- ORIGINAL ARTICLE -

Experiences from a Data Analysis of Crimes against Humanity

Experiencias en Análisis de Datos sobre Crímenes de Lesa Humanidad

Daniela Manrique[1], David Troncoso[1], Agustina Buccella[1], and Alejandra Cechich[1]

[1]*GIISCO Research Group, Departamento de Ingeniería de Sistemas, Facultad de Informática, Universidad Nacional del Comahue, Neuquen, Argentina*
{daniela.manrique,david.troncoso,agustina.buccella,alejandra.cechich}@fi.uncoma.edu.ar

Abstract

Data analysis is a widely researched field, where innumerable applications allow to discover domain particularities that are specially useful. In this paper, we introduce the data analysis process that we applied to two different systems storing information about statements and testimonies of crimes against Humanity. We describe the activities, design decisions and lessons learned from implementing a specific goal, which involves transforming text data into georeferenced information.

Keywords: Data Analysis, ETL Process, Spatial Data Mining, Crimes Against Humanity

Resumen

El análisis de datos es un área ampliamente estudiada y de la cual existen muchas aplicaciones que permiten decubrir particularidades en los dominios donde se aplica y que pueden ser de gran utilidad. En este trabajo describimos un proceso de análisis de datos aplicado a dos sistemas que almacenan declaraciones y testimonios de delitos de lesa humanidad. Siguiendo un objetivo específico, que involucra la transformación de datos de texto en información geo-referenciada, describimos la actividades necesarias para cumplirlo, las decisiones realizadas y las lecciones aprendidas.

Palabras claves: Análisis de Datos, Proceso ETL, Minería de Datos Espacial, Crímenes de Lesa Humanidad

1 Introduction

A data analysis process contains a set of activities or phases that includes the selection of the data sources, the processing of important information, the storing in a specific repository and the analysis of the stored data. Thus, a data analysis development includes firstly a ETL (Extraction-Transformation-Load) process for preparing data, and then a set of configuration tasks for the application of data analysis techniques. They are applied for discovering patterns and relations that could be interesting for a domain and that cannot be obtained by simple queries. In general, the application of a data analysis process generates an important contribution for the domain in which it is applied. We can see this in the many articles in the literature about applications of data analysis processes to different domains, such as education, market, health, climate, etc. [1].

In this article, we introduce an extension of the work presented in [2], in which we described a real experience of a ETL process on text data. The process was applied to two systems storing information about testimonies and trials of crimes against humanity in Argentina during the 70-80's decades. Now, we further elaborate one of the goals of this previous work, aiming at extending the ETL activities and adding spatial data analysis. The first system, named *Sistema Informático de Procesamiento de Declaraciones en Juicios de DDHH*[1] (SIPDJ), was developed by the Faculty of Informatics of the Comahue National University[2] and stores information about statements during the trials Escuelita I and II. The another system, *Sistema de Análisis Sociológico de Querellas*[3] (ASQ) was developed by the Center for Genocide Studies[4] and stores information about part of the ABO Trial (Atlético, Banco y Olimpo), the Operative Trial of Independencia de Tucumán and the Santiago del Estero mega case.

Both systems store similar information; however SIPDJ is focused on human rights trials and their statements, and ASQ is focused on the testimonies for collaborating with the trial complaints. At the same time, both systems require storing information (about testimonies and statements) in very extensive text for-

[1]Computer System for the Processing of Statements in Human Rights Trials

[2]`https://faiweb.uncoma.edu.ar/`

[3]System of Sociological Analysis of Complaints

[4]`https://www.untref.edu.ar/instituto/ceg-centro-de-estudios-sobre-genocidio`

mats without a predefined structure. In spite of this characteristic seems a negative aspect of both systems, the extensive texts are necessary for capturing all possible information witnesses remember about his/her kidnap (specially places and people). With this in mind, we define several goals to be performed with the information stored. In particular, in this article we describe the data analysis process developed for one specific goal: *Analyze the geospatial distribution of the places in which victims were kidnapped related to the location of clandestine detention centers (CCD) (CCD*[5]*)*. To do so, in addition to apply a ETL process for transforming the information in a specific format, we apply spatial data analysis techniques. In this way, this work contributes to show a real experience of a data analysis process on text data, providing decisions and lessons that may be of interest in similar situations.

The article is organized as follows. In the following section we describe the related work. Then, we define a data analysis process to be applied in Section 4 to the both systems. In Section 5 we analyze the lessons learned that can be useful for the application of ETL processes and data analyses with similar characteristics. Finally, conclusion and future work are addressed.

2 Related Work

We divide the related work into two areas, involving methodologies and techniques for: (1) ETL, and (2) data analysis (or data mining). In the first case, ETL processes involve methodologies and techniques focused on three main activities: extraction, transformation and load.

