City of Chicago: Exploring Patterns of Crime and Discovering Crime Hot Spots

Dimitrios Megkos

Abstract—The aim of this study was to identify the changes in patterns of crime incidents that happened in the city of Chicago during the years 2001 to 2021 by analysing crime data provided by Chicago's Police Department. A further investigation of domestic violence during the Covid-19 pandemic period was conducted. A clustering method was applied in order to find Chicago's crime hot spots and identify which areas were the safest. Python was used for data preparation and all computational processes while visualizations were created using Tableau software. The analysis revealed that overall the number of crime incidents is decreasing every year. The Covid-19 pandemic had a negative effect on crime patterns as the percentage of domestic violence incidents was higher than previous years. Several crime hot spots were identified in the southern part of Chicago, showing that the northern part of the city was safer. These findings however also revealed that every district's police department was underperforming due to lack of personel, since the percentage of incidents that resulted in an arrest was small.

1 Problem Statement

Social order and compliance with basic laws can be considered the most important catalysts for a thriving society. However, the world is far from ideal, and it is impossible for societies to maintain a stable state all the time. When there is depression, unemployment, hunger, poverty, and other negative states, there is also crime.

The United States of America is undoubtedly one of the biggest and most advanced societies in the world. However, the total crime rate is higher than other developed countries, specifically those in Europe, with homicide rates being substantially higher [1].

The aim of this study is to investigate the crime patterns of the city of Chicago, the third-most populous city in the United States [2], throughout the years. More specifically, the research questions that were investigated are the following:

1. How have crime patterns changed throughout the years?
2. How did the Covid-19 pandemic affect crime patterns?
3. Which district has the most effective police department?
4. Which wards are the most dangerous and which areas are the safest to live?

The dataset used for the purpose of the analysis was extracted from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting system [3] and proved to be ideal for answering the research questions as each row contains a crime incident with information about the type of crime, date, location, whether an arrest was made or not, and other important details.

2 State of the Art

There are various papers that focus on identifying patterns in crime and attempt to predict the time and space of future incidents, as a means of being better prepared and having a safer society. Two papers were the main source of inspiration for this study's research questions. The first one acted as a stimulus to investigate the types of crime and the location description of incidents that happened specifically during the Covid-19 pandemic period, while the second one provided the idea of using spatial clustering and mapping as part of the analysis.

S. M. Perez-Vincent, and E. Carreras in "Domestic Violence Reporting during the COVID-19 Pandemic" [4] examined the changes in the frequency and characteristics of domestic violence crime reports during the pandemic period in six Latin American countries. The data was provided by different domestic violence hotlines, emergency lines, and police reports, containing information about the type of violence, the date, the relationship between the victim, and the offender and more, obtained from 2018 to 2020. The impact of the pandemic was assessed by comparing the month where the pandemic started, across multiple years and assuming that, if it wasn't for the pandemic, the reports would have had a similar pattern with previous years. A similar method was used for this study, investigating the locations of Chicago crime incidents, the type of crime, whether it was reported as domestic violence or not, and comparing the number of incidents during the pandemic period with the previous years, making the same assumptions.

X. Zhang, Z. Hu, R. Li, and Z. Zheng in "Detecting and mapping crime hot spots based on improved attribute oriented induce clustering" [5] discussed how important is the detection and mapping of crime hot spots, which are high crime density areas. The data used in that paper were similar to those that were used in this study. Each crime event that was investigated was a record that included information about the time the incident happened, the crime type, the location, the description of the crime, and so on. The method that was used is spatial clustering. More specifically, an improved attribute-oriented induce clustering algorithm was selected, due to the crime data having many attributes at different levels. The paper focused more on preprocessing the crime events before mapping and finding the best clustering method rather than the mapping itself. The aim of this study, however, was to map bigger groups of crime incidents in order to discover the most dangerous areas of Chicago. A simpler spatial clustering method was selected with the aim of answering the fourth research question. Several clusters were created using the k-means clustering algorithm, using the

location coordinates of each crime incident in order to identify bigger crime hot spots, which were visualized in a map using Tableau visualization software.

3 Properties of the Data

The dataset contains reported crime incidents that happened in the city of Chicago from January 2001 to December 2021. The data is provided by the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system and downloaded from Google Cloud as a Big Query public dataset. Although the reports included were unverified reports supplied to the Police Department, for the purpose of this analysis, an assumption was made that all reports were verified.

