“Mapping during Conflict: Assessing the Influence of the Russia-Ukraine War on mapping practice of OpenStreetMap Contributors”.

Module Code: CST40090 - Individual Data Science Project
Student name: Shruti Mistry
Student ID: M00880809
Supervisor Name: Dr Giovanni Quattrone

Table of content

Abstract

OpenStreetMap (OSM) is a free and open geo-spatial data platform that is created and maintained by open community of volunteers and collaborators. It stores global map data, which is freely licenced under Open Database licence, providing free access to its dynamic and updated geodata for worldwide locations. Thanks to the large and ever-growing community of cartographers that the volume and quality of the map continues to enhance, providing a valuable data resource for various non-commercial activities like Geo-spatial analysis, disaster response, humanitarian support activities, geographic research and land analysis. Also, in recent decade it is widely utilised for commercial usage like web applications and mobile applications.

Given its open availability, continues updating nature and vast contributor dynamics, OSM has seen a notable success in Volunteered Geographic Information (VGI) landscape, garnering much attention of researchers. Many studies have been carried out in terms of qualitative and quantitative data quality, map evolution through temporal and spatial dynamics, contributor behaviour in natural crisis response, bias in map production activities etc.

However, there is a gap in understanding the mapping behaviour as a response to man-originated crisis such as war. In this paper, we aim to study the contributor behaviour and their editing patterns following the recent events of Russia-Ukraine war and utilise the history data of OSM to conduct our analysis. The intuition behind this study is to acknowledge the changes occurred in mapping activities as an immediate aftermath of this war and how it can be further utilised to design a framework to promote and motivate contributors for active participation in response to such crisis. The tempo-spatial analysis of mapping activities as a preliminary analysis is used for war-affected region and is compared with control region which is not affected by the war. The counterintuitive results of this study show that there has been a decrease in map production over the period following attacks in the war- affected zones. The impact of war is also assessed on contributor engagement and observations disclose lesser number of users participating post-event. This study also constructs the contributor profile expressing the behavioural features extracted from user mapping patterns foreach of the users actively participating during the crisis.

Introduction

Cartography, a discipline dedicated to creating and designing visual presentation of graphical features of the earth over a flat surface such as paper or walls is an ancient form of art used for navigation and routing across different locations. Since its first recorded formation during the Greek civilization, the practice of mapping the geography has been maturing with time and with the advancement in the scientific study related to earth surface, mathematical

research and increasing resource availability. As J.B. Harley explains the significance of maps in civilization history, it is an 'extraordinary authority' of graphical language used to describe spatial representation that can easily be perceived by the eyes.[1] Earth is an ever-changing landscape and therefore cartography practice keeps evolving with time to reflect the changes and include accurate information. And with the breakthrough of internet, the mapping practice sought a major shift from being tangible articles providing only static information and became digital representation of interactive geoinformation. Emergence of Global positioning System (GPS) with satellite images and remote sensers and the introduction of Geographic Information Systems (GIS) tools and software usage has enabled modern cartography to create sophisticated maps of the global locations and made them available on web for online access. Meng [2] has highlighted the freedom of geo-visualisation in digital maps that is possible with continuously improving multimedia technology and growing computation resources. Today online map services like Google Maps, Microsoft Bing Maps and Apple Maps have become essential source for many applications to offer real-time location data.

This technological freedom of creating and annotating geo knowledge has emerged the concept of Volunteered Geographic Information (VGI) which is a crowdsourcing approach of producing geographic data and information by means of open collaboration of individuals, volunteers, non-profit organizations and large communities that contributes towards generating location-based data. Success of VGI is attributed to web 2.0 by Michael F. Goodchild in [3] that expanded web interactions and termed VGI as a special phenomenon of user-generated web content. OpenStreetMap (OSM) is a result of such collective efforts of volunteer mappers. A crowdsourcing project initiated to build an open-source mapping platform; it offers freely accessible geographic database of the world [4]. Since its inception in 2004, OSM has gained huge attention from crowdsourcing community that has led to continuous improvement in its framework and coverage of spatial knowledge. Although it is constantly updating, the application of OSM is diverse [5] and widely used for building road networks, urban planning, land use classification, validating land-cover maps [6] location - based map services. [7] Its widespread expansion and growing prominence has also made it instrumental tool in many disaster response activities like disaster risk reduction and support efforts. [8]

OpenStreetMap is a vast and ever-growing repository of geospatial information contributed by diverse demographic groups having different skills, motivation and interests. However, the nature of this data and its model has made it crucial to evaluate the database with respect to quality, consistency, accuracy, privacy and ethical aspects as well as technical paradigms.
Various study lead by academics and institutions have attempted to understand the integrity and comprehensiveness of OSM database and determine directions for improvements for respective research subjects.

Scholars have been actively studying multifaceted subjects relating to OSM for example, quality assessment of the data is crucial to determine the usability of the map for high -end applications [9]. Equally important is semantic accuracy [10] of the data being produces as that benefits its local users in coordinating relief efforts during crisis and public health campaigns. Logical consistency, timing precision and completeness are also important criteria to assess the map.[11] The objectives of different studies may have different motives, but the

goal is to provide constructive ways to establish concurrent and reliable source of geographic information.

By the definition of crowdsourcing, the success of OSM is solely imputed to its growing crowd of contributors' community. In fact, "Maps are ultimately embodied subjects that cannot be considered separate from the people that created them." [12] The quality of data cannot be viewed in isolation from the actors behind this data production. It is therefore imperative to shed light on the footprints of the OSM contributors to understand their behaviour and demographics presence as well as quantify their actions to scrutinize their participation at a broader level. Previously, scholars have ventured to study the contribution patterns of OSM mappers in terms of information quality. Evolution of the community is understood in response to humanitarian mapping efforts that provides useful insight on developing better mapping methods. Gender bias is also one concern of similar research as it can lead to imbalance in the type of objects being mapped. There are also present several works that analysed the community response in times of crisis with the interest of investigating the participation level and user engagements.

The prospects for understanding contributors' dimension are wide-spread and of growing interest for scholar works, yet it is observed that there are various potential areas which are left untapped and can benefit if well understood. Because of its extensive application during the crisis and disaster response, a thorough study of OSM contributors can prove valuable. Reviewing the previous works proposed in this orientation provides diverse insights on mapping behaviour during natural disasters. However, there is little understood about the change in OSM contribution in the wake of human-triggered crisis like war and conflicts. Unlike natural calamities, war-like situations are recurring and have impact on a range of places that are geographically disparate. Such conflicts also forces border restrictions and may impact humanitarian support efforts.

In this analysis, we consider the recent events of Russia-Ukraine conflicts and study the mapping behaviour of OSM contributors for this timeline. To outline the aim of this analysis, we propose following research questions.

1) *How has the ongoing war influenced the contribution activities for Ukraine? We study the temporal and spatial analysis of mapping activities for the study regions. The impact of bombing on contributor engagement is also evaluated.*
2) *Uncover the user base contributing to the map post bombing event. The resulting user profile contains informative behavioural characteristics of mappers who are active during the wartime.*

The aim of this study is to provide a useful insight on community response during wartime, which can potentially help in harnessing the efficiency and skills of mappers for better map production initiatives. Also, understanding of user behaviour has intrinsic importance from future perspective in terms of quality of map and forecasting map changes.

The following section will inspect on the previous work related to our research subject and evaluate the methodologies. In the subsequent sections we will discuss the data being used for this study and outline the methodology designed for our analysis. Following section will

discuss the results obtained, followed by challenges in current study and presenting its limitations.

Literature Review and Previous Work

OpenStreetMap has garnered attention from various research studies for its geographical coverage and adaptation with mapping technology transformation. These studies have ventured on understanding the different aspects of OSM database that are important factor to assess the overall development. This section reviews some of the previous works that illustrate the ideology behind their work and the significance of studying the theme they represent.

(1) Spatial-Temporal dynamics:

The evolution of the map as a result of humanitarian mapping efforts is studied by Benjamin Herfort et el. [13] to understand the impact of large-scale mapping efforts of HOT tasking manager on OSM community and their map contribution using spatio-temporal analysis of OSM history data over a decade. The study reveals the scale and demographics of the contribution done by mappers and influence of socio-economic attributes on the result. The results highlight the disparity between the distribution activities and the distribution of global population.

The temporality of these mapping activities is exclusively studied in [14] for large cities across the globe. The study seeks to understand better understand the most consistent time period to understand the temporal dynamics of the data contribution activities using time series forecasting and decomposition and determines the temporal correlations with contribution activities. The study also explores the methodology with forecasting models and evaluates different accuracies achievable with the use of these machine learning models. This approach of understanding the changing structure of database is very useful for estimating the data production during the maintenance of applications using the data from OSM.

(2) Quality and reliability of the data

Understanding the quality of data produced is also one of the core features studied by scholars to determine the usability of data in different situations and applications.
For example, the quality of data as a metric of user behaviour and participation pattern is analysed in [11] where the level of participation is compared across three study locations. Ther result suggests that while 90% of data is contributed by only a small percentage of users, the information loss by ignoring the rest of the users' contributions is significant as such information can be valuable in deriving completeness and quality of the whole dataset under consideration.

Intrinsic measure of data quality is further explored in [15], focusing on aggregated expertise of user contributions to outline the credibility of OSM features being mapped. This work seeks to determine accuracy, completeness, and consistency of the

features by creating a reputation model of contributors based on the collaborative contribution and expertise. The study emphasis on the idea of considering the contributor skills, activity time, editing pattern, location as the important features in assessing and predicting future quality of OSM data.

(3) OSM data production in response to disaster events

Given the importance of crowdsourced data during crisis, OSM data is well studied in relation to natural disasters of different scales and community response for these events.
The effect of earthquakes on OSM mapping behaviour is analysed by Ahmed Ahmouda et el. [16] for Nepal and Central Italy regions over short-period and long-periods. The analysis compares the effect of the earthquake on OSM contribution in relation to mapping campaigns. Locality of user is also considered to differentiate between local user and remote user, for which geo-referenced tweets of the mapper were used. The power of social media for real-time information retrieval is well utilised for undiscovering the user demographics for this work. The analysis uses monthly frequency plots of different operations performed on OSM features over a period of year and statistical method to uncover the difference between affected areas and reference location. The results reveal an increase in contributors post disaster- both local and external, as well as strong effect of the disaster on features being mapped by these contributors.