To address those activities, the literature shows several approaches describing the design and modeling of ETL processes. In general these proposals can be classified into those based on UML [3, 4, 5, 6], BPMN [7, 8], or on semantic web technologies, like ontologies or logical models [9, 10, 11]. With respect to the proposals based on UML, in general the approaches extend the language with stereotypes for representing different functions of the ETL process. For example, in [5] authors add tasks to the class diagrams, such as filtering, joins, aggregations, etc., in order to represent transformations and loading. At the same time, in the last years commercial and open source ETL tools for supporting several of the ETL activities have emerged. Some known tools are, for example, *Talen Open Studio*[6], *Apache Nifi*[7], *Snaplogic*[8], etc.

With respect to the second area, data analysis, we are particularly interested in spatial analysis. Spatial data mining (SDM) can be defined as the extension of data mining for extracting implicit knowledge over spatial information [12]. This extension considers the inherent features of geographic information systems (GIS) with respect to the way of representing spatial data and their operations. Among the different techniques in this area, we can cite spatial association rules, spatial clustering, trend detection, etc.

We are particularly interested in the *hotspot analysis*, which is a spatial analysis technique focusing on the detection of higher concentrations of events on specific points. The hotspot analysis has been applied in several domains such as public health, epidemiological research, forest and land fire, etc. [13, 14]. Some of the techniques used by the hotspot analysis are the clustering algorithm K-Means, PAM (Partitioning Around Medoids), CLARA (Clustering Large Applications), CLARANS (Clustering Large Applications based upon Randomized Search), DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Point to Identify the Clustering Structure), etc. For example, CLARANS is an extension of PAM and CLARA based specifically on spatial analysis [15].

These techniques have been applied on different domains. For example, in [16] authors applied K-means as indicator for forest fires occurrence in Indonesia. Also, DBSCAN has been used to detect changes in the distribution pattern of hotspots and land cover of peat land in Sumatra. In [17], a modified algorithm of PAM was applied for mobile network planning, that is, to determine if an area satisfy the mobile requirements. OPTICS was applied to determine outliers in hotspot data in the Riau Province, and PAM together with a LOF (Local Outlier Factor) algorithm were applied in [18] for calculating a land fire indicator in Indonesia.

Similarly, in this work, we analyze and apply different geospatial analysis techniques, such as DBSCAN, KNN (k-nearest neighbor), and heatmaps techniques for preparing, analyzing and visualizing data.

3 The Analysis Process

Firsly, let us clarify our intent, which is not to propose a new analysis process. We know, as we have described in the previous section, that there are several proposals in the literature about ETL, and Data Mining processes and techniques. From them, we extract some of the most important activities (or essential ones, according to the literature) in order to specify the data analysis process that we use in this work. As we can see, in Figure 1, the process is contextualized within the common phases of any software development, but including specific activities related to the ETL process and data analysis.

Following, we briefly describe the activities involved in each phase:

- *Phase 1. Planning*: In this activity, the goals of the process must be defined. These goals describe the final use of the data extracted from data

[5] In Spanish Centros Clandestinos de Detención (CCD)
[6] `https://community.talend.com/`
[7] `https://nifi.apache.org/`
[8] `https://www.snaplogic.com/`

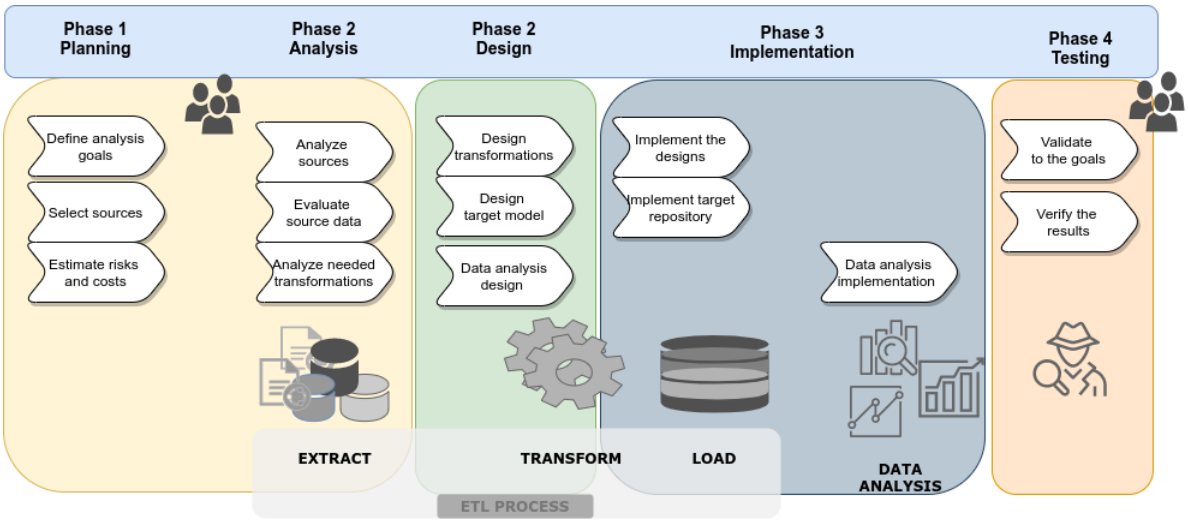


Figure 1: General Analysis Process

sources; that is, the goals denote what engineers (software engineers, data engineers or data scientists) want to obtain from data. Also, with these goals in mind, data sources must be selected. Finally, in this activity is important to estimate costs and the possible risks to achieve the goals.

