DCluster: Geospatial Analytics with PoI Identification

Cláudio Gustavo S. Capanema[1], Fabrício A. Silva[2],
Thais R. M. Braga Silva[2], Antonio A. F. Loureiro[1]

[1] Universidade Federal de Minas Gerais, Brazil
{claudio.capanema, loureiro}@dcc.ufmg.br
[2] Universidade Federal de Viçosa, Brazil
{fabricio.asilva, thais.braga}@ufv.br

Abstract. The generation of geospatial data is an inherent aspect for several applications that aim to track people, automobiles, or other mobile objects. Mining information from this type of data is a crucial factor for the development of Smart Cities. In many cases, it can help improve human mobility and the quality of citizens. In this sense, there is a growing demand for systems capable of extracting information from several data types, including the geospatial one. In this work, we present DCluster, a web system that aims to assist data analysts in exploring and visualizing the main types of data, including the geospatial one. Additionally, DCluster has the capability of discovering points of interest based on data of mobile users and classifying them as *Home*, *Work*, and *Other* locations. Data analysts can take advantage of DCluster to explore their data and extract knowledge from it.

Categories and Subject Descriptors: H.2 [**Data Mining**]: Miscellaneous

Keywords: Points of Interest, Clustering, Geospatial Data, Data Analysis

1. INTRODUCTION

The use of mobile devices, such as smartphones and tablets, is spreading rapidly, increasing the availability of large volumes of data. This is due to factors like the low cost of acquiring mobile devices and the interest of service providers in getting to know their customers better. According to [Statista 2017], in the first quarter of 2012, the number of daily active users on Facebook through a mobile platform was approximately 266 million. In the same period of 2016, this number rose to 1.146 billion.

In general, much of the data coming from mobile devices is georeferenced, i.e., it includes the user's location. One example is the possibility of performing location check-ins when interacting on social networks. In other cases, georeferencing is essential for the services' operation, as with the Waze and Uber applications. Finally, data from different business segments (e.g., banks, telecommunication operators, e-commerce systems, among others) also have georeferenced information used for various purposes, such as fraud detection and geofencing. In fact, location information brings many benefits to companies as well as to the research community.

However, raw location data has to be correctly processed so useful information can be extracted. The growing demand for data analysis has brought many analytical tools to the market. However, currently, there is still a lack in the treatment of georeferenced attributes, which either is not included or comes up against limitations, such as the license price. Also, some of these tools require the user to have technical expertise in data analysis. This makes it difficult for people with non-technical knowledge to use advanced resources.

In this work, we aim to fill this gap and propose DCluster, a system for EDA (Exploratory Data Analysis) of different types of data, including the georeferenced one. Because it is a Web tool, its use is not restricted to users with powerful personal computers. In addition, a simple and intuitive interface is designed, which is characterized by the absence of unnecessary information, so that the user experience runs as smoothly as possible. These aspects aim to contribute to the dissemination and popularization of the analysis of georeferenced data. The proposed system is an extension of the one originally presented in [Capanema et al. 2017], and the main contributions are summarized as follows:

—Support for different types of attributes: numeric, categorical, datetime, and coordinate. All of them are automatically identified by the system.
—Statistics and visualizations: the system provides statistics and graphs for each attribute. Additionally, it provides pairwise attribute analysis.
—Clustering: it is possible to run a clustering algorithm to group location points based on their proximity.
—PoI identification: an algorithm for points of interest (PoI) identification is available. It is designed for sparse location data and is capable of determining the central point of a relevant location and its type (*Home*, *Work*, or *Other*).
—Export data: the filtered dataset and the identified points of interest can be exported using the well-known file format *Comma Separated Values* (".csv").

The remainder of this text is organized as follows. In Section 2, the main related works are presented. In Section 3, the DCluster system is described. From Section 3.1 to 3.9, we present the system's features such as data input, georeferencing support, data filters, graphs and statistics available, pairwise association of data attributes, clustering, PoI identification, and data export, respectively. Finally, the conclusions and future works are presented in Section 4.

2. RELATED WORK

In this section, we describe the characteristics of the main tools for exploratory data analysis and compare them with DCluster. Table I summarizes the related systems according to support for georeferenced data, license, and platform (desktop or web).

One of the most well-known systems is Weka [Hall et al. 2009], a free tool with a simple interface. However, its development started in the 90s and was interrupted with its acquisition by Pentaho in 2006, which does not favor an ideal environment for the analysis of large volumes of data, since its architecture is centralized and does not support georeferenced data.

The Geo-Data Visualizer [Xavier et al. 2017] is a free tool developed in *Javascript / Jquery* that uses georeferenced data to generate maps and associated statistics. Its main features are the visualization of maps with clusters, and the filtering of data based on temporal metrics.