There are one hundred thousand rows, each row representing a crime incident, and twenty-two columns which are the crime incident features. Each feature provides information about each crime that happened, with the most important ones being the type of crime, the location where the crime happened, the date, an arrest indicator, and a domestic violence indicator. The dataset contains four different data types namely Boolean (two columns), float (four columns), integer (six columns) and object (ten columns).

The dataset was checked for any null values. There are 131 crime incidents missing location description, one missing case number, ten missing ward, and 949 that are missing the location. Because the location is important for the analysis and considering the small amount of missing data, these records can be removed from the dataset. The one with missing case number and those with missing ward can be removed too. As for the records with missing location description, they will be added to the "OTHER" category.

There are thirty-four unique primary crime types. A word cloud visualisation was used in order to find which primary crime types appeared more frequently throughout the years. According to figure one below, the five most frequent crime types were Theft, Battery, Criminal Damage, Assault, and Burglary.

Figure 1: Wordcloud of most frequent types of crime.

Figure two below shows the total number of incidents that occurred from 2001 to the present, focusing only on the top five types and how they compare with each other and the remaining twenty-nine types. There was a total of 98982 incidents remaining after dealing with the null values. Twenty-five percent was Theft (25129 incidents), twenty percent was Battery (20175 incidents), thirteen percent was Criminal Damage (12582 incidents) with Assault and Burglary both being eight percent (8133 and 7885 incidents respectively).

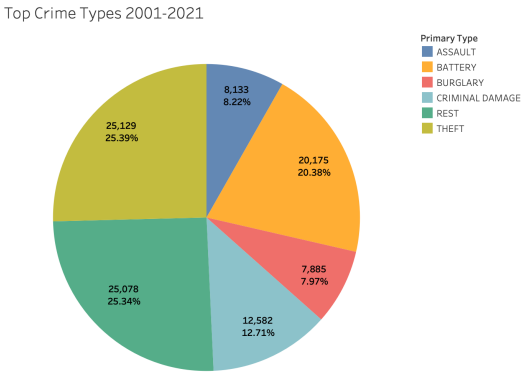

Figure 2: Pie chart of most frequent types of crime.

Overall, this was a well-structured and high-quality dataset, with very few missing values. The initial exploratory data analysis and the visualization of the top crime types suggested that there could be potentially interesting findings in further investigating the crime patterns throughout the years. Each crime incident report contained a great amount of information, which proved to be sufficient for investigating this study's research questions.

4 Analysis

4.1 Approach

The analysis approach that was used to investigate this study's research questions is described in the figure three below. Both the computer and the human mind played an equally important role in the analysis. The computer was used for managing thousands of data records, dealing with difficult computational processes, and visualizing the results. The human mind conceived the research questions by observing the data and the initial exploratory data analysis, and derived knowledge by interacting with the visualizations and applying analytical reasoning.

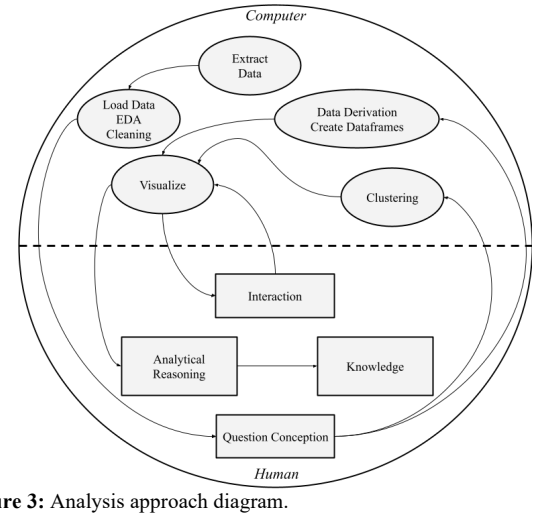

Figure 3: Analysis approach diagram.

Each step shown in the diagram above contributed greatly to the analysis process. The first step was to extract the data from Google Cloud and was loaded using Python. The next step was to clean the dataset and perform exploratory data analysis. The latter helped conceive the initial research question.

The next step was to prepare the data and derive new information in order to answer this study's research question. Initial visualizations and human interaction helped refine and create additional research questions. Plotting every data point on a map, as it was, proved that it lacked the required information to answer the fourth research question, which suggested the idea of using a clustering.