- *Phase 2. Analysis*: In this activity, the data sources selected must be analyzed and evaluated. Here it is necessary a deep analysis about how data are organized and structured on each data source. Also, according to the analysis of the information in the data sources, engineers must to analyze all the transformations needed to format data as required.

- *Phase 3. Design*: With the data sources analyzed, and a concrete knowledge about how the information is available, the transformations are designed. These transformations include changes, deletions, and even new derived information. Also, a target model is designed according to the way the information need to be organized and extracted. With respect to the extraction, also engineers must define and design the data analysis model to be applied.

- *Phase 4. Implementation*: As the previous designs must be implemented, a tool or set of tools must be selected to carry out the transformations (semi) automatically. Also, the target repository must be selected and created according to the model. It can be a simple database or other repository such as a data warehouse or a data lake. Finally the configuration and implementation of the data analysis technique is also performed together with the obtained results.

- *Phase 5. Testing*: In this activity V&V tasks must be performed in order to verify the results with respect to the goals defined in the first phase; and validate them with the final users.

In the next section we describe an application of this data analysis process to analyze text data of crimes against humanity in Argentina.

4 Application of the data analysis process

The application of the analysis process has been performed over two data sources containing information about crimes against humanity in Argentina during 70-80's decades. These data sources, ASQ and SIPDJ, store similar and related information of testimonies of statements during human right trials. As the information is, in general, stored in very long text format fields, it is necessary to transform them (by means of the ETL process) into useful formats to apply spatial data analysis techniques. Thus, by following the general analysis process (Figure 1) specified in the previous section, following we describe all the activities guided by a specific goal.

- *Phase 1. Planning*: In this work we are focused on one data analysis goal: *Analyze the geospatial distribution of the places in which victims were kidnapped related to the clandestine detention centers (CCD)*

  The specific tables selected for achieving this goal are the *secuestro*[9] tables from both ASQ and SIPDJ systems; *testimonios*[10] and *campo*[11] from ASQ; and *victima*[12], *persona*[13], *traslado*[14]

[9] kidnap
[10] testimonies
[11] detention centers
[12] victim
[13] person
[14] transfer

y *cautiverio*[15] from SIPDJ.

For the case of the *secuestro* tables, the information stored in both systems is similar with only some differences in the data recovered from testimonies. For example, some testimonies can contain more or less descriptions of people also kidnapped in the same places or periods. The other tables contain personal information about the victims and their transfers from one CCD to others. The *secuestro* tables contain the information about the place and the date kidnaps were performed. By joining these tables with the testimonies and the CCDs, we can find the place in which victims were kidnapped.

- *Phase 2. Analysis*: To achieve the goal defined in the previous phase, we need to extract the places in which victims have been kidnapped and the CCDs to which they have been transferred. The systems do not store neither GPS points or any type of geographic coordinates in the tables. The addresses are stored in text format with incomplete and missing information. The incompleteness of this information is not an indication of low quality of data. The information is incomplete, because it is what witnesses remember from the moment they were kidnapped; so any data about the addresses are important to be stored. For example, for the *secuestro* table of the ASQ system, in Figure 2 we can see the *lugar_secuestro*[16] attribute. There are several ways the addresses are stored on each field, such as:

  1. San Martín 151, San Miguel de Tucumán
  2. Su domicilio, sito en el Pasaje Ecuador 135, barrio El Palomar en La Banda del Río Salí (Cruz Alta, Tucumán)
  3. Su domicilio calle Uruguay 4532, San Miguel de Tucumán, Tucumán
  4. Me secuestraron el 1ro. de Junio de 1978 en un bar a una cuadra de General Paz y Av. de los Constituyentes.

  The other tables selected, also contains missing information and/or wrong formats to be analyzed. So, the output of this phase involved the specific attributes of each table that must be considered, together with the associated problems.

- *Phase 3. Design*: Based on these problems, we perform the transformation designs. In this work, we only show the design needed to clean and reformat the information of the *secuestro* tables; however more designs for the other tables have been also performed. The design can be observed in Figure 3, in which we apply filters, aggregations, joins and loads for formatting and improving addresses. For the designs we use the approach defined in [5].