A similar study by Kamptner & Kessler [17] covers similar concept and compares user contribution level during small-scale crisis. Here the community response is studied for four fire-events, where the editing pattern of mappers such as contribution frequency, primary edits and spatial coverage is considered to evaluate the contribution pattern. They are further classified based on their level of contribution.
It is noted that in three of the four events studies, there was an increase in mapping after the incident. Also, the active mappers were only a small percentage of total contributors for these locations. This study also presents a method to estimate locality of mappers based on the centroid of their editing, which can help in understanding whether the contributors are done by local or external user when explicit user location data is not available.

(4) Mapping behaviour of OSM contributors and their classification

While previously presented works analyses the quality of data by comparing it with external databases, Jacobs and Mitchell [18] explores the application if unsupervised machine learning for deriving intrinsic quality of the data for the study region. The analysis uses pre-existing metadata of OSM elements and OSM user to group the OSM contributors using clustering algorithm with PCA (Principal Component Analysis) to identify mapping behaviour of expert users, which can further be used to characterize the quality and accuracy of the features mapped. This approach of quality assessment as an intrinsic feature of data opens the possibility to examine the database in absence

of any external data source to validate against. However, the paper does not explore any other classification methods to compare and evaluate the results.

A more refined approach is proposed in [19] for OSM contribution classification. In this paper, the metadata of OSM elements for London area are extracted using Oslandia tool. These elements are weighted and then weighted principal component analysis (WPCA) is applied for dimensionality reduction. The paper also discusses the methods for component selection and weight selection for WPCA. This study further performs the classification of contributors using Gaussian Mixture Model Clustering (GMM), which is an unsupervised clustering algorithm. This combined approach of classification addresses the challenge of ambiguity present in understanding the components of PCA.

For analysing the editing patterns, Zhao & Fan [20] suggests that drawing directions and spatial proximity of their mapping can be useful features for inferring contributor attention and mapping experience for grouping contributors. Result compares the metrics across four countries and shows high level of dispersion in densely populated areas while low dispersion in less populated areas. Their work presents an effective statistical method to understand spatial distribution of mapping activity which can be useful in deriving user behaviour and motivation in mapping disaster-struck areas.

Exploring these papers as an example to summarize the work done so far, we get an overview of the important factors affecting the data well as contributor characteristics being considered to facilitate these works.
However, when we consider the research done for analysing the impact of disasters and crisis, there is only little understood about the community response and majority of it involves studying OSM data contribution post natural disasters such as earthquakes and Tsunami. This reveals a lack of diversity in the events being selected. Human-man crisis such as war and conflicts are not natural phenomenon but rather a consequence of human actions and are recurrent in nature. Therefore, understanding the community response during such events is imperative as it offers valuable insight on resulting changes on map quality as a reflection of ongoing event.
Above papers utilise various properties of contributors. However, most of them are raw features found in the data. Uncovering the contributor profile using their mapping behaviour provides a comprehensive approach to feature extraction for future works.

Data Introduction

For this analysis, OSM history data is used for Ukraine region. OSM history data contains the historical records of each data point being mapped which gives information about all the changes made to that record over the time. The version number associated with each record gives information about the order in which changes were done.

OSM data has three primary elements:
Nodes: nodes are individual location points representing point of interest such as restaurant, bust stop with their latitude and longitude coordinates.

Ways: ways are formed by connected line of nodes and the sequence of nodes define the order in which they are connected. Ways are used to denote linear or polygon geography depending on how they are used. Linear ways are used to map roads, rivers, railway tracks etc. Polygons are mapped using closed loop of way where the first and the last nodes are same. They are called area and used to map features like building.
Relations: They are used to define complex data structures which are related but not physically connected. They are used to define relationships between other elements such as nodes, ways, or even other relations.

All these primary elements can have tags associated with them in the form of key: value pair that hives descriptive details about what this element contains. Key: highway can be used to tag roads or footpath etc There is no predefined set of values that tags can have as more and more new features are being mapped. Some of the important tag keys are highway, building, amenity, place etc.

Data parsing:

The raw data extracted for this study containing the changeset and edit details is available in compressed format of parquet files (. parquet) which is a file storage developed as a part of Apache Hadoop system. It is a splitable columnar based file format to store large data. Since the data is large and stored across different small files, directly working with parquet file adds complexity and therefore, forr the analysis purpose the data is first converted into CSV format using python API PySpark for Apache Spark. This API has package dependencies on other libraries for installation. This is addressed by using Google Colaboratory for the conversion. Google Colab provides server hosted Jupyter notebook services with pre-installed libraries and packages for diverse range of purpose. The resulting csv files from this conversion contains uncompressed history data of Ukraine split in 500+ separate csv files. To build a coherent and meaningful data, these files are imported into uniform database format. To import into structured format, PostgreSQL database with PostGIS extension is used. PostgreSQL is an open-source relational database management system and with PostGIS extension it allows to store and query geographic and spatial data.
These csv files are merged and imported into PostgreSQL database format using direct import feature of PGSQL. The table schema is as below:

User_info:

	changeset bigint	deleted text	id bigint	timestamp timestamp with time zone	uid bigint	user text	version integer	visibility text	tp text	extra bigint
1	90461009	false	7879285261	2020-09-05 19:10:24+01	4808647	Евтушок Алексей	1	true	node	0
2	90461009	false	7879285262	2020-09-05 19:10:24+01	4808647	Евтушок Алексей	1	true	node	1

node_locations:

	id bigint	latitude double precision	longitude double precision
1	161057	50.5264549	30.1474571
2	161057	0	0
3	10980416	46.4651993	30.7353371
4	10980416	46.4651993	30.7353371

node_tags:

	id bigint	tags text
22	10980419	{'highway': 'traffic_signals'}
23	10980419	{'highway': 'traffic_signals'}

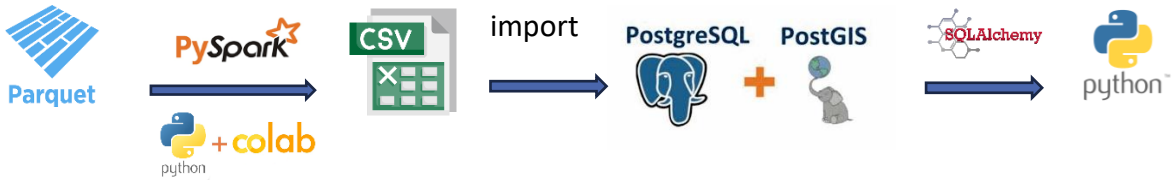
For data access purpose, two new tables are created by joining user_info table with node_locations for two different time periods. Filtering by time allows to join two huge data tables and faster data retrieval.

Table schema for merged table:

	id bigint	changeset bigint	deleted text	timestamp timestamp with time zone	uid bigint	user text	version integer	visibility text	tp text	extra bigint	latitude double precision	longitude double precision
4	2367187013	120317993	false	2022-04-28 19:18:35+01	6617477	deindsin	5	true	node	212826	45.4532797	29.2757653
5	2367187013	120317993	false	2022-04-28 19:18:35+01	6617477	deindsin	5	true	node	212826	45.4532251	29.2757452

In each record, *changeset* value represents the id of the changeset in which the given edit was posted. Each changeset can have multiple edits grouped together in one changeset. *id* is a unique value given to each data location being mapped by contributors. Since, history data contains the history of all the changes made to given location, there can be multiple records present with the same *id* value.t timestamp gives details about precise date and time at which the edit was posted by the user. *Uid* and *user* denotes details about the contributor, uid is the unique identifier and user is username chosen by the user. Since user can modify username at any time, we will consider only *uid* to refer users as *uid* is unique and cannot be updated during its lifetime. *version* is the number associated with each record. It starts at 1 and is increased by one every time an edit is posted for that record. *visibility* and *deleted* gives information about whether the record still exists in the database or deleted. Visibility= false and deleted= true means the record has been deleted. *tp* suggest the type of element, whether it is a node, way or relation. *latitude* and *longitude* present geographical coordinates for that data point.
OSM user can either add new data point, modify it or delete the point. All changes are posted using changeset which is a logical container allowing user to post group of edits in a single timeframe.

Analysis on data is conducted using Python with Jupiter IDE. Data tables are accessed from PostgreSQL server by establishing connection between notebook and server using *SQLAlchemy* toolkit which allows to load the data table by creating engine object and passing SQL query for data filtering and retrieval.



Data cleaning:

When we perform a table merge using the 'id' column as our merging key, an issue of duplicate records is encountered. This is primarily because the 'id' column does not have unique values as historical data contains multiple versions associated with the same 'id.'

After importing the data into a Pandas DataFrame, the initial step involves eliminating these duplicate records by considering a combination of both the 'id' and 'changeset' values to uniquely identify and remove redundant entries. Furthermore, within the data imported from the 'node_tags' table, we observe a prevalence of null values. To streamline our analysis, we opt to filter this table, retaining only those records where the 'tag' column holds non-null values.

Analysis Framework

The research period spans six months before and six months after February 24, 2022, a significant date marking the invasion of Russia and the commencement of major bombing campaigns in the selected regions. This timeline is considered optimal for understanding the temporal dynamics of contributor activities. For the comparison purpose we also study Zakarpattia Oblast, which is located in western part and has reported no active bombing during the ongoing conflict. Based on previous work, optimal timeline of four months is suggested to understand temporal dynamics of mapping pattern. Using this and to cover more subsequent bombing effects, we select the timeline of six months for this study.
According to multiple reputable news sources, it is evident that the Russian invasion and significant bombing campaigns commenced on February 24, 2022, impacting all the key locations under consideration for this research.

Each individual site is analysed separately using the location data associated with each record. The data is filtered for each location using bounding box which contains minimum and value of latitude and longitude such that the box covers the whole region. Bounding boxes are retrieved Nominatim API (https://nominatim.openstreetmap.org/search). Nominatim API is a free-to-use geocoding service that allows users to retrieve geographic coordinates based on location names or addresses via search queries and provides the results in JSON format. This data extraction method ensures the precise geographical coverage of the selected regions, enabling a comprehensive analysis of contributor mapping activities within each area.

Method for RQ1:
To address the first research question comprehensively, the analysis is structured divided into three subsections, and we conduct the analysis over a timeline spanning six months both before and after the pivotal event under consideration.

(1) How much contribution is made and how did the users contribute?
(2) What users have mapped during this period?
(3) Where did the users contribute?