The Lemonade [dos Santos et al. 2017] is a web system that allows users to construct an execution flow by dragging and dropping functional components. Among its functionalities are algorithms of clustering, classification, cross-validation, and regression, graphs for visualization of the results, and other features. However, it has limited support for georeferenced data as it was not originally designed for this purpose.

The APPEL [Coimbra et al. 2019] is an extension of Kepler[1] that is useful to correlate coordinate points to census data of the region that the data belongs to. It can also perform fast spatial operations

[1]https://kepler.gl/ Accessed on 18/3/2021.

Table I. Comparison between systems. The license free* indicates limited free access.

	Parameters		
Systems	Georeferencing	License	Platform
APPEL	yes	free	web
Azure Machine Learning	yes	paid	web
BigML	no	free*	web
Geo-Data Visualizer	yes	free	web
Lemonade	yes	free	web
Pentaho Big Data	yes	paid	web and desktop
Power BI	yes	paid	web and desktop
Qlik	yes	paid	web and desktop
SAS	yes	paid	web
Sisense	yes	paid	web and desktop
Tableau	yes	paid	web and desktop
Weka	no	free	desktop

over points and polygons. Both Lemonade and APPEL are free tools, but they are still limited in terms of algorithms that can explore location data.

The BigML [BigML 2011] and the Azure Machine Learning [Microsoft 2017a] are good alternatives for those who want to use machine learning resources. They have a friendly user interface, a variety set of features, and are available as web applications. However, both tools face a lack of support for georeferenced data.

The Pentaho Big Data [Hitachi 2017], Tableau [Tableau 2017], Microsoft Power BI [Microsoft 2017b], SAS Business Intelligence and Analytics [SAS 2017], Qlik [Qlik 2017] and Sisense Business Analytics Software [Sisense 2021] systems fit into today's most comprehensive set of paid data analysis tools: support multiple data sources, have online versions, intuitive user interfaces, and work with georeferenced data. These tools fall into the BI (Business Intelligence) category, and which requires more time for learning how the system works. Also, the price of licenses is often not viable for small and medium-sized businesses.

Compared with the related systems, DCluster stands out for integrating common types of data (e.g., numeric, categorical, and datetime) with the georeferenced one. In addition, it detects points of interest and annotates their semantics. Those features are included in a free system in favor of students and researchers.

3. DCLUSTER

DCluster is a web system designed for exploratory analysis of different types of data, with a focus on georeferencing. Its client-server architecture aims to facilitate the use of the tool since installation on a local machine is not necessary. DCluster also stands out for the use of interactive features for visualizing graphs, and for being developed in the Python language, which offers a list of APIs for analysis, data visualization, and machine learning. The system has a set of features for processing, viewing, and exporting data. Additionally, we make the system available through a Docker Image[2][3]. DCluster has also the option of using a demo dataset, in order to make it possible to explore its functionalities. The available dataset contains the following attributes: $\langle user\_id, latitude, longitude, datetime, state, speed \rangle$. It was artificially generated because user tracking data that contains fine-grained data (e.g., GPS coordinates from mobile devices) may face privacy concerns. However, the pattern of the synthetic data was generated based on a real-life and private user tracking dataset. Figure 1 illustrates the main components of DCluster, which are centered on

[2]https://nesped.caf.ufv.br/producao-cientifica/. Accessed on 18/3/2021.
[3]https://github.com/claudiocapanema/dcluster-docker. Accessed on 18/3/2021.

supporting georeferenced data. The specific features of the system, such as visualization of maps, PoI identification, and others, are detailed in the following sections.

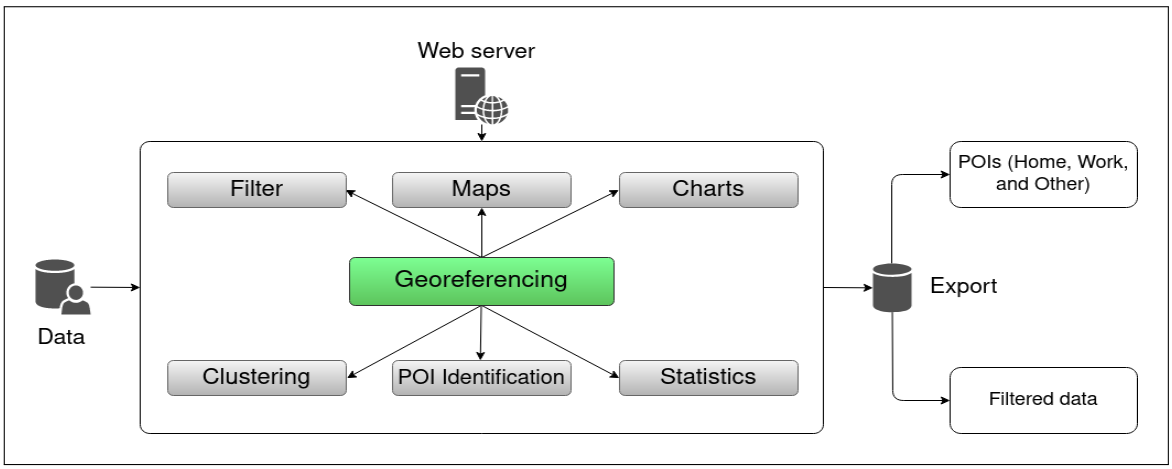

Fig. 1. Main components of DCluster.

