping of those land parcels?

5. From those steps, can you arrange from 1 to 10 (10 is the longest time taking) which one is the step that spends the most time in the process?

6. Based on your experience, which steps of that method if done automatically will accelerate the whole process of cadastral quality improvement? And why?

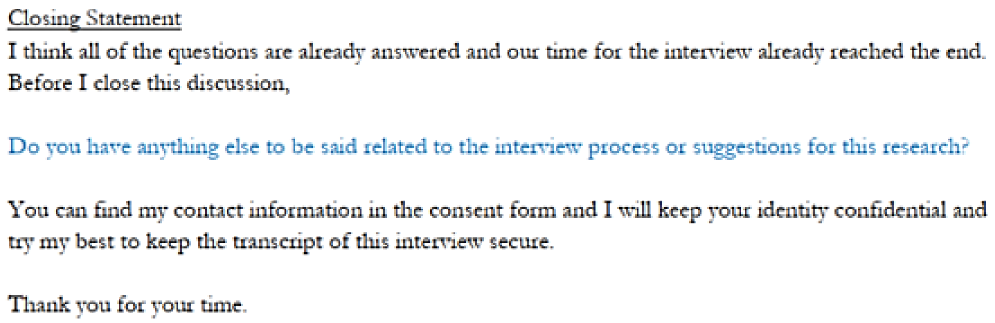
Prompt:

- It is related to plotting the unplotted land parcel which is mapped in local coordinates or tied to "moving" natural features?

7. From your experience, do you think machine learning will work to accelerate the current process?

Prompt:

- Is there any special case that the "human touch" is still needed in this process?
- Is it feasible to be implemented in the organization?



Closing Statement

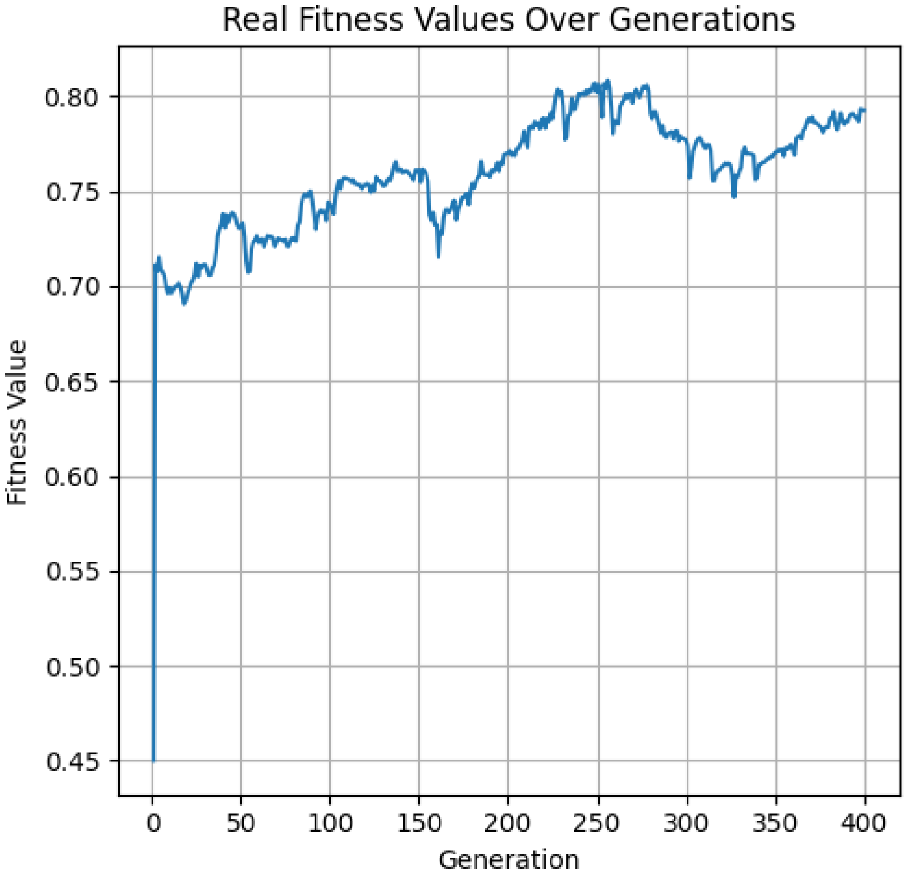
I think all of the questions are already answered and our time for the interview already reached the end. Before I close this discussion,