In the first subsection, we focus on analysing the temporal dynamics of data for each specific location. This involves examination of daily contributions made over distinct time intervals: one week, one month, and six months. These intervals are chosen to provide a comprehensive view of how mapping activities have evolved over both the short and long term, both before and after the significant event under study.
Using statistical measures- mean ($\bar{x}$), standard deviation(σ) calculated for each time intervals and comparing them across two different timelines give us an overview of the change in mapping frequency.

To delve even deeper into these temporal patterns, we use time-series decomposition disentangle the data into its underlying components, namely trend and seasonality. This process is invaluable in understanding how mapping activities have changed over the entire duration of our analysis.

Trend Component: The trend component reveals the overarching, long-term pattern in mapping contributions. It helps us identify whether there has been a sustained increase or decrease in mapping activities over the entire period.

Seasonality Component: The seasonality component uncovers if there is any recurring patterns or cycles within the data.
After testing for both, additive and multiplicative model, we select additive model as the magnitude of seasonality remains constant over the time period. The window size selected for calculating the moving average is 2 for weekly data and 7 for rest of the period.
The additive model is given by,

$$Y(t) = T(t) + S(t) + R(t)$$

where y(t) is the raw series, T(t) is the trend-cycle component, S(t) is the seasonality and R(t) is the residual component at time t.

To gain insight into how these contributions were made, we distinguish between human contributors and automated scripts. This differentiation is crucial in understanding the nature of the contributions. To achieve this, we define a function that assigns each contribution to one of two categories: "human" or "bot." This assignment is based on the number of edits contained within each changeset. The function is given by,

$$f(x) = human, if\ x < 1000$$
$$bot, if\ x \geq 1000,$$

where x is the count of edits contained in each changesets posted on the given date,
f(*date*,*changeset*)=*Count of occurrences for the combination of date and changeset*

In the second phase of our analysis, we use the tag data associated with each record to gain insights into the types of locations that contributors are adding to the dataset. The tags are stored in dictionary format and may contain number of *key:value* pairs and there are no specific keywords used to define these keys. To make this data more comprehensible and accessible, we first use data transformation process and convert the tags from their dictionary format into a table format. Each column of the table corresponds to a unique key, and each row contains the value associated with that key, provided it is available for the 'id' being referenced in that row.

In the next step, we aggregate the total number of times each tag is added to the dataset. This aggregation allows us to quantify the frequency of each tag and gain insights into which tags are being mapped with high frequency. To visualize and interpret this distribution, we generate a histogram that displays the frequency distribution of tags. we set a threshold where tags having a count of more than 5000 occurrences are selected for further evaluation. This threshold helps us focus on tags that are significantly represented in the dataset. Also, the count of unique values added for each of these selected tags is calculated. This additional metric provides information about the variability within each tag category.
By comparing the tags and their frequencies before and after the event, we can capture the changes in semantic annotation.

In the third analysis, the location data is used to understand the spatial mapping patterns of contributors. We leverage the latitude and longitude coordinates associated with each location record and plot them onto a map of Ukraine, utilizing a shapefile as a geographical reference.
To understand mapping behaviour at contributor level, we create two different plots:
First plot visualizes the locations where each distinct user has mapped their first nodes. The second plot uses mean value of the location coordinates of nodes created by the user. Mean value captures the focal area of mapping for each user.

To evaluate the dispersion of mapping, complete dataset (pre-bombing and post-bombing) are considered. The dispersion degree is measured using the calculation of population coefficient of variance (CV) for both these datasets. The CV measure is a useful method to understand the relative variability of each dataset and is expressed as a percentage.

$$CV = {}^{\sigma}/_{\mu} * 100$$

where:

- σ is the standard deviation of the population.
- μ is the mean of the population.

Using the coefficient variance values, comparison of different datasets having different means and standard deviation can be made and the variability within each dataset can be explained. A higher CV indicates greater variability, while a lower CV suggests less variability.

Impact on contributor engagement:

To assess the impact of bombing on contributor retention and activation, following steps are performed.

Group the data by user ID (uid) for both the pre-event and post-event periods.
Collect the user IDs into two separate lists, denoted as A and B, respectively.
Calculate the intersection of these lists (A ∩ B) to determine the number of users who remained active post-event.
Calculate B - (A ∩ B) to identify contributors who started mapping for the first time in this region, which refers to both – entirely new users and reactivated users.

Method for RQ2:

In this analysis, we create a comprehensive contributor profile by utilizing feature extraction technique by quantifying the mapping patterns of each contributor across various parameters. For this purpose, the Pandas library in Python is used, which provides powerful tools for data manipulation and analysis. We calculate following measures to understand the participation level of each user.

1. Total contribution: To calculate the total contribution made by each user, the data is grouped together based on contributor ID. This grouping operation clusters together all contributions made by each individual contributor. Subsequently, we calculate the total count of these contributions associated with each contributor ID.

2. Changesets: We again perform grouping on data and count the unique values of changesets to calculate total unique changesets contributed by each user.

3. Contribution level: To understand whether the contributors are actively mapping or hit and run, we classify them into three different categories using the values from 'Total contribution' feature. Quantile values Q1 and Q2 for editing frequency of six months are used to determine groups as follows:

   - *inactive mapper: < 2000 edits*
   - *casual mapper: 2000 - 6000 edits*
   - *committed mappers: > 6000 edits.*

4. Deleted nodes: To calculate the total number of deleted nodes by each user, we focus on the data column 'deleted' where the value is marked as 'True.' We aggregate this data by grouping it based on user IDs. Within each group, we count the occurrences where the 'deleted' attribute is marked as 'True.'

5. Node versions: The version history of each node provides the details whether the node is a new node or an edited node. Using this information, the frequency of new nodes and edited nodes is calculated by grouping data on each user and for each group, a list

of versions is extracted. Function to calculate the new nodes and edited nodes are given by,$f(x)_N$ and $f(x)_E$ as below:

$$f(x)_N : \sum_{i=1}^{n} x_i , x=1$$

$$f(x)_E : \sum_{i=1}^{n} x_i , x \neq 1$$

Where, i= *ith* element in the version list and
x= version number

6. Lifespan: To calculate total number of days for which the contributor was active, the date of first contribution and last contribution is extracted. This information is derived from the timestamp associated with each of the user's contributions The difference of these two dates gives the number of days the user was active.

7. Tag contribution: To assess a contributor's involvement with different tags, we utilize the tag dataset. A dictionary is created by grouping the data based on tag keys. Within each group, we calculate the frequency of each unique value of key and repeat it across all user IDs. Calculating how often a specific tag is used by a user, we can understand the tag preference of each user.

Result and Evaluation

Evaluation for RQ1

(1) How much contribution is made and how did they contribute?

The table below summarizes the contribution made during different timeframe at different study areas. The time series decomposition plotted for three periods- 1 week, 1 month and 6 months shows the trend and seasonality components in contribution level.

	Timeframe	1 week		1 month		6 months	
Region	**Measure**	**Before**	**After**	**Before**	**After**	**Before**	**After**
Kyiv Oblast	Mean	3113.42	3951.57	5676.48	1652.50	4957.90	2882.13
	Total edits	21794	27661	175971	49575	912254	527430
	Days_mapped	7	7	30	31	184	183
Kherson Oblast	Mean	437.42	1872.28	3368.87	1011.79	2771.81	705.56
	Total edits	3062	13106	104435	29342	510044	112185
	Days_mapped	7	7	31	29	184	159
	Mean	6381.2	164.85	8559.45	902.83	5595.73	821.94

Kharkiv Oblasts	Total edits	44669	1154	265343	27085	1029616	146307
	Days_mapped	7	7	31	30	184	178
Donetsk Oblast	Mean	2291.42	728.71	3992.51	1862.81	2782.72	1513.19
	Total edits	16040	5101	123768	50296	512021	266323
	Days_mapped	7	7	31	27	184	176
Zakarpattia Oblast	Mean	2684.85	573.57	1214.03	559.96	2987.41	1105.72
	Total edits	18794	4015	37635	16799	549685	202347
	Days_mapped	7	7	31	30	184	183

By comparing the data from below table for 1 week, it is observed that while the mapping activities increased for Kyiv and Kherson regions, for Kharkiv and Dontesk sites the numbers drastically decreased.

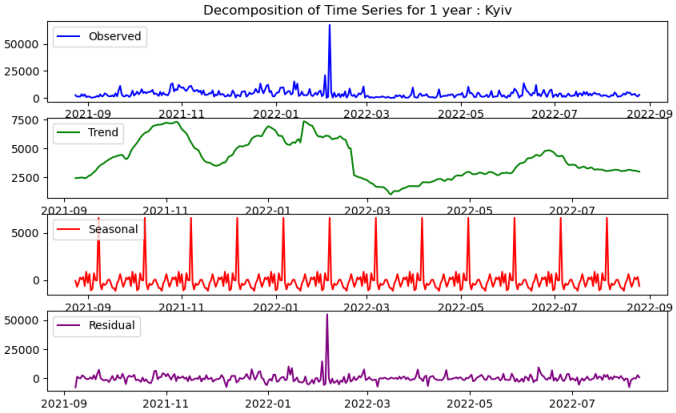
One-Month Period: During the one-month period before and after the event, all four sites experienced a major decrease in contributions, with each site witnessing a substantial decline of at least 60% in mapping activities compared to the previous month. This sharp drop in mapping efforts indicates a significant disruption in mapping activities during this period, likely as a direct consequence of the event, and reflects the impact of the conflict beginning to be experienced in this timeframe.

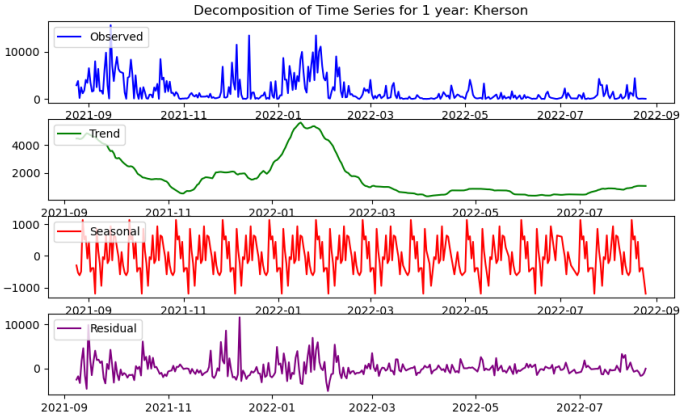
Six-Month Periods: When assessing mapping activities over a longer six-month timeframe, it becomes apparent that the decline in contributions was not limited to the immediate aftermath of the event but also due to the impact of recurrent destruction from ongoing attacks. it is noticed that new contributions at each location were only 50-60% of total contributions made during the previous timeframe. This suggests a sustained reduction in mapping efforts over long-term.