3.1 Data input

The execution flow starts with the user sending data to be processed to the server, which can come from a CSV (Comma Separated Values) format, whether compressed (zip or tar.gz) or not, or by connecting to a MySQL or SQL Server database, which requires users to provide the query. Then, the system automatically identifies the types of each attribute (column) of the dataset, which can be: numeric, categorical, datetime, and spatial coordinate as shown in Figure 2.

3.2 Georeferencing

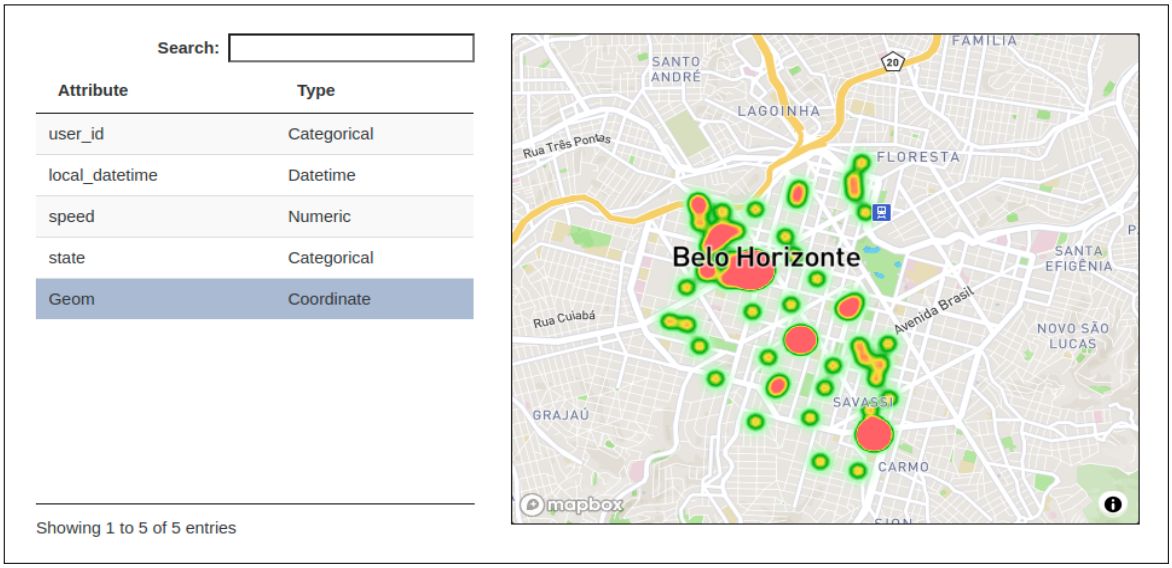

Fig. 2. Heatmap of coordinate attribute.

To handle georeferenced data, we create a new type called coordinate, composed of two attributes representing the latitude and the longitude. The system automatically identifies and creates a coordinate attribute by recognizing columns that have the following names: “latitude”, “lat”, “longitude” and “lng”. In case the dataset has different column names, the user must explicitly indicate which attributes have the latitude/longitude association, thus forming a composite attribute of the coordinate type.

Once the coordinate types have been defined, it is possible to visualize the data on a map (See Figure 2). In this regard, the central regions of the points are highlighted in red on the map. The support for georeferenced data is better explored in the next sections.

3.3 Filters

DCluster allows the user to filter the data, selecting items (rows) based on the values of the attributes (see Figure 3). Filters are flexible and can be edited after being created to reflect the user’s needs.