The new data frames that were created for the purpose of the analysis were loaded into Tableau visualization software. The main reasons why Tableau was used instead of a Python library were first the great number of visualization options and second the ease of use. Because Tableau has a user interface, it is easier for a human to interact with the visualizations on the fly and find the ones that are the most beneficial to the analysis.

The most important step in the above analysis approach was the analytical reasoning, which helped extract the desired knowledge form this study, by combining the visualization of results and their interpretation from the human mind.

4.2 Process

The analysis of this study was conducted according to the analysis approach that was described in the above section. The dataset used for the analysis was extracted by Google Cloud using a Big Query script. Due to size limitation reasons, only one hundred thousand rows were retrieved.

Following the loading and cleaning of the dataset, exploratory data analysis was performed that contributed towards the conception of the first two research questions. The dataset contained crime incident reports from year 2001 to year 2021. However, because there were only fifty-eight incident reports available for 2001, a number that does not seem to be realistic compared to the following years, all incidents that happened during that year were removed from the dataset. The next step to the analysis was to begin working towards answering the research questions. For every research question a new data frame was created using Python and then loaded into Tableau for visualization.

1. How have crime patterns changed throughout the years?

The question was answered by grouping crime incidents by year, counting the number of total incidents that happened during each year and visualizing the results using a line plot.

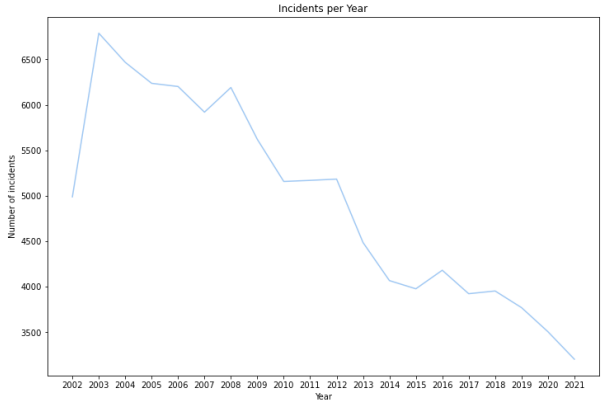

Figure 4: Line plot of total number of incidents, years 2002-2021.

According to figure four above, 2003 was the year with the most reported crime incidents of all time with more than 6500 incidents, following a rapid increase from 2002. One possible reason for this outcome could be retaliation for the 2003 invasion of Iraq from the United States, though there is not enough evidence to back this claim. From that point onwards, crime has been steadily decreasing every year, with years 2008, 2012, 2016, and 2018 being an exception having a minor increase in the number of incidents compared to the year before.

Looking at the behaviour of the top five crime types throughout the years, similar patterns were identified as shown in figure five below. The number of reported Theft, Battery, Criminal Damage, and Burglary incidents decreased while Assault incidents, though they slightly decreased at first, remained mostly the same.

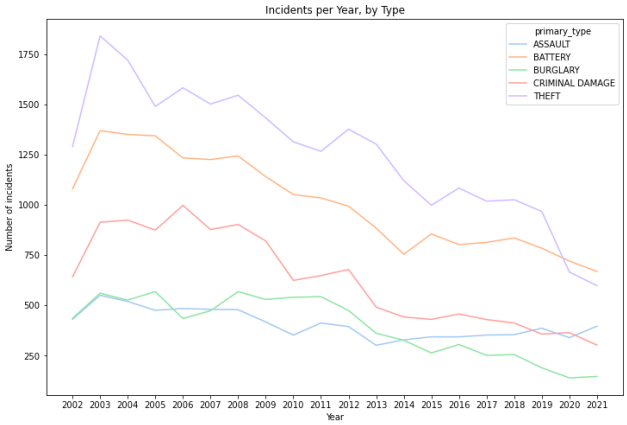

Figure 5: Line plot of number of incidents per type, years 2002-2021.