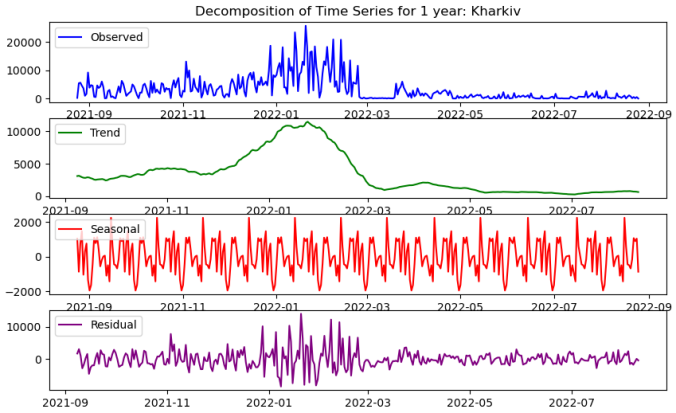
Kharkiv, despite initially having a high number of contributions, has seen a sharp and continuous decline in mapping activities across all three time periods. This decline is particularly huge in the post-bombing period, highlighting the lasting impact of the conflict on this region and reported only 14% new mapping activities post-event as compared to the value of previous six months.

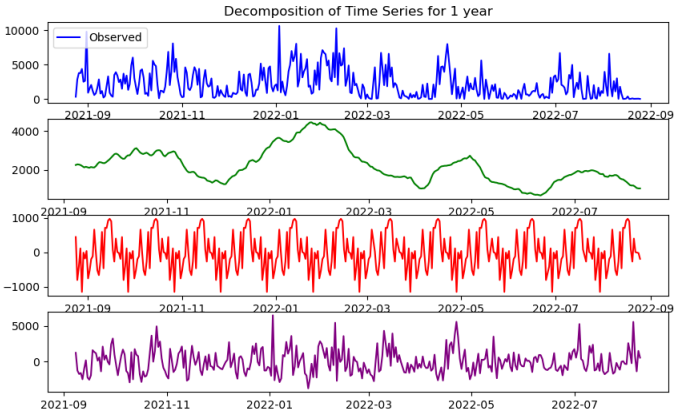
Days Mapped: The absence of mapping for several days in Kherson, Kharkiv, and Donetsk further signifies the disruption caused by the conflict. This period of inactivity could be linked to various factors, including safety concerns, infrastructure damage, or displacement of contributors.

It is important to note that at control site- Zakarpattia oblast also had decreasing contributions during all three time periods even when there was no active bombing reported at this location. These results draw a pattern of overall decrease in mapping activities across the Ukraine and shows the substantial impact of ongoing conflict not just as immediate effect on targeted regions but its far-reaching consequences at regions not affected.
Further, we analyse the mapping trend from below time-series decomposition taken for one year timeframe by combining 6 months pre-bombing and post-bombing data.









Trend: The trend component for Kyiv non-uniform cyclicity with downwards trend in activities around June month. Similarly, Donetsk also observes non-uniform cyclicity with decreasing trend after July month. All four sites has observed a spike during the timeline of January-February month however, Kherson and Kharkiv only had downward trend in activities post March month.
Seasonality: The seasonality component for all sites has unchanged monthly periodicity with magnitude being constant.

From the trend analysis we uncover sudden changes in mapping activities which can be linked to the beginning of war during the same time.

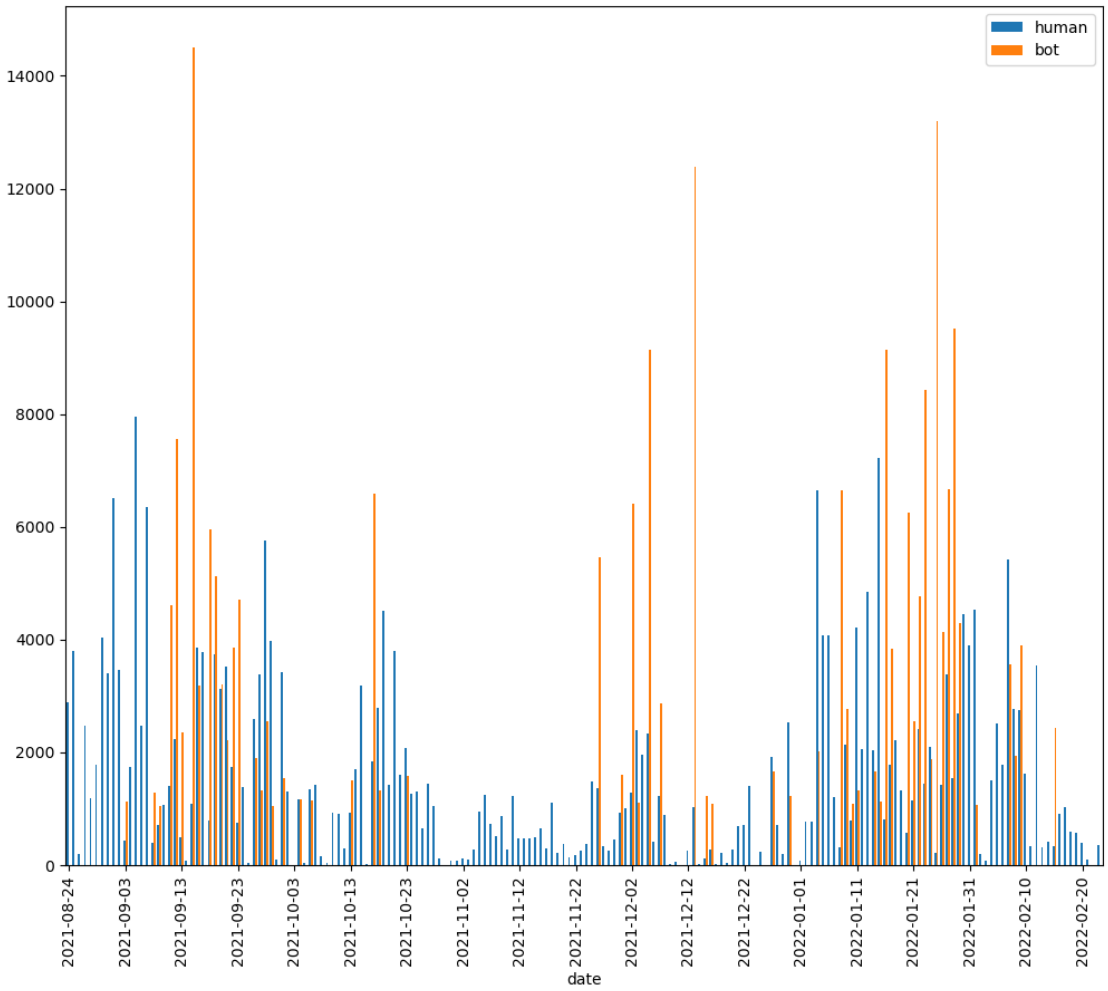
Further, we summarize the number of contributions posted manually and by automated scripts with below table.

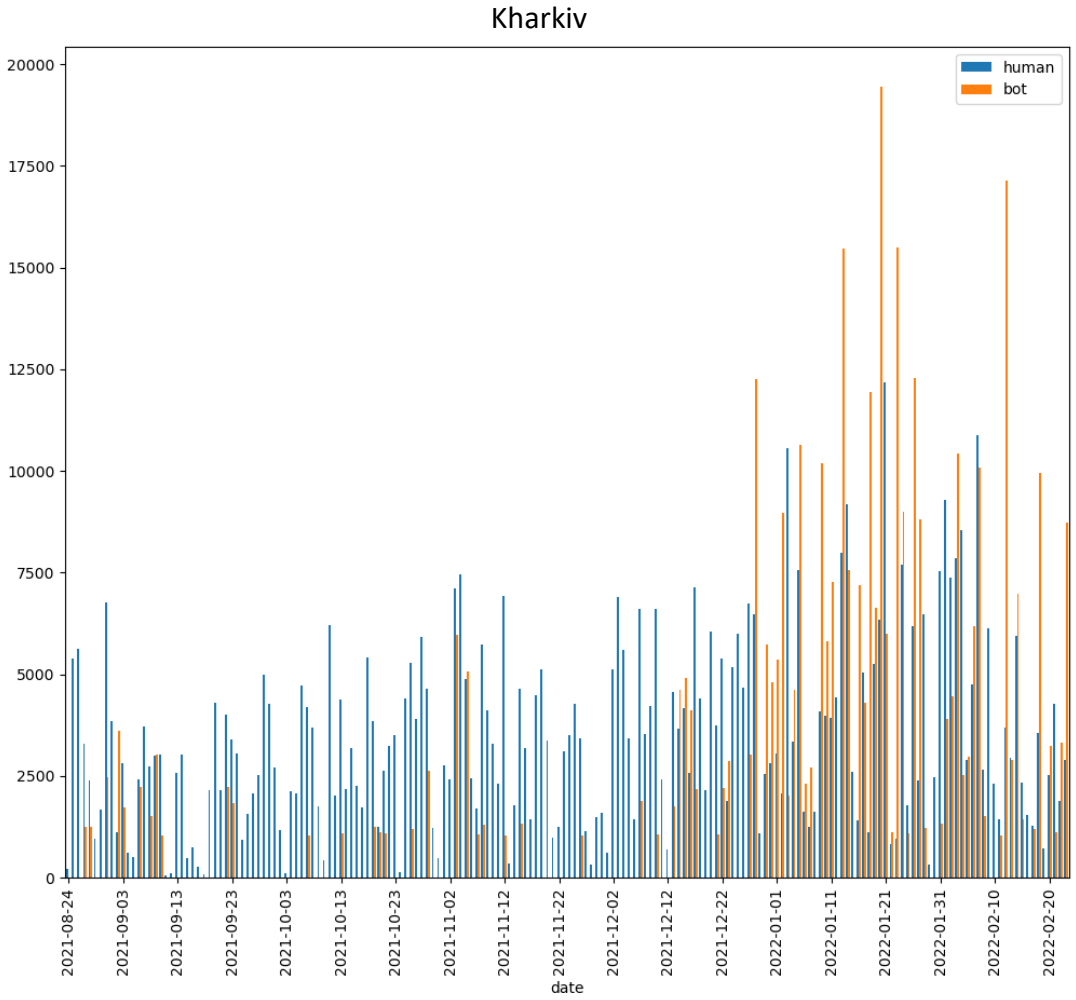
	Before bombing		After bombing	
Region	**Human**	**Automated**	**Human**	**Automated**
Kyiv Oblast	12220	80	7204	65
Kherson Oblast	2981	120	1129	13
Kharkiv Oblast	5499	179	1994	9

Donetsk Oblast	5904	38	2603	35
Zakarpattia Oblast	3799	131	5414	21

Charts: The number shows that in Kherson and Kharkiv, relatively high number of changesets were added by automated scripts in pre-bombing period but only few contributions were done in the latter 6 months. For Kyiv and Donetsk, the numbers for automated contributions are comparatively unchanged, with higher number of changesets contributed manually. This can be used to imply that higher volume of data before the bombing at these sites were majorly due to automated contributions.
The bar charts showing daily total contributions categorised by human and bot shows high spikes for 'bot' category which can be used to imply the same. These charts also explain the increasing trend during January-February.