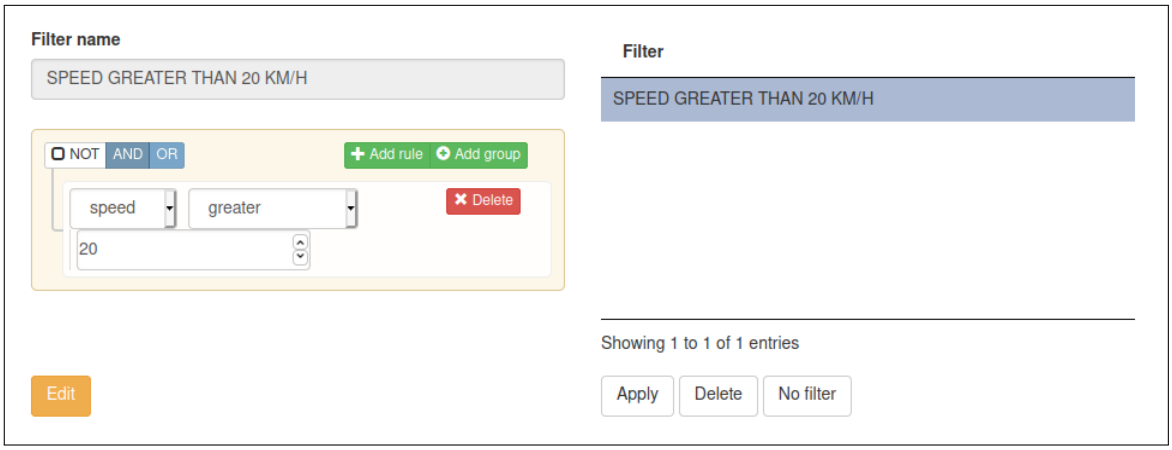

Fig. 3. Filter example.

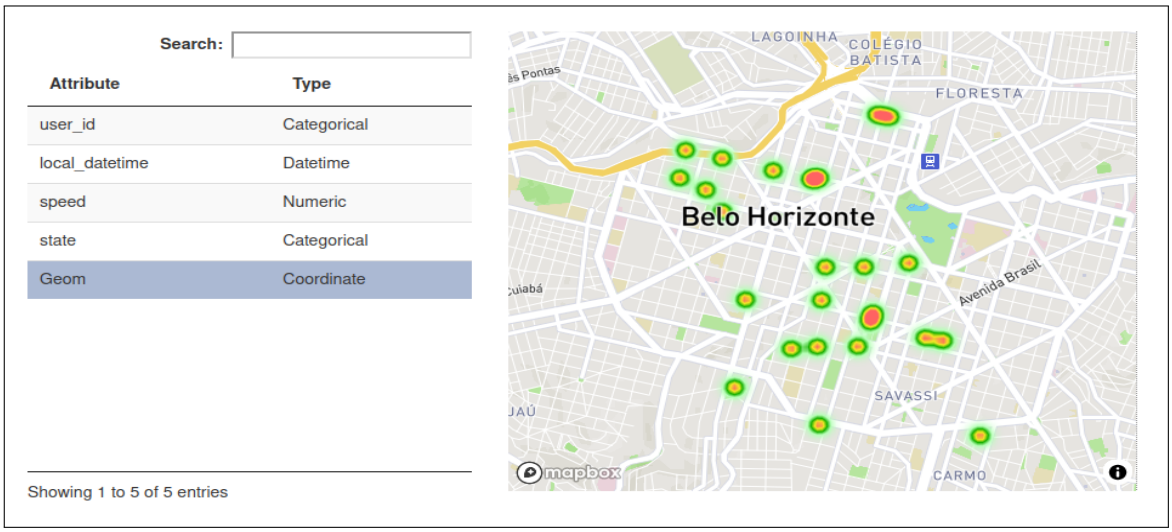

Fig. 4. Heatmap with filtered data.

A filter is defined by rule associations, which can be through the logical operations NOT, AND, or OR. Each rule corresponds to an attribute associated with an operation and a value, which can be numeric or categorical, depending on the type of the selected attribute. The rules can be nested, allowing complex logical expressions to be elaborated. The set of operations for numeric attributes are: different from, equal, greater, less, greater or equal, less or equal, range of values, and verification of nullity or non-nullity of items. For categorical attributes, the operations available are: equal, different, and checking for nullity or non-nullity items.

Figure 3 illustrates a filter that selects the items that have a speed value greater than 20. In Figure 4, the heatmap contains the filtered data, showing the regions where the users were moving faster than 20 km/h. The difference can be seen when comparing Figure 2 with Figure 4. The filtered data is used all over the system, but this can be reversed at any time in the filter section.

3.4 Graphs

DCluster has available different kinds of graphics for each of the accepted types of attributes: numeric, categorical, and datetime (See Figure 5). For numeric attributes, bar and line graphs are available. The datetime and categorical attributes can be visualized in bar graphs; the latter can also be visualized in pie graphs. The graphics have the advantage of being interactive, allowing the user to obtain information by hovering the cursor over certain regions. In addition, the graphics can be downloaded and edited in the graphics editing panel offered by the *Plotly* library [Inc. 2015].