[15] captivity
[16] kidnap_place

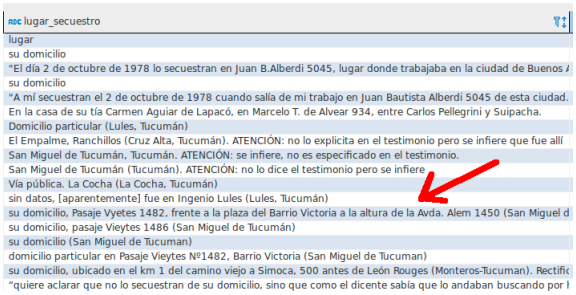


Figure 2: Data stored in the ASQ system about the *lugar_secuestro* attribute

Filters and join. Considering the *secuestro* tables of both systems, we apply three **filters** and a **join** to unify the similar attributes. Thus, we firstly apply two *filter* components for recovering only the needed attributes; *id*, *id_secuestro*, *id_testimonio*, *lugar_secuestro* and *momento_del_dia_secuestro* of *secuestro* table of ASQ system, and *id_secuestro*, *numerodeclaracion*, *lugarsecuestro* and *momentodia* attributes of SIPDJ system. These two filters are combined in a new filter (in blue en the figure) responsible for recovering only data that can represent spatial addresses. These data are stored in a new attribute called *direccion_secuestro*. Also, we use the *su_domicilio* attribute because its values store data indicating if the testimony contains the address of the victim.

Aggregation. The aggregation component (in green in Figure 3) is responsible for creating two new attributes, *latitud*[17] y *longitud*[18]. These attributes store information in a georeferenced mode.

Load. All the new and formatted attributes are loaded in a new table named *direccion_secuestro*[19]. This table contains a formatted attribute (*lugar_secuestro*) and other useful attributes such as the date of the kidnap event, an identification, the latitude and longitude derived, etc.

- *Phase 4. Implementation*: Here, we use some of the functionalities provided by Talend Open Studio for filtering and joining actions defined in the designs. For the case of the transformation design showed in Figure 3, we use the tool for the two

[17] latitude
[18] longitude
[19] kidnap_place

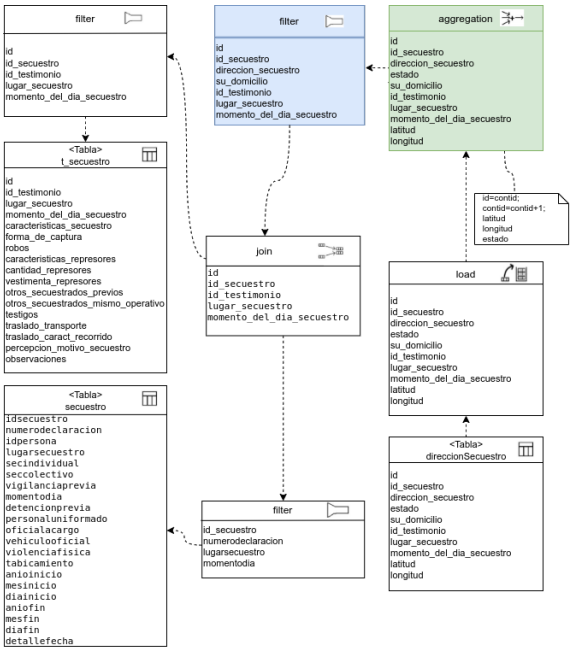


Figure 3: Transformation design for deriving georeferenced attributes (latitude and longitude) from *secuestro* tables

filters (in white in the figure) and the *join* component. Then, for the aggregation and generation of the georeferences attributes, we create a new file including the addresses in text format. This file is the input of a new algorithm developed in Java, which normalizes the addresses to a specific format according to a regular expression defined. This regular expression was the result of several tests on the data from which we maximize, as far as possible, data involving possible geographical representations of events. This is achieved by adding rules to the expression, or modifying them to restrict some special cases[20]. The algorithm validates the strings against the regular expression generating an output file with a set of georeferenced addresses. Thus, data that could represent streets and numbers or a combination of 2 streets (either intersection or parallel streets) are placed at the beginning of a string by the *lugar_secuestro* attribute. This string is followed by the strings corresponding to city, province and the *Argentina* word. Finally, the algorithm creates a new attribute, named *state* with a numerical value from 1 to 4. These values are generated according to the precision obtained in the georeferencing process. Value 1 represents that the address has all needed data for specifying a location on a map. On the other hand, higher values indicate that the addresses have some incomplete values. Thus, the value of this attribute is defined as follows:

$$\left.\begin{array}{c} street + number + city + province = 1 \\ street + number + [city]|[province] = 2 \\ street + number = 3 \\ \emptyset = 4 \end{array}\right\} address_precision \quad (1)$$

When the algorithm finishes, the addresses are processed by the Google API ($<< googleMapsClient geocode >>$). Those records with a 4 value are not exposed to the next detection process, since it is known that they will not be able to be georeferenced. From the analysis with the API, it was possible to obtain approximately 300 georeferenced results. Following, the georeferenced values are stored on the *latitude* and *longitude* attributes of the *direccion_secuestro* table and dumped into a CSV file (by using Talend).