Kherson





War impact on Contributor counts:

The table enumerates the contributors who participated before and after the crisis, for each study location. Results show that Kyiv had the highest number of active contributors before the commencement of attacks showcasing the widespread community of mappers in this location, but the number of mappers reduced significantly after the crisis.
Similarly, Kharkiv also lost notable number of contributors post event. Other sites had comparatively small number of active participants and was not affected intensely.

Furthermore, comparing committed mappers with total mappers in each case shows that these users constitute only 1-2%. This result is in line with the concept of 90-9-1 which is generally used in describing participation level of online contributors and states that only 1% of total users are highly engaged and most active members of the community. The result implies that the community is vulnerable to external events such as one studied here highlights the need to support this community. Evidently, we note that a small but dedicated group of contributors plays a crucial role in maintaining and advancing mapping efforts.

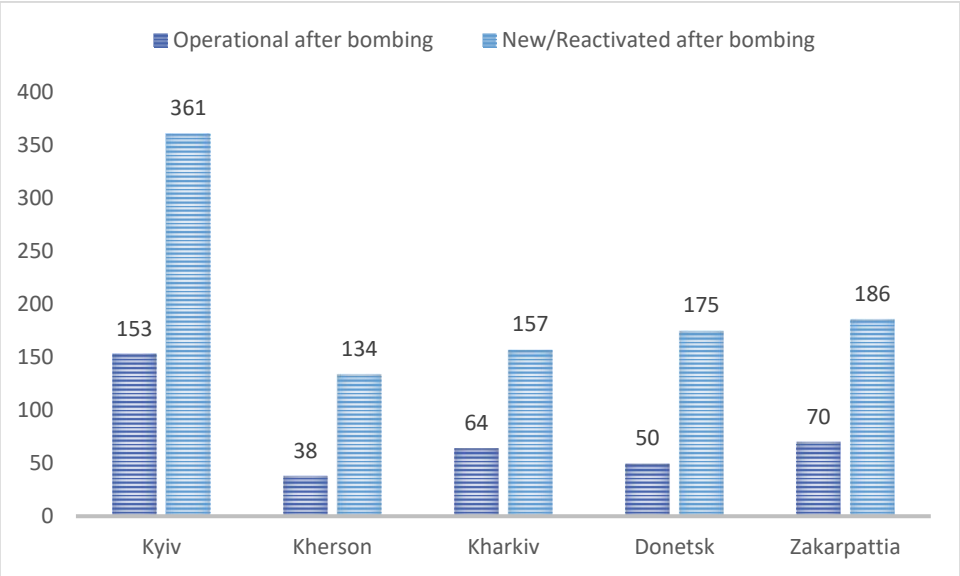
	Kyiv		Kherson		Kharkiv		Donetsk		Zakarpattia	
Contributor type	Before	After	Before	After	Before	After	Before	after	Before	after
Committed mapper	16	14	5	3	8	3	8	5	10	8
Casual mapper	32	20	10	9	12	10	12	7	6	8
Inactive mapper	693	480	180	160	334	208	231	213	272	240
Total mappers	741	514	195	172	354	221	251	225	288	252

Contributor addition/reactivation

From the chart describes the number of users who were operational before the attacks and are still active as well as the number of users who began mapping for the first time after the series of events. These users may be entirely new users or may have been reactivated due to the impact of ongoing war.

The decline in existing contributors still operation post attacks can be linked to number of factors such as displacement, change in priorities, lack of resources or may have been directly affected by the damage.

Comparison of values shows that while the total number of mappers is reduced post-event, the number of new/reactivated users are higher than the number of existing mappers still active during this period. This could be due to a number of reasons, such as an increased awareness of the importance of mapping during times of crisis, a desire to contribute to humanitarian efforts, or a need for accurate mapping data for emergency response.
This shows a positive impact of ongoing crisis on contributor activation.



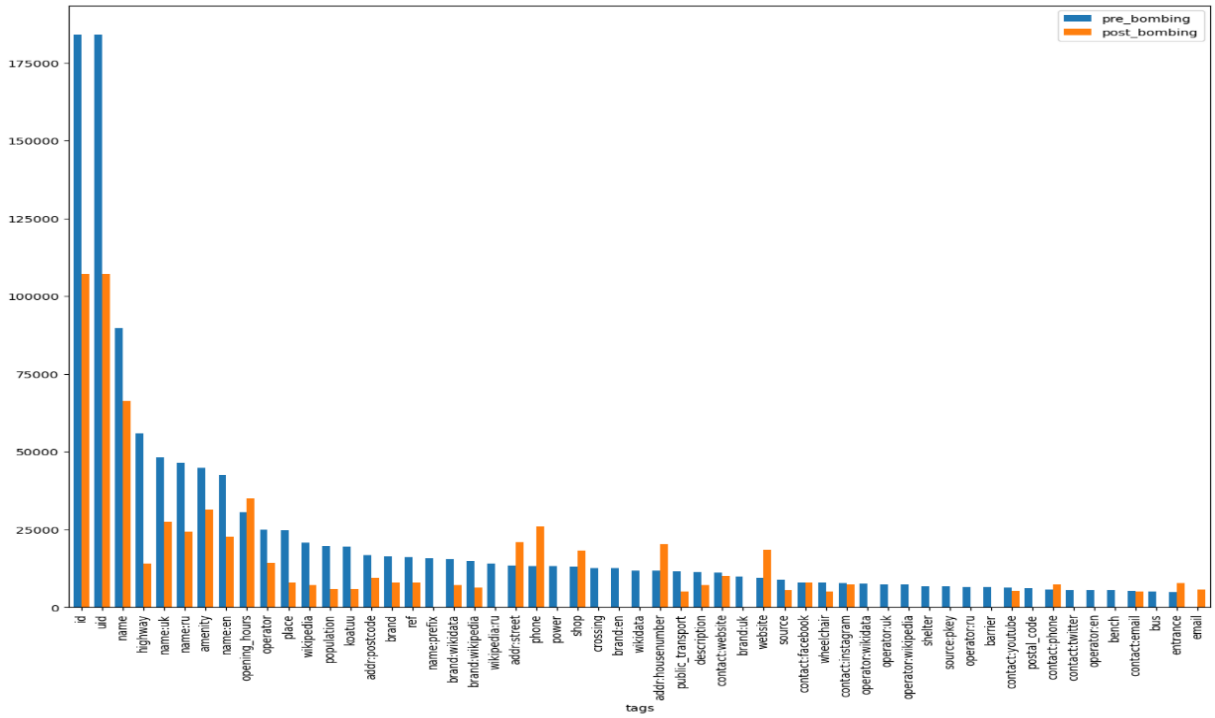
(2) What users have mapped during this period?

Results obtained for tag analysis can be summarized using below comparison charts which show the frequency of tags meeting the threshold value(> 5000 counts) for both 6-months intervals.

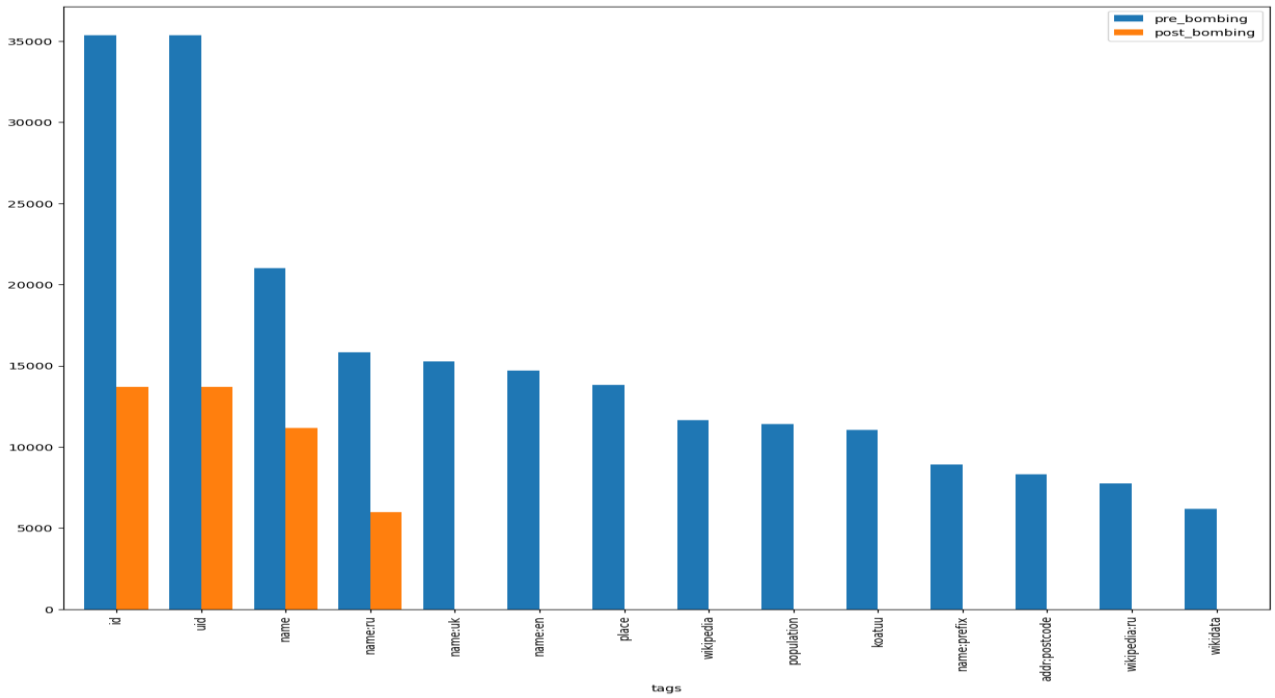
Although the, the number of distinct tags added by contributors in Kyiv were lesser post-bombing, overall,there were more unique tags added and shows more variability. Kharkiv and Donetsk had similar pattern in terms of tag addition. Tags contribution for Kherson is reportedly very less in number and variability. in fact, only two tags had more than threshold count for Kherson. Some of the most frequently added by the contributors are 'highway, amenity, name, place. These tags capture very useful information that can help during disaster assistance.