3.5 Statistics

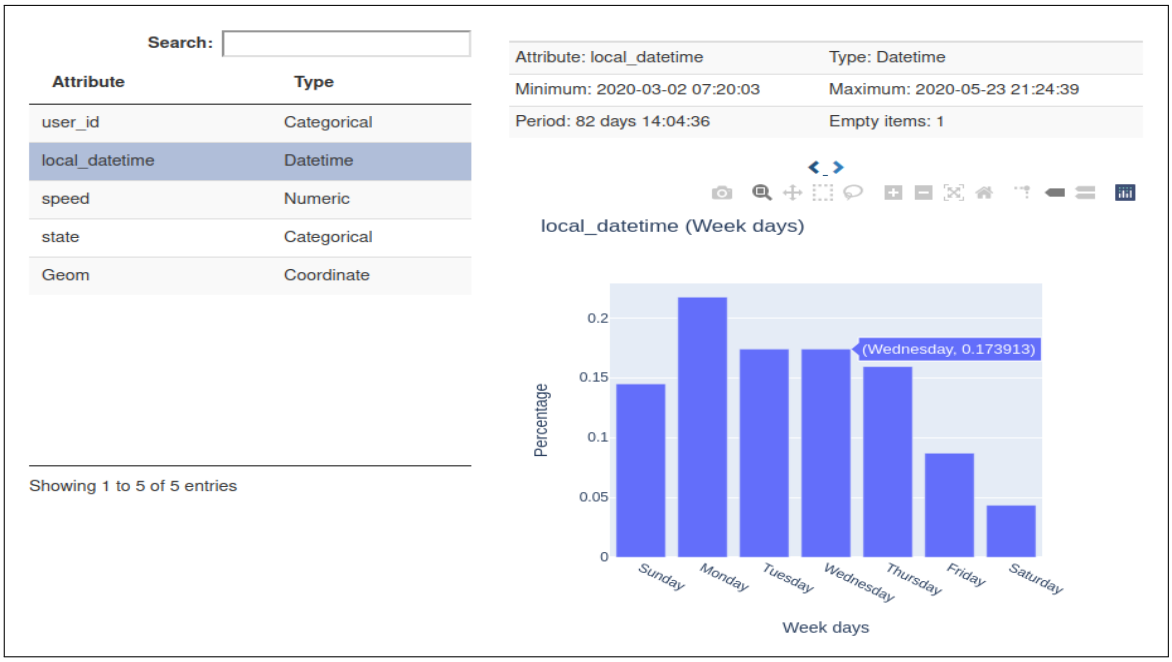

Fig. 5. Graph and statistics of datetime attribute. Each attribute has its respective graph and table of statistics.

In addition to the graphs, descriptive statistics are generated according to the type of each attribute in the dataset. For numeric attributes, average, variance, minimum, and maximum values are calculated. For datetime, the ranges of values, the minimum and the maximum, are displayed. In the case of categorical attributes, information on the number of different items is computed. For all types of

attributes, the number of items with the respective empty or null value is also informed. An example is illustrated in Figure 5 for the datetime type, where the respective graph and table of statistics are shown.

3.6 Pairwise Association

Table II. Associations between pairs of attributes and their respective graphs that can be generated by the tool.

	Y axis			
X axis	Numeric	Categorical	Datetime	Coordinate
Numeric	scatter	barplot e boxplot	barplot e boxplot	heatmap

The user has the option of selecting pairs of attributes and visualizing their association. The possible pairs of attributes and their respective visualizations are Numeric x Numeric, generating a scatterplot; Categorical x Numeric, generating bar and boxplot graphics; Datetime x Numeric, generating bar and boxplot graphics; Numeric x Coordinate, generating a heatmap. Table II summarizes the pairwise associations.

Fig. 6. Categorical and numeric association. In this case, the possible visualizations are bar and boxplots.

Figures 6 and 7 show two examples of associations between pairs of attributes based on the demo dataset, which is described in Section 3. The first figure contains a boxplot generated by the association of the attributes “state” and “speed”, which are categorical and numeric attributes, respectively. The second figure shows a map generated by the association between numeric and coordinate attributes. In this example, “geom” and “speed” are selected. The color of each circle varies according to the speed associated with each location point, that is, the faster the speed the redder the circle while the slower the speed the greener the circle.