Finally, we use the programming language R[21] for applying the algorithms DBSCAN and KNN (k-nearest neighbor). These algorithms are used together to find and delete the outliers points in the map. The $dbscan()$ function has two parameters, ε (eps) defining the radius of neighborhood around a point x, and $MinPts$ defining the minimum number of neighbors within ε radius. As it is important to define an optimal value for ε (to find correct clusters), we compute the k-nearest neighbor distances in a matrix of points. Thus, we can calculate the average of the distances of every point to its k nearest neighbors. In Figure 4a we can see the result of applying DBSCAN (by using the fpc function) with outliers points. Then, Figure 4b shows the k-distance plot[22] for determining the *knee*[23], which corresponds to the optimal ε parameter. Finally, once *dbscan* is applied, we can delete the outliers points and obtain the correct clusters. In Figure 4c we can see the graphical representation without outliers or noise. In this point, we have the dataset ready for the spatial analysis activity.

For the data analysis activity, we use the libraries *ggmap* and *ggplot2* of the R programming language[24]. These libraries allow to create static maps for representing different features about data. For example, in Figure 5a we can see three maps generated according to the address precision (*state* attribute) defined in the transformation process. Here we use the red color for a high precision ($state = 1$), green color for a medium precision ($state = 2$), and blue for a low precision ($state = 3$). Following, in Figure 5b the same maps are represented as *heatmaps*, which

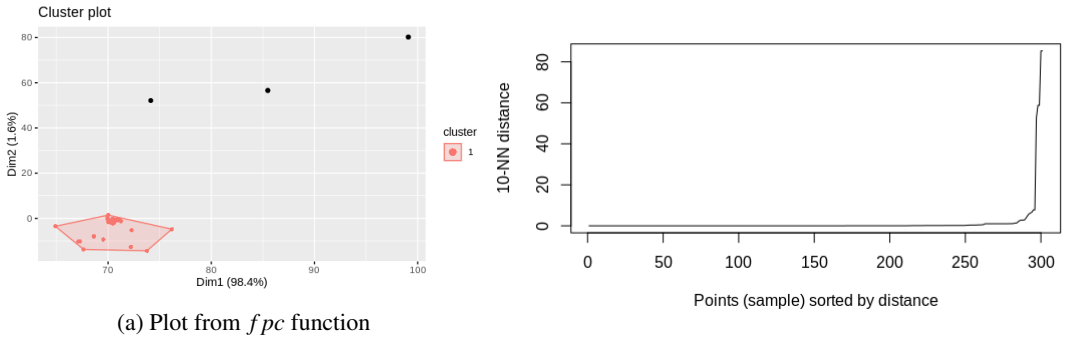
[20] Because of space reasons, we do not include the regular expression. This expression includes more than 20 different patterns for representing the different formats in which addresses were stored.

[21] `https://www.r-project.org/`

[22] by using the $kNNdistplot()$ function

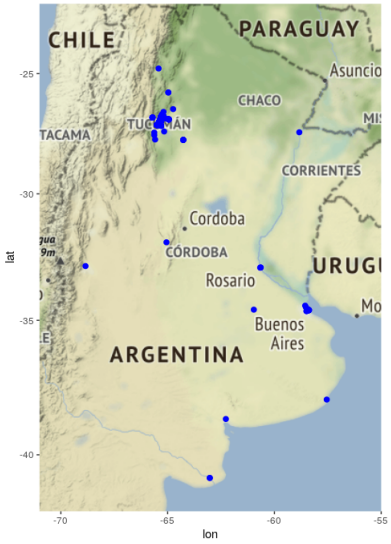
[23] A knee corresponds to a threshold where a sharp change occurs along the k-distance curve

[24] `https://cran.r-project.org/web/packages/`



(a) Plot from *fpc* function

(b) k-distance plot



(c) Representation after noise deletion

Figure 4: Application of DBSCAN

add a layer for the intensity of data at geographical points by a set of colors, from red to yellow for denoting more-less intensity (red denotes the highest intensity of points, and yellow the lowest). In this way we can identify different data distribution where kidnappings were registered.

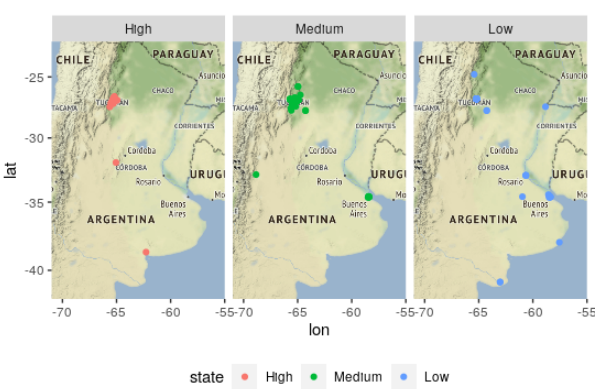
Following, we include the locations of the CCDs (extracted from an external table) and the places in which victims were transferred the first time they were kidnapped. This information was obtained from joining the information in testimonies and clandestine center tables already cleaned and formatted. In addition, as some CCDs in these testimonies were not found, we extracted this information from the *Informe de Investigación RUVTE-ILID* [25]. The information of these tables was included in the datasets by following similar transformation processes as those described in the design activity (for the CCD addresses).