2. How did the Covid-19 pandemic affect crime patterns?

The dataset contains information on the description of the location each incident occurred. A few of the locations where most incidents occurred are Residence, Street, Apartment, Sidewalk, Parking, and Others. In order to investigate whether the Covid-19 pandemic had any effect on the crime patterns or not, the last two years of the pandemic were compared to the previous years. According to the figure six below, years 2016 to 2019 had similar patterns, with Residence incidents being around 35% of the total incidents, Street incidents being around 30%, and Apartment incidents being around 11%. However, after investigating 2020 and 2021, the years where lockdowns and home isolation were introduced, a few interesting differences were identified. Although the percentage of Street incidents remained the same, Residence incidents increased from 35% to 43% in both years and Apartment incidents increased from 11% to 17% in 2020 and to 20.5% in 2021.

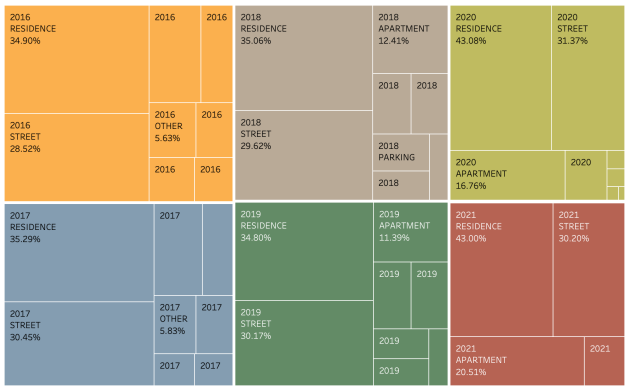

Figure 6: Tree map with locations where most incidents occurred, years 2016-2021.

Figure seven below shows the percentage of Domestic and non-Domestic incidents for years 2016 to 2021 where similar findings were identified. More specifically, domestic incidents increased from 17% (years 2018 and 2019) to 18.22% in 2020 and 21.19% in 2021.

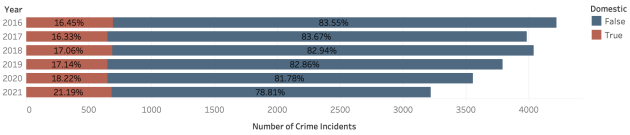

Figure 7: Bar plot of domestic crimes percentage, years 2016-2021.

4.2.1 Preparing for the next two research questions

Further data preparation was required in order to answer the last two questions. Columns that were not required for the analysis were dropped, arrest column was converted from Boolean type to integer (one where an arrest was made, zero where not), and a new column was created to indicate if an incident is a serious crime or not. For the purpose of the analysis, a crime incident is considered serious if it belongs to one of the following types: CRIM SEXUAL ASSAULT, ASSAULT, CRIMINAL SEXUAL ASSAULT, KIDNAPPING. The analysis focused on crime incidents that happened the last five years to explore the trends. A new data frame was created with district information, to answer question three. The data frame contains the mean location coordinates for each district, the total number of incidents and the number and percentage of arrests.

In order to find the most dangerous areas and the areas that are safer for citizens to live and answer the second part of question four, the k-Means clustering algorithm was implemented. All incidents were split into larger groups, based on the longitude and latitude coordinates of each incident. The dataset already contains information on each district; however, clustering was used to cover greater areas on the map and get a holistic view. The elbow method was used in order to find the best number of clusters. Five clusters were used in order to split the map into five big areas, where crimes took place in the last five years. Two new data frames were created with ward and cluster information, in order to answer question four. The data frames contain the mean location coordinates, the total number of incidents, and the total number and percentage of serious incidents, for each ward and cluster accordingly.

3. Which district has the most effective police department?

Using the arrest column of the districts data frame, the district with the most effective police department was identified, by calculating the percentage of successful arrests of each district and visualizing it on a map. A red-white-green color range was used to better distinct lower (red) and higher (green) percentages. According to figure eight below, though district six had the higher percentage of successful arrests (23.08%), it also had a very small number of crime incidents, hence there were not enough data to draw a conclusion of its effectiveness. Therefore, considering the number of total incidents, the district effective police department is district five (19.40%), followed by district four (14.53%) and district nine (13.88%).