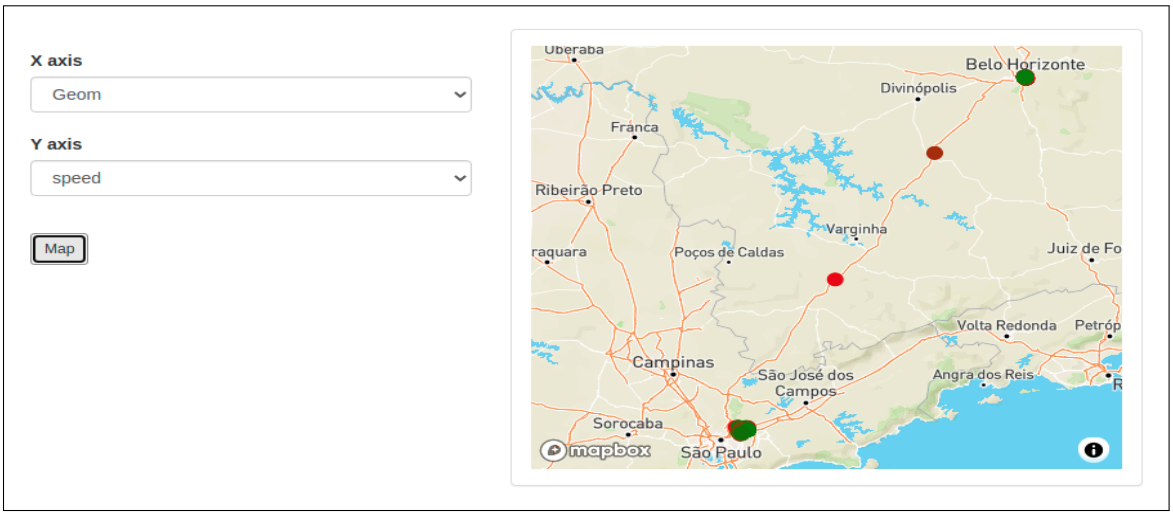

Fig. 7. Coordinate and numeric association. Each circle represents one data point and its color scale varies according to the selected numeric attribute.

3.7 Clustering

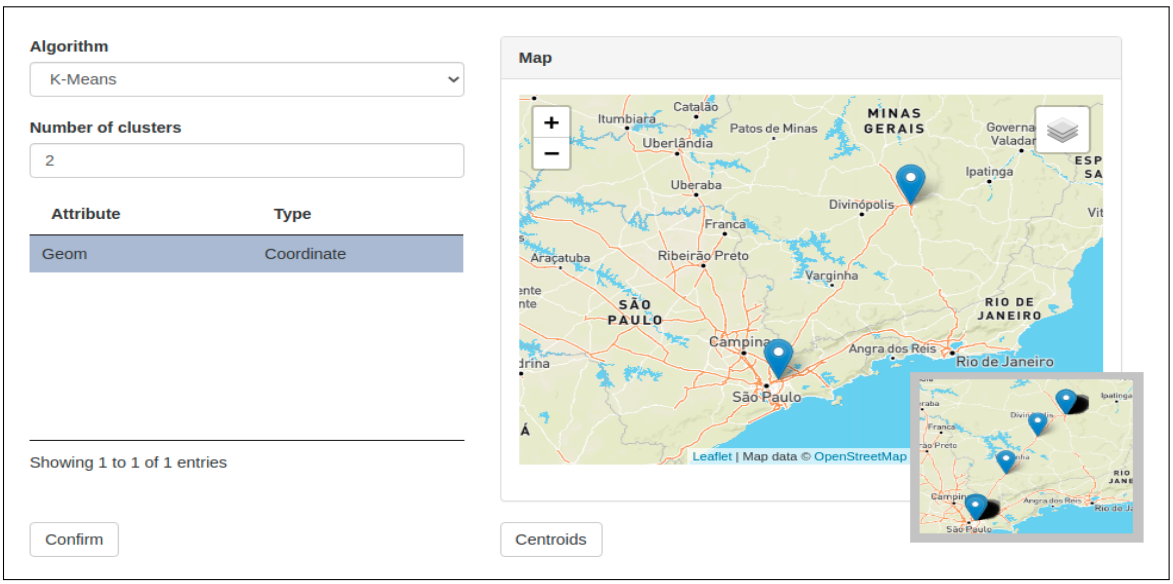

Fig. 8. Clustering of coordinates points, resulting in two clusters illustrated by the blue markers. The smaller map represents the original data, before clustering. It contains two main concentrations of markers.

Clustering is an important task for discovering relevant locations when analyzing georeferenced data. To perform the georeferenced data clustering, DCluster implements a parallel version of the algorithm *K-means*. The clustering process is done in the back-end, which is a more appropriate scenario when working with large amounts of data. Compared with the clustering provided by the Mapbox API [Mapbox 2021], in DCluster only the centers of the clusters are sent to the front-end, which avoids overloading the personal computer of the user. The concept of parallelism used as the basis for the implementation of the algorithm is the master/slave processing, as presented in the work [Hadian

and Shahrivari 2014]. The master process is initially responsible for dividing the data into parts that are sent to the slaves that are responsible for the first step of the clustering. Each slave returns the center found to the master, who later uses this data received as input for the *K-means* for the final processing. Figure 8 presents an example of clustering using a coordinate attribute.

To assess the parallel implementation of the cluster, tests were performed on two datasets of different sizes. The results showed a performance improvement of 26.6% for a base of 73 MBytes and 35% for a base of 310 MBytes, when compared with the traditional version of *k-means*.