As the maps generated in Figure 5 are statics, we use the LeafLet library[26] in order to generate dynamic maps with layers interactions. In Figure 6 we can see part of this map showing different layers (with marks) in different colors. Each mark has two colors, the outer one represents the address precision (green for high, orange for medium, and red for low), and the inner color (rounded the x) represents the specific place in which the victims were transferred. The references to these colors are in the lower right corner, in which the specific name of the CCD is indicated. Also, the marks can be showed/hided by interacting with the layers (in the upper left corner). When a mouse is hovered over the marks, a menu is displayed showing descriptive data. Finally, the last layer represents the CCDs as circles with the same colors as in the (left) reference. To interact with the map, readers can visit the following url: `https://bit.ly/3rDbYRr`.

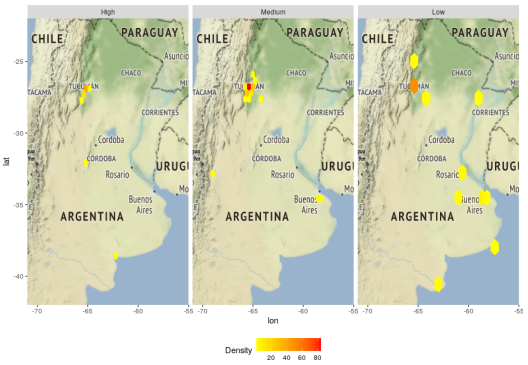
Remember that the application of the data analysis process is to analyze geospatial information

[25] Informe de Investigación sobre Víctimas de Desaparición Forzada y Asesinato, por el accionar represivo del Estado y centros clandestinos de detención y otros lugares de reclusión clandestina. `https://diegokoz.github.io/presentes/index.html`

[26] `https://leafletjs.com/`



(a) Kidnap places according to the address precision



(b) Heatmaps according to the address precision

Figure 5: Maps generated by using the libraries *ggmap* and *ggplot2*

about places in which victims were kidnapped and their relation to the CCDs. From the interactive map, we can extract several analyses; for example, we can obtain the zones in which more victims were kidnapped and the more used CCDs. This is useful because we can obtain information about the repressive routes; that is, the starting points in which the victims were brought into the repressive system. For example, in Figure 7 we can see a view of the *San Miguel de Tucumán* city with two layers, the marks with higher precision, and the CCDs. This last layer, marked with colored circles, shows the clandestine centers where victims said they were transferred (in testimonies). If we focus on the *Jefatura de Policia de Tucumán* center (the purple circle highlighted with number 1), and the markers of the kidnap places that have the inner circles of the same color (with number 2), we can observe that this clandestine center was the destination of numerous kidnappings that were not necessarily close to each other; but they came from very different points of the city, even when there were other CCDs closer to these points. This may suggest that this CCD was an ”entry” point into the repressive routes for victims of kidnapping.

In addition, we can change few parameters of the libraries to add more layers and/or marks for analyzing concentrations of victims in CCDs, or the routes in which victims were transferred.

- *Phase 5. Testing*: In this activity we verify that the system performed the analysis correctly according to the goal defined; and we also validate the results with expert users of the SIPDJ system. Validating the web site with expert users of the ASQ system is currently work in progress.

5 Lessons Learned

From the application of our data analysis process to a real case study we can highlight the following lessons learned:

- *Knowledge about the domain was crucial for the process*: the work performed between engineers and domain experts was crucial to understand the way the information was stored, as well as the correct way to normalize and restructure it. By the nature of the system, the attributes were stored as very long strings with many descriptions. Although these descriptions seemed redundant at first glance, the experts highlighted the importance of storing them in the statements or testimonies since they collect not precise, but important information about the detentions. Therefore, a very careful extraction of the data had to be carried out so as not to eliminate important information of the testimonies. For this reason, we decided to use, in the case of kidnap places, colors to indicate the address precision.

- *The definition of a data analysis process with well-defined phases for ETL and data analysis activities helped organize the team and reproduce practices*: for the ETL design approach we used a modeling approach based on UML. This selection was useful for the engineers because of UML is a language widely used in software development and known by the team. They only just had to learn the particular mechanisms of the ETL desing process. At the same time, the methodology allowed design decisions to be replicated among the problems founded. Particularly when the CCDs have to be loaded on the map, where many addresses were also incomplete and mechanisms performed for the *secuestro* tables were repeated.