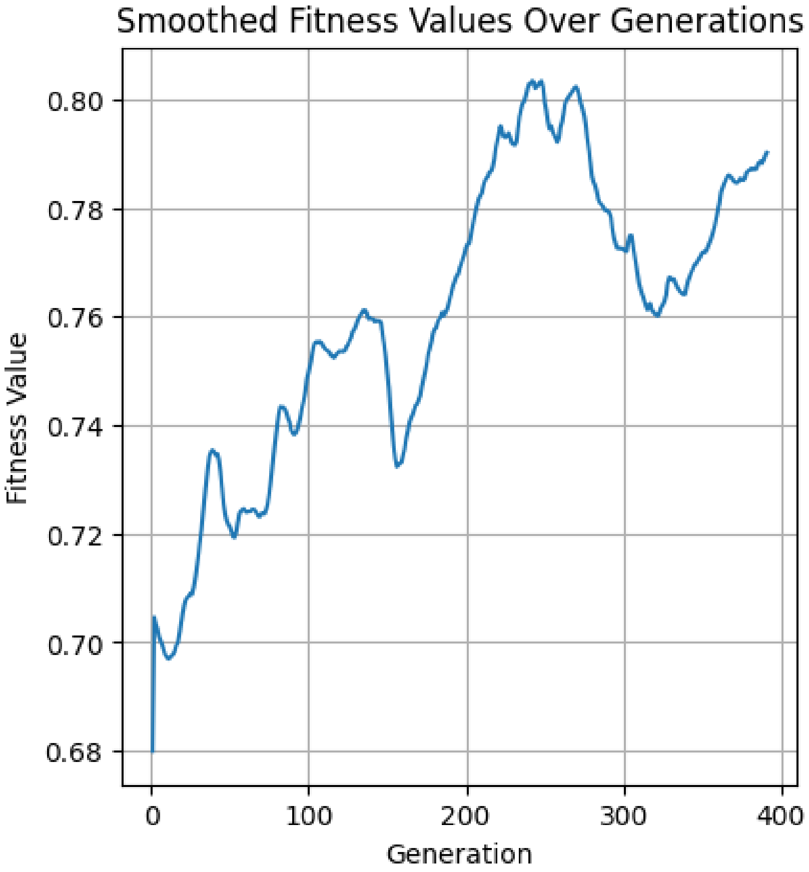
Do you have anything else to be said related to the interview process or suggestions for this research?

You can find my contact information in the consent form and I will keep your identity confidential and try my best to keep the transcript of this interview secure.

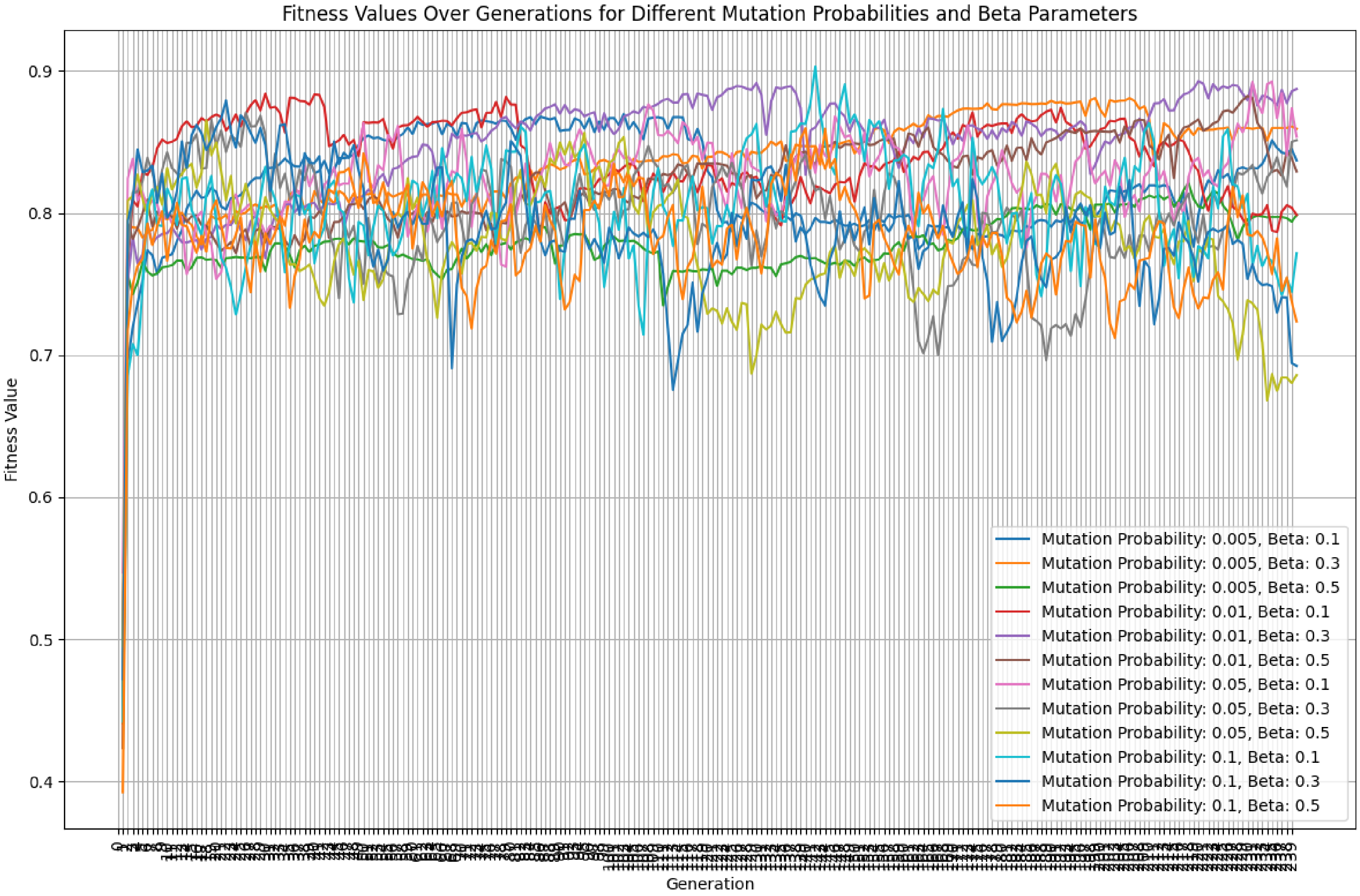
Thank you for your time.