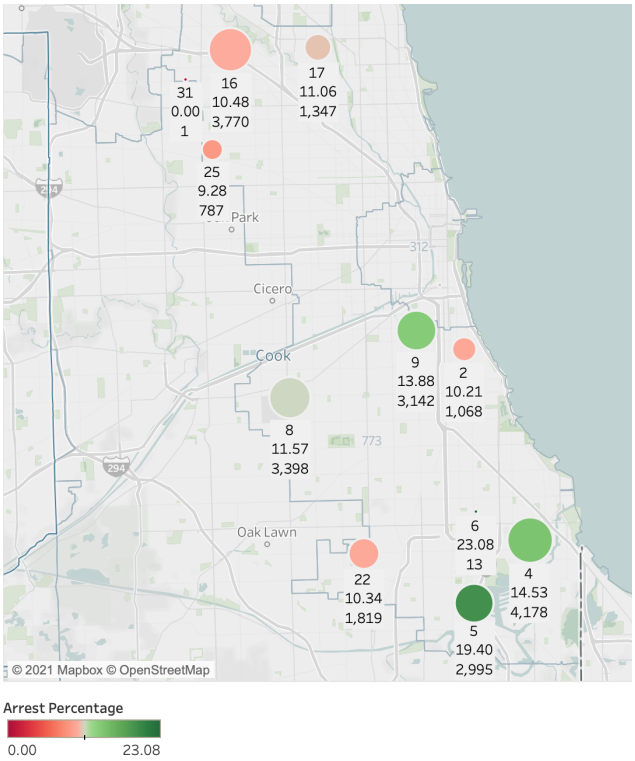

Figure 8: Arrest percentage per district, plotted on the map of Chicago.

4. Which wards are the most dangerous and which areas are the safest to live?

Using the new wards data frame, all twenty-two wards were visualized as circles on the map of Chicago. The size and the color of the circles were adjusted based on the number of incidents and the percentage of serious crimes that happened on each ward, during the last five years. A bigger circle means greater number of incidents. Similarly, deeper red color means a higher percentage of serious crimes. According to figure nine below, ward nine is the most dangerous ward of Chicago, followed by ward three and wards eight and ten. There are more wards with a deep red color on the map; however, their circle size is very small due to their number of incidents, therefore their percentage of serious crimes is not representative.

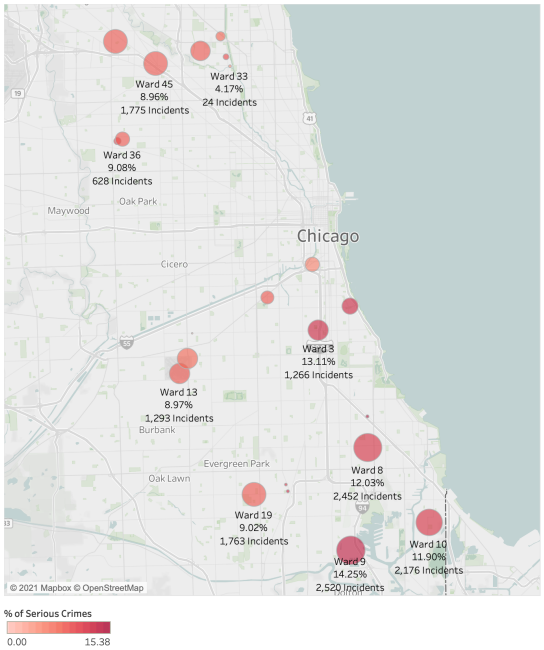

Figure 9: Most dangerous wards in the city of Chicago.

In order to find the most dangerous areas on the map and identify the locations that are safe for anyone that wants to move to Chicago, all incidents were grouped into five clusters. Each cluster's centroid was placed on the map as a circle. Each circle's size and colour were adjusted based on the number of incidents and the percentage of serious crimes. According to figure ten below, the most dangerous area is area two, having both greater number of incidents and higher percentage of serious crimes, followed by area three. Analytical reasoning suggests that these two areas should be avoided by those who worry about their physical integrity. Anywhere outside the circles of these areas should be safe to live.