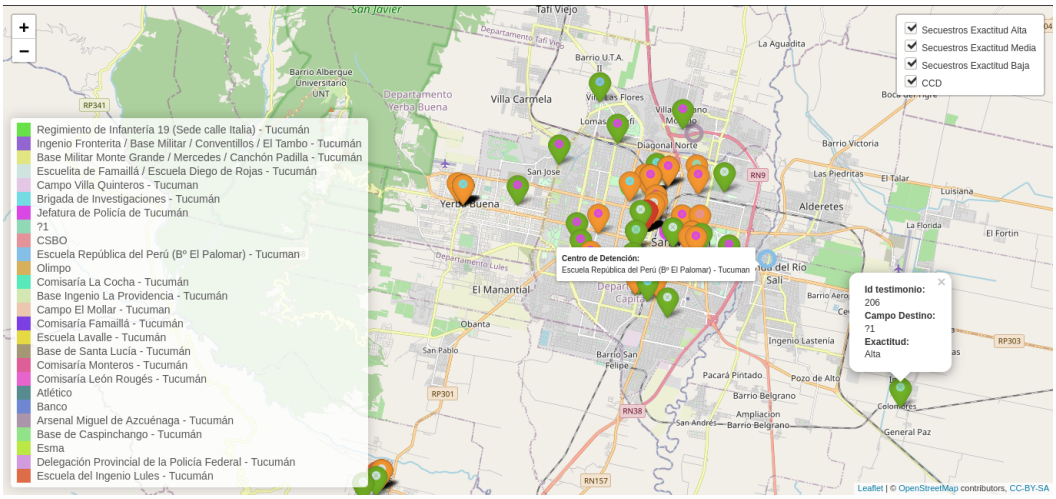


Figure 6: Map with layers according to the address precision, the CCDs and places in which victims were kidnapped

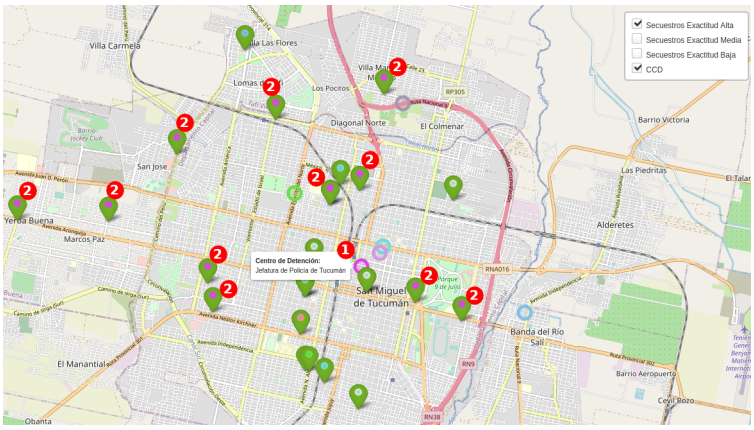


Figure 7: A view of kidnappings in Tucumán together with the CCDs involved

- *The application of the Talend Studio tool sped up the implementation of the process*: this tool is very intuitive and simple, so the implementation was done relatively quickly. At the same time, with the transformation designs already outlined, the mapping between them and their implementations in Talend was often trivial. Likewise, in some cases we had to add some extra programming to precisely capture the particularities of this domain. This was the case of the addresses for the goal presented here.

- *The result of the ETL process promotes the application of a spatial data analysis with useful results in this domain*: the pre-processed information from the data sources was correctly processed to be used for spatial data analysis techniques. These techniques allowed to visualize and analyze the recurrent kidnap places and the starting points of the repressive circuits. At the same time, with this pre-processed information we can extend the geospatial analyses such as denoting the heatmaps for CCDs through which the same person passed, places where they were kidnapped and routes they took during detention, etc.

6 Conclusion and Future Work

In this work we have defined and applied a data analysis process for two systems storing information about crimes against humanity in Argentina during the 70's-80's. Particularly, we have defined a specific goal for analyzing the distribution of places in which victims were kidnapped, and we have showed the activities performed to achieve it. Also, we have highlighted the lessons learned from performing the whole processes.

As future work we are working on extending the analysis by applying other spatial analysis techniques.

Competing interests

The authors have declared that no competing interests exist.

Authors' contribution

D.M and D.T analyzed the tools for ETL and for spatial data analysis . They specified the ETL designs and implemented the activities of the data analysis process. A.B and A.C

defined the data analysis process and supervised the activities and the findings of this work. All authors discussed the results and contributed to the final manuscript.

Acknowledgements

This work is partially supported by the UNComa project 04/F009 "Reuso de Software orientado a Dominios: Parte II" part of the program "Desarrollo de Software Basado en Reuso: Parte II

References

[1] P. Neelamadhab, M. Pragnyaban, and P. Rasmita, "The survey of data mining applications and feature scope," *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, vol. 2, 11 2012.