APPENDIX 2: TESTING THE FITNESS VALUES OVER GENERATIONS

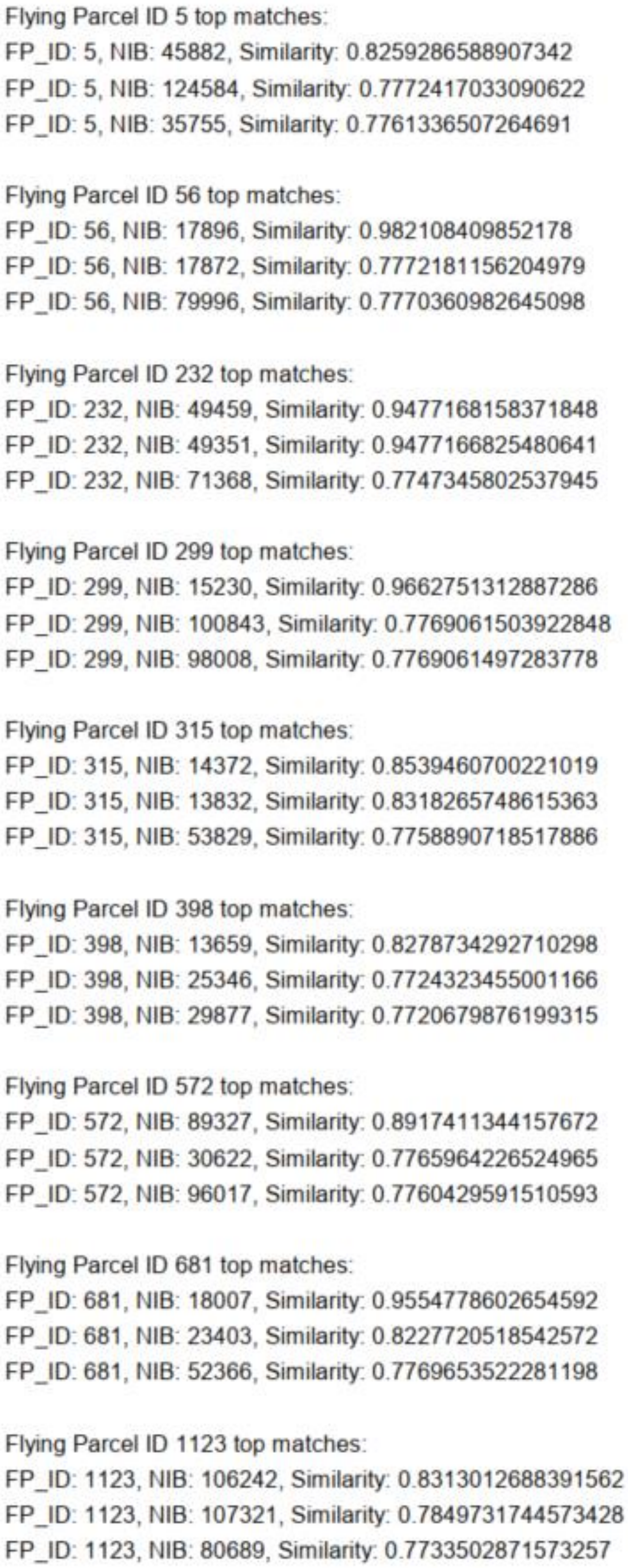




APPENDIX 3: TESTING THE FITNESS VALUES OVER MUTATION RATE (β) AND MUTATION PROBABILITIES



APPENDIX 4: SAMPLE RESULTS OF THE GEOMETRIC MATCHING



```
Flying Parcel ID 5 top matches:
FP_ID: 5, NIB: 45882, Similarity: 0.8259286588907342
FP_ID: 5, NIB: 124584, Similarity: 0.7772417033090622
FP_ID: 5, NIB: 35755, Similarity: 0.7761336507264691

Flying Parcel ID 56 top matches:
FP_ID: 56, NIB: 17896, Similarity: 0.982108409852178
FP_ID: 56, NIB: 17872, Similarity: 0.7772181156204979
FP_ID: 56, NIB: 79996, Similarity: 0.7770360982645098

Flying Parcel ID 232 top matches:
FP_ID: 232, NIB: 49459, Similarity: 0.9477168158371848
FP_ID: 232, NIB: 49351, Similarity: 0.9477166825480641
FP_ID: 232, NIB: 71368, Similarity: 0.7747345802537945

Flying Parcel ID 299 top matches:
FP_ID: 299, NIB: 15230, Similarity: 0.9662751312887286
FP_ID: 299, NIB: 100843, Similarity: 0.7769061503922848
FP_ID: 299, NIB: 98008, Similarity: 0.7769061497283778

Flying Parcel ID 315 top matches:
FP_ID: 315, NIB: 14372, Similarity: 0.8539460700221019
FP_ID: 315, NIB: 13832, Similarity: 0.8318265748615363
FP_ID: 315, NIB: 53829, Similarity: 0.7758890718517886

Flying Parcel ID 398 top matches:
FP_ID: 398, NIB: 13659, Similarity: 0.8278734292710298
FP_ID: 398, NIB: 25346, Similarity: 0.7724323455001166
