Location prediction using forward geocoding for fire incident

Faris Ade Irawan$^{1,2*}$, Projo Danoedoro$^3$, and Nur Mohammad Farda$^3$

$^1$Doctoral Student of the Faculty of Geography, Universitas Gadjah Mada, Yogyakarta, Indonesia
$^2$Geomatics and Survey Engineering Technology, Department of Civil and Earth Engineering, Politeknik Negeri Banjarmasin, Indonesia
$^3$Department of Geographic Information Science, Faculty of Geography, Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract. Urban fires, although not a natural disaster, are a severe threat that often occurs in urban areas. Banjarmasin City, the capital of South Kalimantan Province and one of the most populous cities in Kalimantan, recorded 159 fire cases between 2020 and 2022, averaging nearly 53 cases yearly. In today's digital era, people often share ongoing fire incidents using smartphones and update information on social media and online news. However, the resulting data could be more structured to serve as a dataset. This research addresses these issues by applying geocoding, a digital service that translates street addresses into geographic coordinates. This research uses three geocoders: Google Maps API, Bing Maps API, and Smart Monkey Geocoder. The accuracy of the three geocoders was tested using the Root Mean Square Error (RMSE) statistical method by comparing the geocoding results with valid locations. Prediction analysis was used to identify the next fire event through the density approach of the previous fire event points. This research is expected to provide insights into efficient data collection and structured data conversion, recommendations regarding the best geocoding service, and prediction of fire vulnerability locations based on recurring factors of fire incidents in the area. In conclusion, accurate data is the key to effective fire prediction.

1 Introduction

Urban fire incidents are not natural disasters, as they can be predicted, and scientific methods can be applied to reduce the associated risks [1]. Instead, they represent a significant threat that must be carefully considered [2], especially in urban areas. Notably, urban centers like Kota Banjarmasin, the capital of South Kalimantan Province and one of the most densely populated cities on the island of Kalimantan, have experienced a concerning trend. In 2020, 2021, and 2022, 159 fire incidents were recorded, averaging nearly 53 incidents per year, scattered across five districts, as described in Figure 1 below.

* Corresponding author: farisadeirawan1984@mail.ugm.ac.id

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).
2 Background

Nowadays, Technology is needed to obtain the data users quickly require and share, process, and analyses the data [3]. When fire incidents occur, one of the initial responses from the public is to document the events using smartphones and share this information on various social media platforms and online news outlets. However, these textual data collections could be more structured, necessitating further processing to render them usable. To achieve this, language-based program services, such as Application Programming Interfaces (APIs), serve as intermediaries connecting one application to another for geocoding purposes [4].

Geocoding, the process of converting textual data into geographic coordinates, is regarded as a significant and influential middleware service that facilitates various location-based services [5–9]. This technique is familiar [10], numerous online platforms offer geocoding services for free [11], reflecting its widespread availability. Given this diversity, there is a need to experiment with high-accuracy geocoding services. The data of addresses transformed into locations are gathered from social media platforms like Instagram, Twitter, and online news portals. This process is known as forward geocoding. This study employs geocoders such as Google Maps API, Bing Maps API, and Smart Monkey Geocoder. Despite the variety of geocoding applications available, one of the primary challenges in accurate geocoding is the availability of suitable reference data [12]. Position accuracy is typically measured as root mean square error (RMSE) [13]. RMSE is computed by assessing the accuracy of predicted points against several known or valid location points to obtain the most miniature possible score [14].

Effective prediction relies on sourcing data of the highest accuracy. The better the model, the higher its accuracy in predicting the occurrence and timing of events [15]. Prediction becomes feasible when a location frequently experiences incidents or exhibits a high intensity, as the likelihood of similar incidents recurring in the same area is significantly elevated. Forecasting theory argues that current and past information can be used to predict the future, especially time series. There is a belief that patterns in historical values can be found and effectively applied to predict future values [16]. The historical data of fire incidents are depicted through density analysis, which serves as empirical evidence of fire occurrences in Kota Banjarmasin and is a crucial variable in this study. Tools such as ArcGIS Kernel Density aid in estimating how the intensity of specific points is distributed across a defined radius [17]. Thus, this study aims to provide insights into the rapid data collection process and the transformation of unstructured data into a structured format. It also seeks to
recommend accurate geocoding services and predict locations susceptible to fire incidents due to the recurrence of such events in the region.