Evaluating post-bombing data shows that contributors started mapping unique tags like 'address', 'opening_hours', 'contact details' , 'phone number' , 'website' and so on for the locations they are adding. This reflects on community understanding of the importance of these tags for easier access to information and seek assistance during recovery efforts.

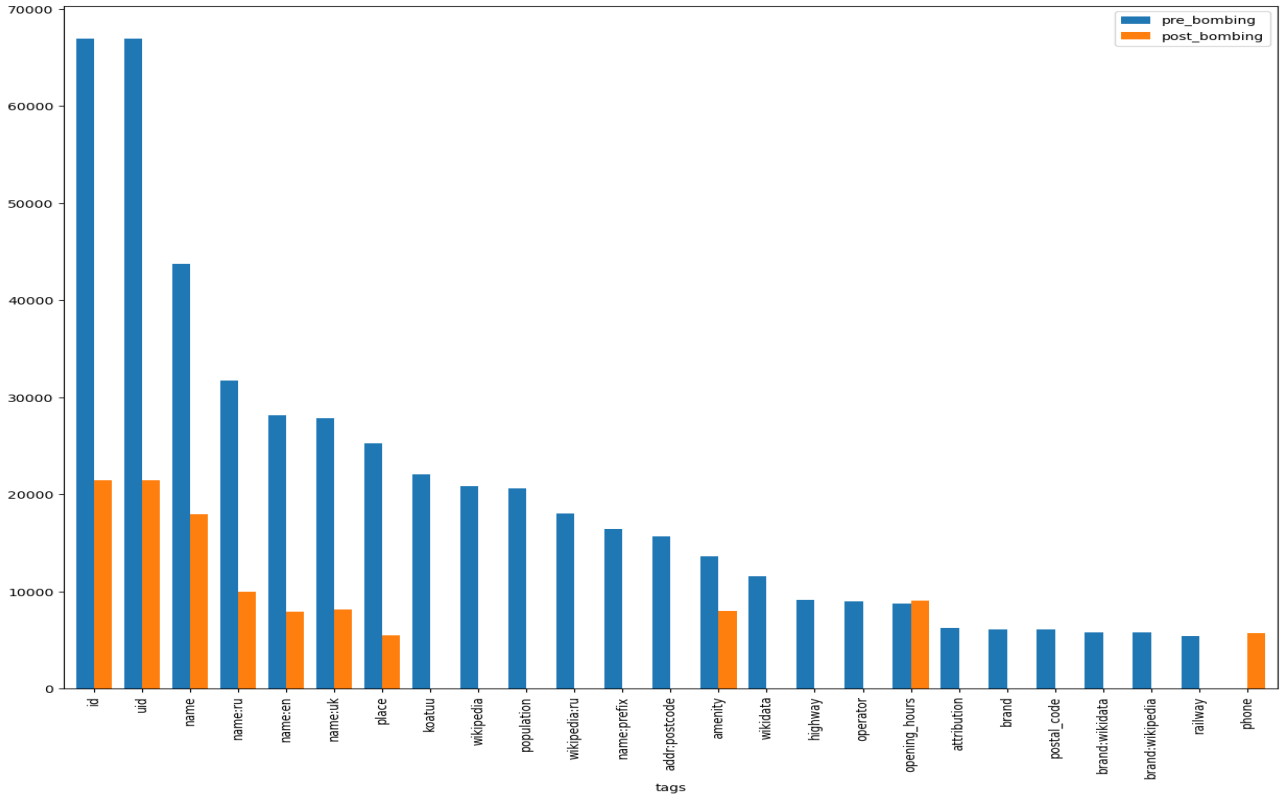
KYIV



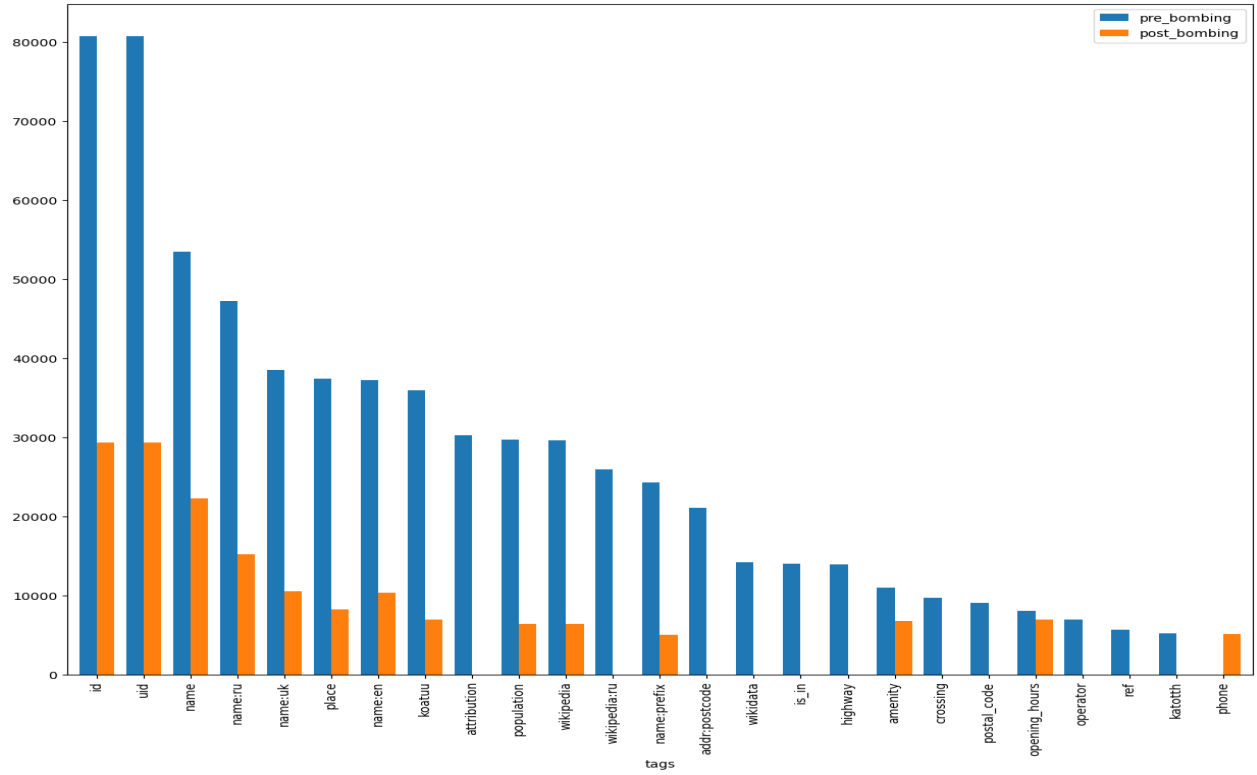
KHERSON



KHARKIV



DONTESK



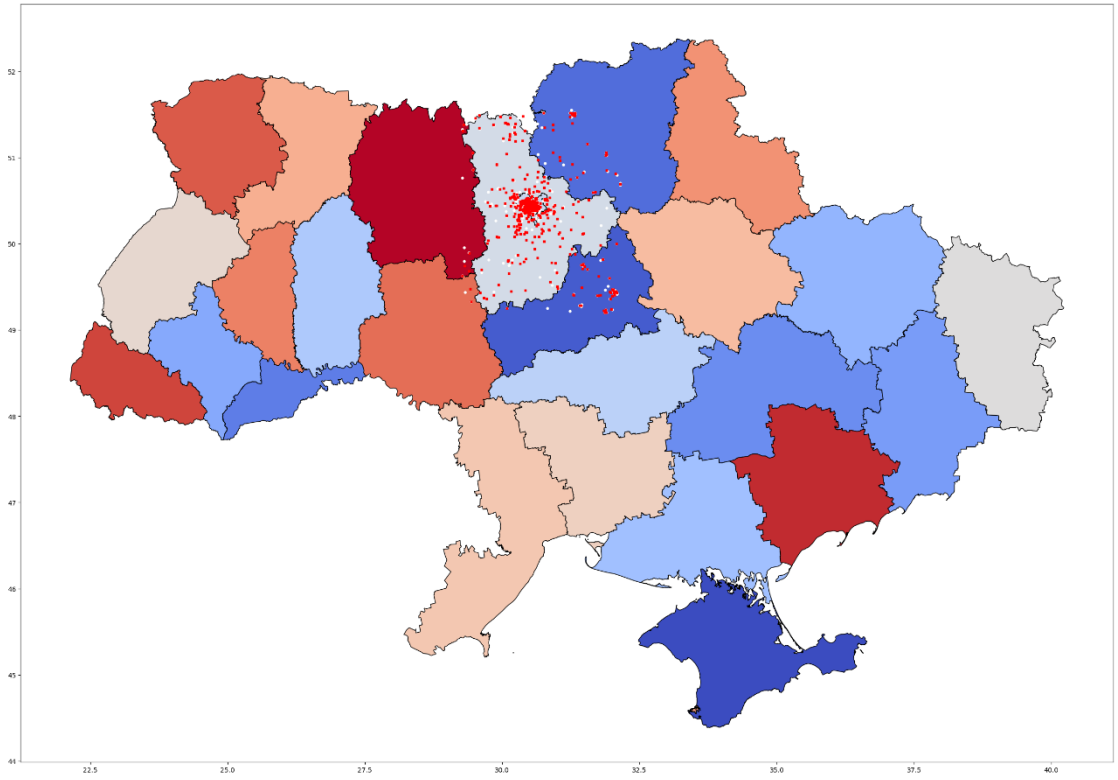
(3) Where did the users contribute?