FP_ID: 398, NIB: 29877, Similarity: 0.7720679876199315

Flying Parcel ID 572 top matches:
FP_ID: 572, NIB: 89327, Similarity: 0.8917411344157672
FP_ID: 572, NIB: 30622, Similarity: 0.7765964226524965
FP_ID: 572, NIB: 96017, Similarity: 0.7760429591510593

Flying Parcel ID 681 top matches:
FP_ID: 681, NIB: 18007, Similarity: 0.9554778602654592
FP_ID: 681, NIB: 23403, Similarity: 0.8227720518542572
FP_ID: 681, NIB: 52366, Similarity: 0.7769653522281198

Flying Parcel ID 1123 top matches:
FP_ID: 1123, NIB: 106242, Similarity: 0.8313012688391562
FP_ID: 1123, NIB: 107321, Similarity: 0.7849731744573428
FP_ID: 1123, NIB: 80689, Similarity: 0.7733502871573257
```

APPENDIX 5: SAMPLE RESULTS OF THE TEXTUAL ATTRIBUTE MATCHING USING THE VILLAGE NAME

Flying Parcel ID (FP_ID)	Matched NIBs	KELURAHAN (FP)	KELURAHAN (GeoDataFrame)
232	49459, 49351	gubeng	gubeng
299	15230	gundih	gundih
1598	85146	menur pumpungan	menur pumpungan
1812	78626	semolowaru	semolowaru
2327	16126	lingkungan tembok dukuh	lingkungan tembok dukuh
114	17898, 17913	lingkungan alun-alun contong	lingkungan alun-alun contong
69	17942	lingkungan alun-alun contong	lingkungan alun-alun contong
436	96089	keputih	keputih
489	97692	keputih	keputih
1054	41167, 41264, 37777	mojo	mojo
2090	16261, 16098	lingkungan tembok dukuh	lingkungan tembok dukuh
1093	37369, 42638, 37205	mojo	mojo
1289	111661	prapen	prapen
1967	120978	panjangjiwo	panjangjiwo