3.8 PoI identification

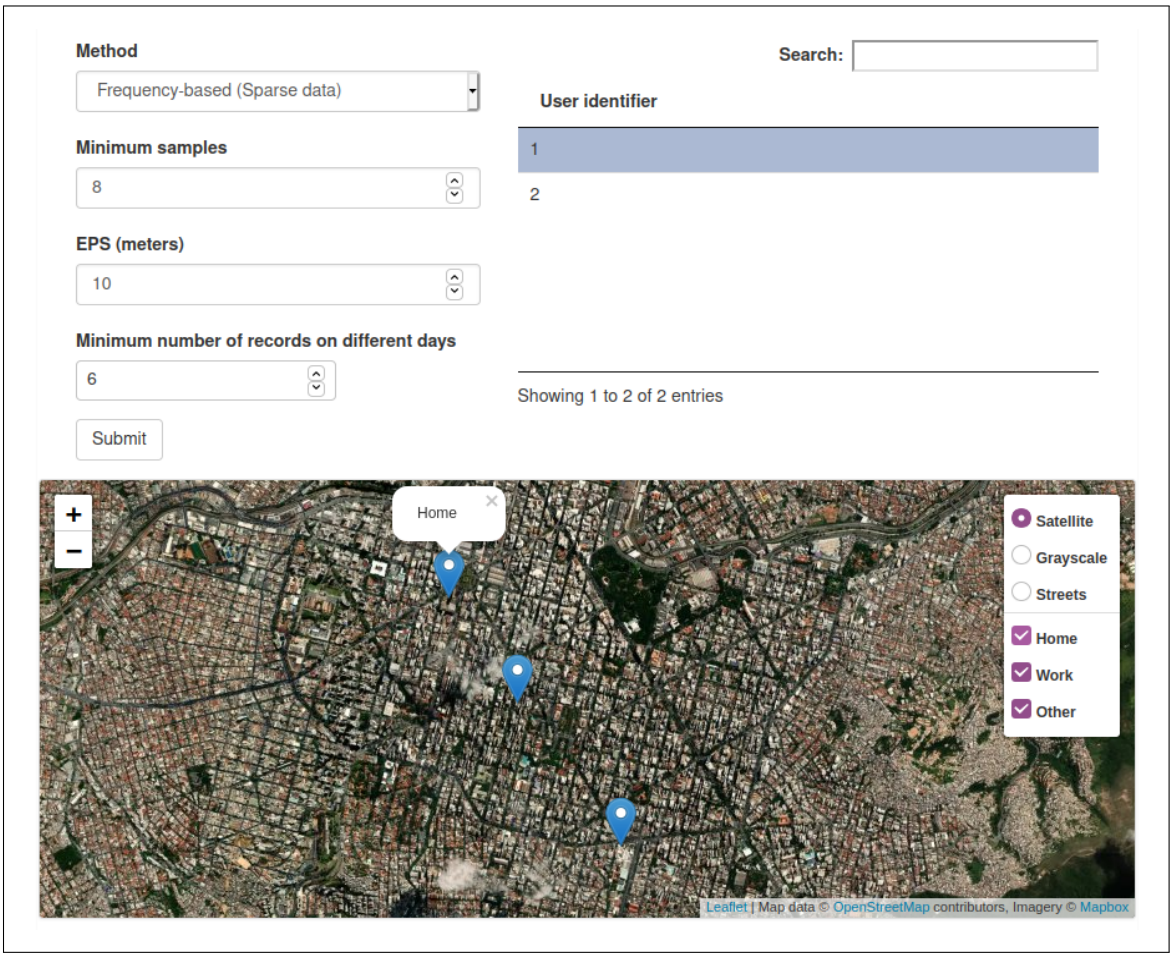

Fig. 9. PoI identification.

Given a georeferenced dataset containing GPS-based location from mobile users, it is possible to discover personal Points of Interest (PoI) and their respective categories (*Home*, *Work*, or *Other*). DCluster provides that functionality in the "PoI identification" option. There are four input parameters, as illustrated in Figure 9. *Method* is the algorithm to perform PoI identification. Currently, the only available option is the "Frequency-based (sparse data)" that is designed to sparse data and was evaluated in [Capanema et al. 2019]. This algorithm has the advantage of finding *Home*, *Work*, and *Other* locations of users that have different types of routines. For example, in general, people go to work in the morning and go back to home in the early night. However, in some cases, individuals

can have an inverted routine, that is, they go to work at night, and stay home during the day. Thus, those nuances are captured by the [Capanema et al. 2019] algorithm to improve the identification of PoIs types.

The *Minimum samples* and *EPS (meters)* are parameters of the DBSCAN algorithm, which is internally used by the PoI identification algorithm. The *Minimum number of records on different days* is a parameter used to avoid locations that were frequently visited in a small number of days to be classified as points of interest. The higher the values of those parameters, the more restrictive the algorithm is. Therefore, if the intention is to find the most important PoIs, the user should enter high values.

The identified PoI for the selected user are shown on the map (See Figure 9), where each marker has a pop-up that describes the PoI category (*Home*, *Work* or *Other*). As mentioned before, the demonstration dataset has synthetic data and does not represent a real person. The user of the given example has three PoI. From top to bottom, the first PoI corresponds to the *Home*, the middle one is a location that has a generic meaning which we call as *Other*, and the bottom PoI is the *Work* location. Those semantic annotations of PoIs are very important, as they can help us to know the most important locations that involve the mobility of each individual.