[2] D. Troncoso, A. Buccella, and A. Cechich, "Decisiones y lecciones aprendidas en un proceso etl aplicado a sistemas con testimonios de delitos de lesa humanidad," in *Proceedings of the CACIC'20: XXVI Congreso Argentino de Ciencias de la Computación*, (Universidad Nacional de La Matanza), RedUnci, 2020.

[3] S. Luján-Mora, P. Vassiliadis, and J. Trujillo, "Data mapping diagrams for data warehouse design with uml," in *Conceptual Modeling – ER 2004* (P. Atzeni, W. Chu, H. Lu, S. Zhou, and T.-W. Ling, eds.), (Berlin, Heidelberg), pp. 191–204, Springer Berlin Heidelberg, 2004.

[4] A. Simitsis, D. Skoutas, and M. Castellanos, "Natural language reporting for etl processes," in *Proceedings of the ACM 11th International Workshop on Data Warehousing and OLAP*, DOLAP '08, (New York, NY, USA), p. 65–72, Association for Computing Machinery, 2008.

[5] J. Trujillo and S. Luján-Mora, "A uml based approach for modeling etl processes in data warehouses," in *Conceptual Modeling - ER 2003* (I.-Y. Song, S. W. Liddle, T.-W. Ling, and P. Scheuermann, eds.), (Berlin, Heidelberg), pp. 307–320, Springer Berlin Heidelberg, 2003.

[6] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos, "Conceptual modeling for etl processes," in *Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP*, DOLAP '02, (New York, NY, USA), p. 14–21, Association for Computing Machinery, 2002.

[7] Z. El Akkaoui, J.-N. Mazón, A. Vaisman, and E. Zimányi, "Bpmn-based conceptual modeling of etl processes," in *Data Warehousing and Knowledge Discovery* (A. Cuzzocrea and U. Dayal, eds.), (Berlin, Heidelberg), pp. 1–14, Springer Berlin Heidelberg, 2012.

[8] J.-N. Mazón, E. Zimányi, Z. El Akkaoui, and J. Trujillo, "A bpmn-based design and maintenance framework for etl processes," *Int. J. Data Warehous. Min.*, vol. 9, p. 46–72, July 2013.

[9] D. Skoutas, A. Simitsis, and T. Sellis, *Ontology-Driven Conceptual Design of ETL Processes Using Graph Transformations*, pp. 120–146. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[10] A. Simitsis, "Mapping conceptual to logical models for etl processes," in *Proceedings of the 8th ACM International Workshop on Data Warehousing and OLAP*, DOLAP '05, (New York, NY, USA), p. 67–76, Association for Computing Machinery, 2005.

[11] M. Niinimäki and T. Niemi, *An ETL Process for OLAP Using RDF/OWL Ontologies*, pp. 97–119. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[12] M. Perumal, B. Velumani, A. Sadhasivam, and K. Ramaswamy, "Spatial data mining approaches for gis – a brief review," in *Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2* (S. C. Satapathy, A. Govardhan, K. S. Raju, and J. K. Mandal, eds.), (Cham), pp. 579–592, Springer International Publishing, 2015.

[13] S. Tay, W. Hsu, K. Lim, and L. Yap, "Spatial data mining: Clustering of hot spots and pattern recognition," vol. 6, pp. 3685 – 3687 vol.6, 08 2003.

[14] T. H. Grubesic, "On the application of fuzzy clustering for crime hot spot detection," *Journal of Quantitative Criminology*, vol. 22, no. 1, pp. 77–105, 2006.

[15] R. T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, (San Francisco, CA, USA), p. 144–155, Morgan Kaufmann Publishers Inc., 1994.

[16] I. S. Sitanggang, T. Fuad, and Annisa, "K-means clustering visualization of web-based olap operations for hotspot data," in *2010 International Symposium on Information Technology*, vol. 1, pp. 1–4, 2010.

[17] L. Fattouh and M. Alharbi, "Using modified partitioning around medoids clustering technique in mobile network planning," *International Journal of Computer Science Issues*, vol. 9, 02 2013.

[18] M. B. C. Imas Sukaesih Sitanggang and Shofyan, "Data mining approach for outlier detection on hotspot data as forest and land fire indicator: A case study in riau province indonesia," *ARPN Journal of Engineering and Applied Sciences*, vol. 12, no. 13, 2017.

Citation: D. Manrique, D. Troncoso, A. Buccella, and A. Cechich. *Experiences from a Data Analysis of Crimes against Humanity*. Journal of Computer Science & Technology, vol. 21, no. 1, pp. 20–28, 2021.
DOI: 10.24215/16666038.21.e03
Received: February 10, 2021 **Accepted:** March 8, 2021.
Copyright: This article is distributed under the terms of the Creative Commons License CC-BY-NC.