3 Methodology

For forward geocoding, users can provide addresses and place names with one or more address component parameters (e.g., city, state, country, etc.). For backward geocoding, the latitude and longitude parameters are specified with two parameters [18]. Although geocoding applications are diverse and include many types of applications, one of the main challenges of accurate geocoding is the availability of suitable reference data [12].

Based on Figure 2, the following stages of work are shown. In this research, we study the behaviour of fire events from past data to predict fire-vulnerable locations by utilizing address data of fire events documented on social media such as Instagram, Twitter, and online news portals. The first step in this process is data collection. Search and retrieve the address data of fire incidents reported and recorded on social media using the keywords: "Fire in Banjarmasin" along with the year of the incident. This data will form the basis of our efforts to understand the pattern of fires in the area we are studying.

Next step is data cleaning. Data cleaning is critical to ensure that our data is of high quality and free from errors or duplication [19]. This will help us minimize bias and improve the accuracy of the result. Apply forward geocoding to the cleaned fire address data. Forward geocoding is a process that converts these addresses into precise geographical coordinates, i.e., longitude and latitude coordinates on the earth's surface. That will help to visualize the fire events geographically. We chose Google Maps API, Bing Maps API, and Smart Monkey geocoder to perform the service from the many options available. Once we have the geocoded data, the next step is to test the accuracy of our model. The Root Mean Square Error (RMSE) statistical method measures our geocoding results' accuracy [13]. This is important because geographic data's accuracy significantly affects our predictions' quality.

Finally, the tested geocoded data to create a fire density model. The model will allow us to predict fire incident locations based on the recorded history of fire incidents. That is an essential step in our endeavour to help identify areas that need special attention in fire prevention efforts.

Fig. 2. Flowchart of the stages of work on the research.
Figure 3 shows an example of some addresses collected from social media and then cleaned for geocoding using smart monkey geocoder application. The address column is the address of the fire incident, the coordinates column is the geocoding result, and the address found column is the address found by the geocoder search engine.

<table>
<thead>
<tr>
<th>No</th>
<th>Address</th>
<th>Coordinates</th>
<th>Address found</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jl.Zafri Zam-zam Banjarmasin</td>
<td>-3.5141234,114.5704905</td>
<td>Jl. Zafri Zam Zam, Kota Banjarmasin,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Kalimantan Selatan, Indonesia</td>
</tr>
<tr>
<td>5</td>
<td>Jl.Sultan Adam, seberang kantor Capil, Banjarmasin</td>
<td>-3.3186057,114.5943784</td>
<td>Banjarmasin, Banjarmasin City, South Kalimantan, Indonesia</td>
</tr>
</tbody>
</table>

3.1 Validation

It is important to note that the geocoding results produced by various geocoders may vary in accuracy depending on the reference data and algorithm used. Position errors can occur if addresses are not written correctly and completely. Incomplete addresses constitute a significant determinant of geocoding position errors [20].

Root Mean Square Error (RMSE) is one of the metrics often used to measure the accuracy or error of the model in predicting points in the validation data [13]. RMSE measures how well the model maps the validation points into the predicted values by calculating the root mean square of the squared difference between the predicted values and the validation values.

In many cases, the RMSE value is calculated using a series of check measurements; these are coordinate values from independent sources that are more accurate for the same point. An error vector can be plotted for each measurement, as shown in Figure 4 below. The error vector has constituents in the x and y directions. These constituents can be recombined by vector summation to produce an error vector that indicates the location error.

![Fig. 4. The error vector [13.]

The formula commonly used for the validation of prediction points is:
\[ RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^\wedge)^2}{n}} \]  

Description:
- \( n \) is the number of data points to be validated.
- \( y_i \) is the actual or valid value at the i-th data point.
- \( y_i^\wedge \) is the predicted value at the i-th data point.

### 4 Results

#### 4.1 Validation of spatial accuracy

Unfortunately, recent research initiatives continue to use geocoded data without considering how the accuracy could introduce inconsistencies or bias in the results [21]. However, a wide range of literature exists that is specifically geared toward exposing how minor errors in geocoding accuracy can affect results based on detailed spatial models [12]. Geocoding quality relies heavily on the proper spelling of the user's address; if the address is insufficient, the result will also be harmful. Manual address correction is a cheap and efficient way to improve the quality of geocoded data [22].