Spatial representation of mapping behaviour is pictured in following plots created with location values. Observing the mapping behaviour of contributors across four different locations reveals different nature of contributor groups.

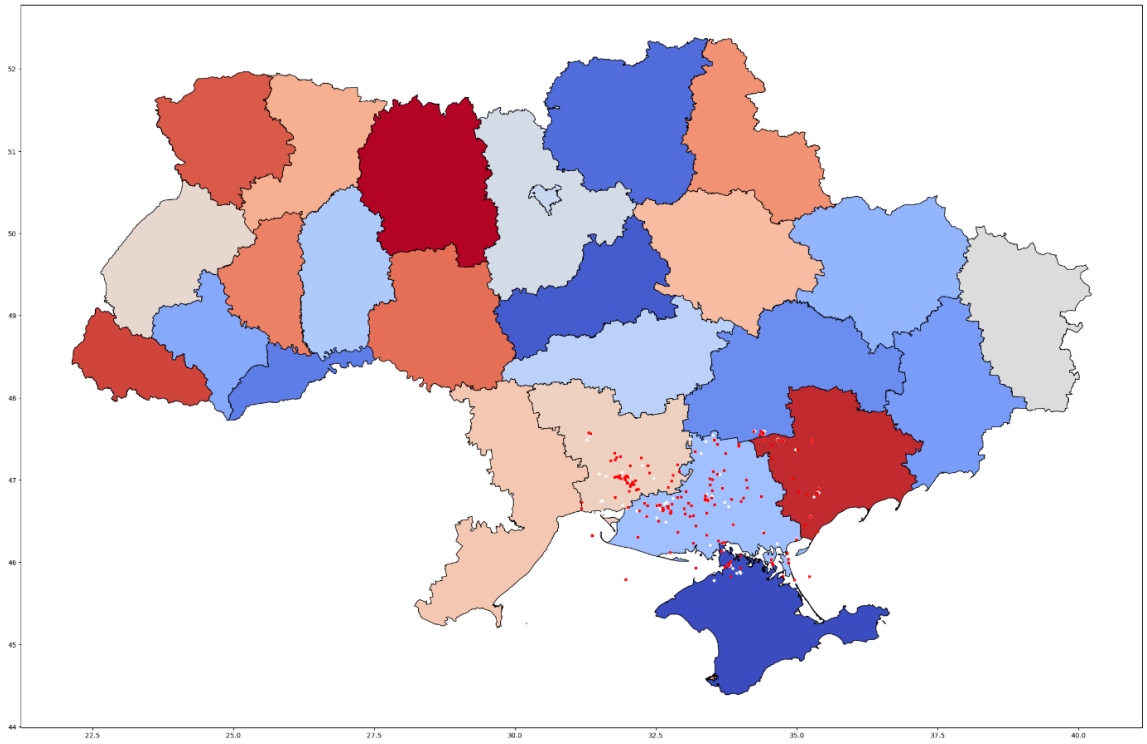
The mapping behavior in Kyiv is characterized by a more clustered pattern around the city center. This suggests that contributors in Kyiv exhibit a focused mapping behavior.
This clustering can be attributed to the greater number of active users in Kyiv.
Similarly, Donetsk also displays focused mapping, although it had comparatively less number of active contributors after the event.
In contrast, the graphs for Kherson and Kharkiv are highly scattered in nature. i

It is also observable from the maps that the location of first node created and location of mean point within each user set are in close proximity, giving similar patterns as seen on map. From this it can be assumed that majority of users will contribute around the location at which they begin mapping.

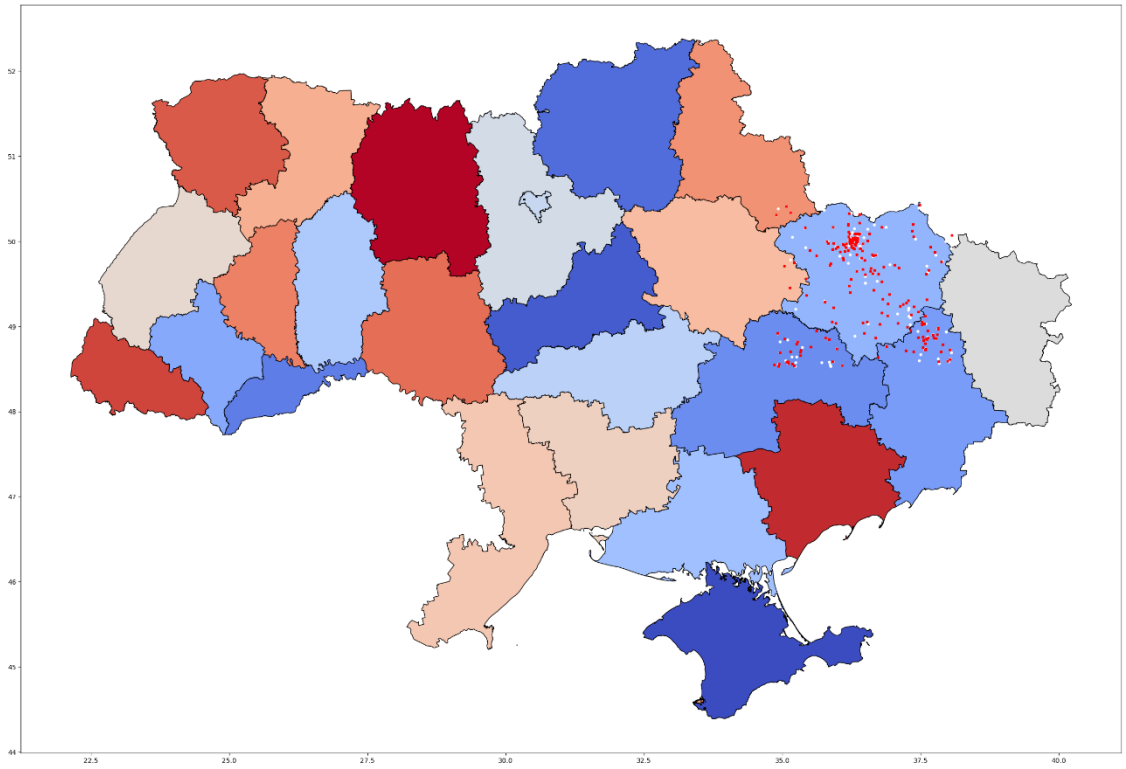
KYIV



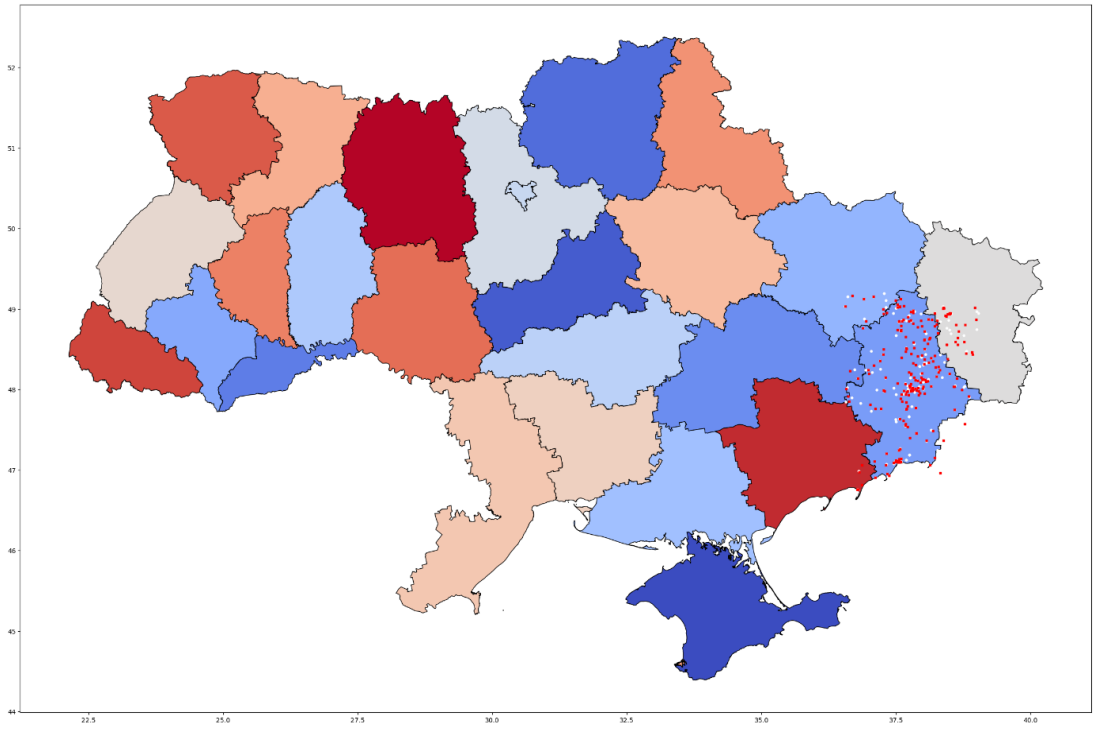
KHERSON



KHARKIV



DONETSK



Degree of dispersion:

The table here compares the Coefficient of Variance (CV) calculated for population data.
By comparing the CV value at each site, the common observation is that the variability increased after the war began. This implies that mapper was more concentrated and followed orderly mapping practice however, post -bombing they started contributing more randomly.

Subsequently, Kherson had more variability in map as compared to other sites. The similar pattern for dispersion is also seen in the previous graphs depicting mapping behaviour of sample data.