3.9 Export

After exploring the data, the user can export the filtered data and all the discovered points of interest for further analysis, through a CSV file format. Additionally, all graphs generated by the tool can be downloaded as ".png" file extension.

4. CONCLUSION AND FUTURE WORK

This work presented DCluster, a web system designed to analyze different types of data with a focus on georeferencing. DCluster aims to unite four essential aspects: support for georeferenced data, accessible license, usability, and availability via the Web. These factors are necessary for the popularization of data analysis to make it more accessible to researchers and data analysts who can not afford paid tool licenses but at the same time need a comprehensive set of features.

There are several challenges to be addressed on DCluster. First, a sound enhancement is to recognize and provide visualizations for different spatial geometries (i.e., Polygon, Line, among others). Subsequently, DCluster will be integrated with Big Data tools such as Hadoop and Spark to improve scalability. Finally, other algorithms for the analysis of georeferenced data will be made available, such as a PoI identification algorithm for dense data, a method to predict the category of the next place that a user is likely to visit [Capanema et al. 2020], and possibly an algorithm for detecting stops of automobiles [Nogueira et al. 2018].

Acknowledgements

We would like to thank the research agencies CAPES, CNPq, FAPEMIG, and grants 15/24494-8 & 18/23064-8, São Paulo Research Foundation (FAPESP).

REFERENCES

BigML. Bigml: Machine learning made easy. `https://bigml.com/`, 2011. Accessed on 11/2/2021.

Capanema, C., Silva, F. A., and Braga, T. M. Identificacão e classificacão de pontos de interesse individuais com base em dados esparsos. In *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. SBC, pp. 15–28, 2019.

Capanema, C. G. S., Silva, F., and Silva, T. Dcluster: Um sistema para análise exploratória de grandes volumes de dados georreferenciados. In *Satellite Events of the 32nd Brazilian Symposium on Databases (SBBD)*, 2017.

Capanema, C. G. S., Silva, F. A., and Silva, T. R. d. M. B. Mfa-rnn: Uma rede neural recorrente para predição de próximo local de visita com base em dados esparsos. In *Anais do XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. SBC, pp. 127–140, 2020.

Coimbra, G. T., Capanema, C. G. S., Silva, F. A., and Silva, T. R. B. Appel: Uma extensão do kepler para enriquecimento de dados geoespaciais. In *GEOINFO*. pp. 176–181, 2019.

dos Santos, W., Carvalho, L. F., Avelar, G. d. P., Silva, Á., Ponce, L. M., Guedes, D., and Meira, W. Lemonade: A scalable and efficient spark-based platform for data analytics. In *2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*. IEEE, pp. 745–748, 2017.

Hadian, A. and Shahrivari, S. High performance parallel k-means clustering for disk-resident datasets on multi-core cpus. *The Journal of Supercomputing* 69 (2): 845–863, 2014.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter* 11 (1): 10–18, 2009.

Hitachi. Pentaho hitachi vantara. `https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho.html`, 2017. Accessed on 11/02/2021.

Inc., P. T. Collaborative data science, 2015.

Mapbox. Maps, geocoding, and navigation apis sdks. `https://www.mapbox.com/`, 2021. Accessed on 18/3/2021.

Microsoft. Azure machine learning. `https://azure.microsoft.com/en-us/services/machine-learning/`, 2017a. Accessed on 11/2/20121.

Microsoft. Power bi. `https://powerbi.microsoft.com/`, 2017b. Accessed on 11/2/2021.

Nogueira, T. P., Celes, C. S., Martin, H., Loureiro, A. A., and Andrade, R. M. A statistical method for detecting move, stop, and noise: A case study with bus trajectories. *Journal of Information and Data Management* 9 (3): 214–214, 2018.

Qlik. Qlik: Data analytics and data integration solutions. `https://www.qlik.com/us/`, 2017. Accessed on 11/2/2021.

SAS. Sas: Analytics, artificial intelligence and data management software. `https://www.sas.com/en_us/home.html`, 2017. Accessed on 11/2/2021.

Sisense. Sisense: Business itenlligence (bi), software and analytics tools. `https://www.sisense.com/`, 2021. Accessed on 11/2/2017.

Statista. Statista: the portal of statistics. `https://www.statista.com/statistics/346195/facebook-global-mobile-dau/`, 2017. Accessed on 11/2/2021.

Tableau. Tableau. `https://www.tableau.com/`, 2017. Accessed on 11/2/2021.

Xavier, W. Z., Xavier, F. H. Z., and Marques-Neto, H. T. Visualizing and analyzing georeferenced workloads of mobile networks. In *IEEE International Conference on Pervasive Computing and Communications Workshops*. pp. 306–310, 2017.