Although the coordinates produced by geocoding use a geographic coordinate system, as shown in figure 3, when testing accuracy using the RMSE technique in ArcGIS, the two points, namely the prediction point and the validation point, are converted to the UTM coordinate system to obtain the value of the distance difference in meters, as shown in Figure 5 below.

Figure 5 shows the distance between the geocoded predicted coordinates and the validation points, comparing the RMSE values of the three geocoding applications. There is a significant difference between the geocoding coordinates using Google Maps API and Bing Maps API, but not with Smart Monkey Geocoder. The results are very similar to Google Maps API. Sub-figure A shows the RMSE value of Google Maps API geocoding, sub-figure B shows the RMSE value of Bing Maps API geocoding, and sub-figure C shows the RMSE value of Smart Monkey Geocoder. The RMSE value of geocoding points generated from Google Maps API is smaller among the three.
4.2 Fire incident modelling

From the previous accuracy test results, the lowest RMSE value is geocoding using Google Maps API. Banjarmasin city fire incident data in 2020, 2021 and 2022 geocoding results using Google Maps API are used to create fire models using the help of kernel density tools. It is a statistical method for estimating the probability distribution function of a variable in the absence of a known random shape or distribution model. In principle, the mathematical function used to calculate the Kernel Density is to estimate how the intensity of a point is distributed on a plane of a given radius [17]. Figure 6 shows the yearly distribution of fire occurrence points in 5 subdistricts in Banjarmasin City. Fires in 2020, 2021, and 2022 show similar patterns. The red colour indicates high density. When the three data are combined, the pattern of fire incidence can be observed. The red-coloured area is where fire incidents frequently occur yearly and may remain vulnerable in subsequent years.
5 Conclusion

Previous research mentions Cui et al., [22] several causes of geocoding errors and low match rates as follows: the quality of the submitted address [12], which may be inaccurate, such as spelling errors, inconsistencies in the use of prefixes, suffixes, or abbreviations, and incorrect postcode information; the quality of the reference file used, which may contain errors such as missing street segments, incomplete address ranges, or inconsistencies in the use of prefixes, suffixes, or abbreviations. The ability of geocoding software to match submitted addresses with existing addresses in the database is an example of the software's ability to identify that "PKWY" is short for "Parkway." Tolerance for imperfect matches varies depending on the algorithm or software used. Methodology used for the geocoding process.

The methodology used for the geocoding process in this research has used three geocoder applications with the same address, concluding that using Google Maps API has the best accuracy rate. This application can be an effective option as an approach and the initial stage of a more comprehensive spatial analysis.

Fire incidents in 2020, 2021 and 2022 were analysed using the point distribution from the Google Maps API geocoding. Banjarmasin Central and Banjarmasin West sub-districts have been fire-vulnerable areas in the last three years and will likely remain vulnerable to fire incidents in the following years.

6 Discussion

A review of geocoding studies, Abe and Stinchcomb found that average position errors ranged between 58 and 96 metres in urban areas and 129 to 614 metres in rural settings [23]. Bonner et al., recorded an average geocoding error of 96 m in urban areas and 129 m in rural
areas [24], while Cayo and Talbot found an average error of 58 m in urban areas and 614 m in rural areas [25]. Ward, Nuckols et al., recorded an average position error of 77 m in urban areas and 210 m in rural areas [26]. This study's highest geocoding position error was 18.742 m, or almost 19 km, by Cayo & Talbot [20]. The quality of geocoding depends on the correct spelling of the user-supplied address. If the quality of the given address is terrible, the geocoding result will implicitly be equally abysmal. The geocoding results are still better because the input addresses are in big cities such as Jakarta, Surabaya, and Yogyakarta. The geocoding result will likely be abysmal if the input addresses other small cities in Kalimantan due to the map database quality in those locations [11]. The geocoding spread point can be very far, even towards the nearest city with a similar address name.

References
3. K. Bok, Y. Kim, D. Choi, and J. Yoo, Sensors (Switzerland) 21, 1 (2021)
4. (n.d.)
13. (n.d.)


22. Y. Cui, Appl. Geogr. 41, 87 (2013)