	Pre-bombing		Post -bombing	
Location	**CV_Latitude**	**CV_Longitude**	**CV_Latitude**	**CV_Longitude**
Kyiv	1.0989	2.4168	1.07113	2.6365
Kherson	0.5170	2.1913	1.144	3.3883
Kharkiv	0.9847	2.3450	1.1006	2.3279
Donetsk	1.2740	1.2151	1.0862	1.7538
Zakarpattia	0.6123	3.094	0.5929	3.3840

Evaluation for RQ2

Using the methodology explained in RQ2, the mapping activities of user actively adding new data during the conflict is captured as quantitative measure. Below resulting profile is created by combining the values of each measure. The features extracted here express multiple dimensions of contributor characteristics in numbered and categorical format.

	uid	id	contribution level	version	new nodes	edited nodes	changeset	active days	name	opening_hours	amenity	name:uk	phone	name:ru	name:en	addr:street	addr:housenumber	website	shop	operator	highway
0	5359	4	inactive	[6, 10, 2, 2]	0	4	4	149 days	2.0	3.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0
1	6389	3	inactive	[18, 11, 13]	0	3	3	32 days	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	11238	14280	committed	[5, 4, 3, 4, 3, 3]	12096	2184	42	96 days	13.0	1.0	5.0	8.0	0.0	5.0	2.0	3.0	22.0	0.0	1.0	1.0	6.0
3	12054	458	inactive	[2, 2, 2, 2, 3, 4]	284	174	3	1 days	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

This analysis gives a structured way of deriving important contributor characteristics that can be used for further analysis.

Challenges and Limitations

1. Data Volume Challenge:

The primary challenge in this analysis stems from the high volume of data. The study relies on historical data from Ukraine, which consists of 170 million records. Dealing with data of this

magnitude demands efficient processing and substantial memory resources. To address this challenge effectively within resource limitations, a strategic approach is taken. The data is stored in a PostgreSQL server, and SQL queries are used for data retrieval and manipulation. This strategy minimizes memory usage and enhances processing efficiency. Furthermore, the raw data is initially distributed in small, compressed files therefore the initial process of data consolidation is an important step in data handling. Data consolidation and data import into PostgreSQL server is successfully implemented by converting the compressed file into csv which more readable format and allows direct import into server using supported commands. This approach allows for the seamless handling of vast amounts of data.

2. Data Structuring Challenge (Tag Data):

Another significant challenge arises when dealing with tag data. Structuring this data into a readable and accessible format is imperative for meaningful analysis. The tags are stored in dictionaries, which necessitates a critical step in the data processing pipeline: converting these dictionaries into a tabular format. Additionally, the tags are not uniform and each tag dictionary is differently structured. This further adds complexity to the conversion method while ensuring that each unique key is captured.

This analysis is conducted under some limitations as discussed here.

1. Geographical Focus:

The analysis is conducted with a specific focus on evaluating the impact of war on locations at the administrative division level-1 scale, as defined for Ukraine. This means that the assessment is carried out on a broader geographical scale, covering larger regions rather than individual cities or specific bombing sites. While this approach allows for a high-level overview of the impact, it may not capture the nuances or localized effects that can be observed by selecting individual sites.

2. Limited Timeframe Consideration:

The analysis is limited to six months, for evaluating the initial impact of the attacks. This timeframe immediate consequences of the conflict. However, for a more comprehensive understanding of the war's effects, a longer timeframe could be considered. Extending the analysis over a more extended period can reveal shifting patterns of mappers which may not have been captured.

3. Uniform Time Intervals:

To facilitate effective comparison between different study areas, the analysis takes similar start-points and endpoints for each time interval: 1 week, 1 month, and 6 months. This standardization helps in monitoring changes consistently across various locations. However, t this approach does not differentiate between the cumulative impact of each individual attack at the given location.

4. ‘Node' Type OSM Elements:

The scope of this study is limited to 'node' type OpenStreetMap (OSM) elements. Due to time constraints and the complex structure of the other OSM element types- ‘Ways’ and ‘Relations’ are excluded from the study.

5. Frequency of bombing and magnitude of damage

Since each study area may have experienced different magnitude of bombing at different time intervals, it may have varied degree of impact on gaining contributor attention. These elements are not covered while assessing the overall impact.

Discussion and Future Scope

OSM is a dynamic map evolving along with its community. Studying the impact of human -led crisis on OSM community, using the recent events of Russia-Ukraine conflicts, this study provides a holistic approach to understanding OSM community and its collective response. Understanding this community and their intrinsic behavioural measures during war has significant implication for future.

By comparing mapping activities before and after the onset of bombing incidents, the study seeks to shed light on how these adverse events impact the efforts of contributors who play a crucial role in creating and updating mapping data. This research is valuable not only for understanding the consequences of conflict on digital mapping but also for gaining insights into the resilience and adaptability of contributor communities in the face of adversity.

The quantitative measures of mapping patterns described in this paper assesses temporal and spatial dynamics of map. The methodology presented in this analysis for feature extraction from user’s mapping history builds an important characteristic profile of community of active users.
The analysis of tag annotation gives some insight on local area knowledge of the contributor, for example tags associated with highway,place, name:ru, shows that users have knowledge on local languages and is located in the close proximity of the location since relocating to bombing site is not intuitive.

In comparison with similar studies, the contradicting result suggesting negative impact of war on overall contribution of OSM community strikes a challenging question of how to make the community more engaged in future in case of similar events. This study can be a guideline to large-scale data production events led by OSM and VGI projects. It can inform best practices for better map production and supporting contributor communities during crises.

This study can be further extended to explore other elements of OSM data and identify how they contribute to the map. To uncover the possible impact of each subsequent attacks and its impact, a more granular approach is to scale down on the spatial location and study each individual sites, this approach enables understanding what immediate effects have been observed.

Since relocating to affected areas is not feasible during such crisis, studying the likelihood of remote mapping can is a relevant aspect that can be inferred by gathering user location details, which is not in scope of this analysis due to unavailability of location data. Such information not only differentiates between the locality of the map producers but also shows the possibility and magnitude of remote mapping activities in wake of wartime.

Leveraging temporal changes in mapping data for time-series forecasting is a forward-looking approach. Using the historical trends, this analysis can provide predictions on future data production. Such forecasts can be valuable for planning and decision-making, especially in crisis or post-crisis scenarios.
Furthermore, applying machine learning algorithms such as clustering and classification algorithms to the contributor features extracted in this analysis can group the contributors in more related groups.

Conclusion:

This paper studies a new approach for analysing the community response to crisis and uncovering the underlying community characteristics. The findings can benefit various stakeholders associated with OpenStreetMap such as OSM members, Humanitarian and relief organizations and Researchers and Academic institutes, Mapping event organizers.

References:

[1] Harley, J. B., and David Woodward, eds. *Cartography in Prehistoric, Ancient, and Medieval Europe and the Mediterranean.* Vol. 1 of *The History of Cartography*. Chicago: University of Chicago Press, 1987.

[2] Meng , L. 2009 . “ Affordance and reflex level of geovisualizattion ” . In Virtual geographic environments”A primer” , Edited by: Hui , L. and Batty , M. 136 – 150 . Beijing, , China : Science Press .

[3] Goodchild, M.F. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221 (2007). https://doi.org/10.1007/s10708-007-9111-y

[4] Bennett, J., 2010. OpenStreetMap. Packt Publishing Ltd.

[5] J. E. Vargas-Munoz, S. Srivastava, D. Tuia and A. X. Falcão, "OpenStreetMap: Challenges and Opportunities in Machine Learning and Remote Sensing," in *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 1, pp. 184-199, March 2021, doi: 10.1109/MGRS.2020.2994107.

[6] C. C. Fonte, L. Bastin, L. See, G. Foody and F. Lupia, "Usability of VGI for validation of land cover maps", *Int. J. Geogr. Inform. Sci.*, vol. 29, no. 7, pp. 1269-1291, 2015.

[7] A. Schilling, M. Over, S. Neubauer, P. Neis, G. Walenciak and A. Zipf, "Interoperable location based services for 3D cities on the web using user generated content from OpenStreetMap", *Proc. Urban Regional Data Management: UDMS Annu.*, pp. 75-84, 2009.

[8] Muhamad, N., Muhamad Hakimi, F. H. and Mahadzir, M. D. A. (2023) “Community Activation for Disaster Risk Reduction Through OpenStreetMap: A Scoping Review”, *Journal of Advanced Geospatial Science & Technology*, 3(2), pp. 25–50. doi: 10.11113/jagst.v3n2.68.

[9] Barron, C.; Neis, P.; Zipf, A. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. Trans. GIS 2013, 18, 877–895.

[10] Fan H, Zipf A, Fu Q (2014) Estimation of building types on Open- StreetMap based on urban morphology analysis. In: Connecting a Digital Europe Through Location and Place. Springer,pp 19–35

[11] Forati AM, Ghose R (2020) Volunteered Geographic Information Users Contributions Pattern and its Impact on Information Quality

[12] Pavlovskaya, M. E., & Martin, K. S. (2007). Feminism and geographic information systems: From a missing object to a mapping subject. Geography Compass, 1(3), 583–606.

[13] Herfort, Benjamin, Sven Lautenbach, João Porto de Albuquerque, Jennings Anderson and Alexander Zipf. “The evolution of humanitarian mapping within the OpenStreetMap community.” *Scientific Reports* 11 (2021): n. pag.

[14] Tessio Novack, Leonard Vorbeck & Alexander Zipf (2022): An investigation of the temporality of OpenStreetMap data contribution activities, Geo-spatial Information Science, DOI: 10.1080/10095020.2022.2124127

[15] Muttaqien, Bani & Ostermann, Frank & Lemmens, Rob. (2018). Modeling aggregated expertise of user contributions to assess the credibility of OpenStreetMap features. Transactions in GIS. 22. 823-841. 10.1111/tgis.12454.

[16] Ahmed Ahmouda, Hartwig H. Hochmair & Sreten Cvetojevic (2018) Analyzing the effect of earthquakes on OpenStreetMap contribution patterns and tweeting activities, Geo-spatial Information Science, 21:3, 195-212, DOI: 10.1080/10095020.2018.1498666

[17] Kamptner, Erika & Kessler, Fritz. (2019). Small-scale crisis response mapping: comparing user contributions to events in OpenStreetMap. GeoJournal. 84. 10.1007/s10708-018-9912-1.

[18] Jacobs, Kent & Mitchell, Scott. (2020). OpenStreetMap quality assessment using unsupervised machine learning methods. Transactions in GIS. 10.1111/tgis.12680.

[19] Zhao, Yijiang & Wei, Xingcai & Liu, Yizhi & Liao, Zhuhua. (2021). An OSM Contributors Classification Method Based on WPCA and GMM. Journal of Physics: Conference Series. 2025. 012040. 10.1088/1742-6596/2025/1/012040.

[20] Zhiyao ZHAO,Hongchao FAN. Towards Exploring Patterns of Editing Behavior on OpenStreetMap[J]. Journal of Geodesy and Geoinformation Science, 2022, 5(2): 85-97.