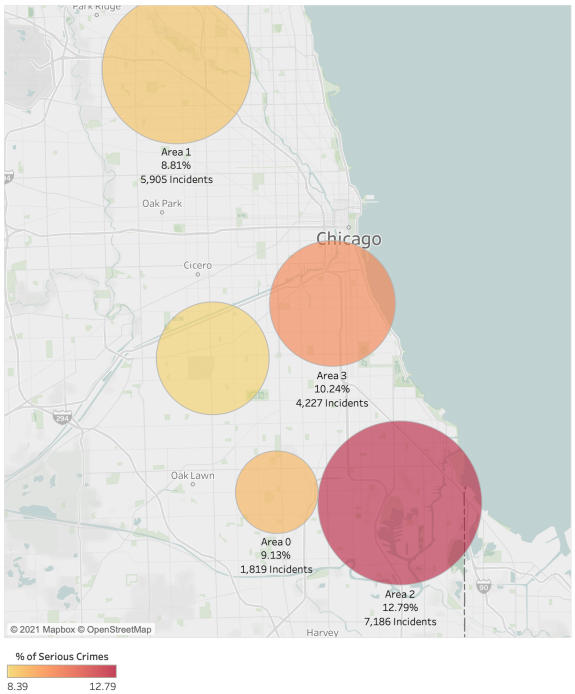

Figure 10: Areas to avoid living in the city of Chicago.

4.3 Results

Regarding research question one, the results overall were encouraging since there are now less than half crime incidents happening in the city of Chicago, compared to year 2003, and, according to the figure four, it looks like this number will further decrease.

The findings of research question two suggest that the Covid-19 pandemic had a negative effect in patterns of crime. Although there were fewer people on the streets because of the lockdowns, more people were self-isolating in houses. These circumstances contributed greatly towards the increase of domestic violence incidents.

Research question three results overall were not encouraging since, at best, only twenty percent of the total incidents ended up in an arrest. A conclusion was drawn with analytical reasoning that every district's police department lacks personnel, especially those in northern Chicago.

The map of research question four showed that most of the incidents happened on the central and southern parts of Chicago. Therefore, the best suggested area for someone to live, especially for families, is the northern part of Chicago, anywhere outside of the area one circle.

5 Critical Reflection

The dataset overall proved to be suitable for the investigation of this study's research questions. However, the following points need to be taken into consideration. First, an assumption was made that all incident reports were verified crime events. Second, only one hundred thousand rows were extracted from Google Cloud, due to computational limits. It would be great, as a future work, to use verified data without any limitations and investigate the findings.

Regarding research question two about the effect of the Covid-19 pandemic on crime patterns, the original hypothesis was that the difference in the number of residence and non-residence incidents that happened during the lockdown period would be great, compared to the previous years. Although the number of residence incidents were greater compared to previous years, the number of non-residence incidents were slightly smaller. A further investigation of this phenomenon could be conducted as a future work, using more data, in order to explain it.

Each incident report already included ward and district information, therefore research question four could be investigated without applying clustering. However, since the goal of this study was to also experiment and learn, k-means clustering algorithm was applied. Using this algorithm proved to be a great lesson for learning the basics of clustering; however, it would be interesting to explore different and harder clustering methods, as a future work, like the improved attribute oriented induce clustering algorithm, used by X. Zhang, Z. Hu, R. Li, and Z. Zheng in their paper.

The use of maps as a means of visualization and the application of analytical reasoning proved to be ideal for identifying safe zones in the city of Chicago. The next step of this study would be to investigate whether it would be possible to predict future crime hot spots on a map, by applying above methods and lessons learned.

Table of word counts

Problem statement	236
State of the art	466
Properties of the data	427
Analysis: Approach	307
Analysis: Process	1218
Analysis: Results	178
Critical reflection	305

REFERENCES

[1] UNODC (2014). Global Study on Homicide 2013. United Nations. ISBN 978-92-1-054205-0. Sales No. 14.IV.1.

[2] "Decennial Census P.L. 94-171 Redistricting Data". United States Census Bureau, Population Division.

[3] CPD, Reported crime incidents in the City of Chicago, 2001-Present, Provided from Google, Accessed on: 14/12/2021. Available at: https://www.kaggle.com/chicago/chicago-crime

[4] Perez-Vincent, Santiago M. & Carreras, Enrique, 2021. "Domestic Violence Reporting during the COVID-19 Pandemic: Evidence from Latin America," IDB Publications (Working Papers) 11716, Inter-American Development Bank.

[5] X. Zhang, Z. Hu, R. Li and Z. Zheng, "Detecting and mapping crime hot spots based on improved attribute oriented induce clustering," 2010 18th International Conference on Geoinformatics, 2010, pp. 1-5, doi: 10.1109/GEOINFORMATICS.2010.5568075.

[6] D. Megkos, 2021. “DM VA Jupyter Notebook”