Feature Selection within Time Series Clustering

M. Giohanna Martinez, Diego H. Stalder
School of Engineering
Asuncion National University
San Lorenzo, Paraguay

Christian E. Schaerer
Polytechnic School
Asuncion National University
Asuncion, Paraguay

Juan V. Bogado
Caaguazu National University
Caaguazu, Paraguay

Corresponding Author: mgmartinez@fiuna.edu.py

Abstract

In recent decades, the world has experienced a health crisis due to increased infectious disease cases such as COVID-19, Dengue, Zika, etc. Dengue is a neglected tropical disease transmitted by mosquito vectors, mainly by the Aedes Aegypti. This work is focused on Paraguay, where the virus has surpassed 16,000 notifications so far, this 2022. This disease has an incidence throughout the country, which results in a large amount of available data. The time series clustering can find a subjacent structure within a large amount of data, simplifying analysis and interpretation of it. This article contrasts two different clustering methods (Shape-based and Feature-based), followed by a feature selection procedure. Initially, both methods are tested, getting the highest Silhouettes scores with the feature-based approach. Subsequently, one feature is removed in each experiment and the results are ranked, getting higher silhouette scores by eliminating the least important feature. Results show that better clustering is obtained by performing an adequate feature selection through the ranking procedure.

Keywords

Clustering, Time Series, Epidemiology, Dbscan, K-means, Hierarchical.

1. Introduction

Technological advances allow us to collect and store data at a very large scale, using remote sensors, satellites, cameras, and telescopes. The epidemiological data allows us to map the incidence of diseases (COVID-19, Influenza, Dengue, Zika, etc.) and combine them with the mobility of people. Combining and recognizing groups and similarities in the data is a challenge for engineers and scientists who need to develop new algorithms and methods of analysis.

Much of the available data is stored as a time series $X = (x_t: t = 1, \ldots, N)$ collected over a period, usually at regular intervals, where N is the number of observations and x_t is the time-varying value measured at an instant t. The analysis of these time series requires the implementation of filtering, noise reduction, anomaly identification, pattern identification, and clustering algorithms.

Clustering of time series identifies time series with similar characteristics (Aghabozorgi et al. 2015). Similar instances are organized in clusters or groups. The time series clusters can provide relevant information for identifying

geographic regions where a disease outbreak's development occurs similarly. Clustering algorithms can also identify atypical elements, sometimes called outliers, which do not fit into any identified groups.

Dengue is a disease transmitted mainly by the Aedes Aegypti mosquito. It represents a major public health challenge, demanding a rapid and efficient response from the institutions in charge. The identification and intervention of critical clusters increase the possibilities of taking measures in time and reducing the impact of the disease on society and the institution's budget.

Considering the work of Bogado et al. (2021) as a baseline, a comparative study of clustering algorithms in epidemiological data time series is proposed to find the most efficient metrics and methods to identify similarities and anomalies in the number of dengue cases in different districts of Paraguay.

1.1 Objectives

The main aim of this work is to conduct a comparative study for time series clustering by selecting and applying clustering algorithms. Then using internal validation metrics, evaluate the incorporation of more attributes in the time series.

2. Literature Review

Time series are an increasingly common data type, and much research has been dedicated to time series data mining in the past few decades (Guha et al., 2015, Oates 1999, Chen 2005). Analyzing data using clustering methods can help researchers gain further insights from the stored data e.g., (Bogado et al. 2020, Bogado et al. 2021). Several features are used on time series (Hyndman et al. 2015); nevertheless, in the case of disease outbreaks (e.g. Dengue fever) it is not clear which feature should be used to have meaningful clusters.

3. Methods

The clustering algorithm's performance may vary considerably for different applications and types of data, often due to noisy, incomplete, and sampled data, which is important to comprehensively compare methods to determine the most suitable for a given scenario (Rodriguez et al. 2019)**.** Two types of representation have been applied: shape-based and attribute-based. The original data is used for the shape-based representation as a succession of observations ordered by time. In contrast, in the attribute-based representation, this collection of observations is transformed into an N-dimensional vector, where N is the number of selected attributes. This vector is constituted by parameters of the series, which are generally statistical, as proposed by (Bogado et al. 2021; Hyndman et al. 2015). Figure 1 describes the methodology applied in this work.

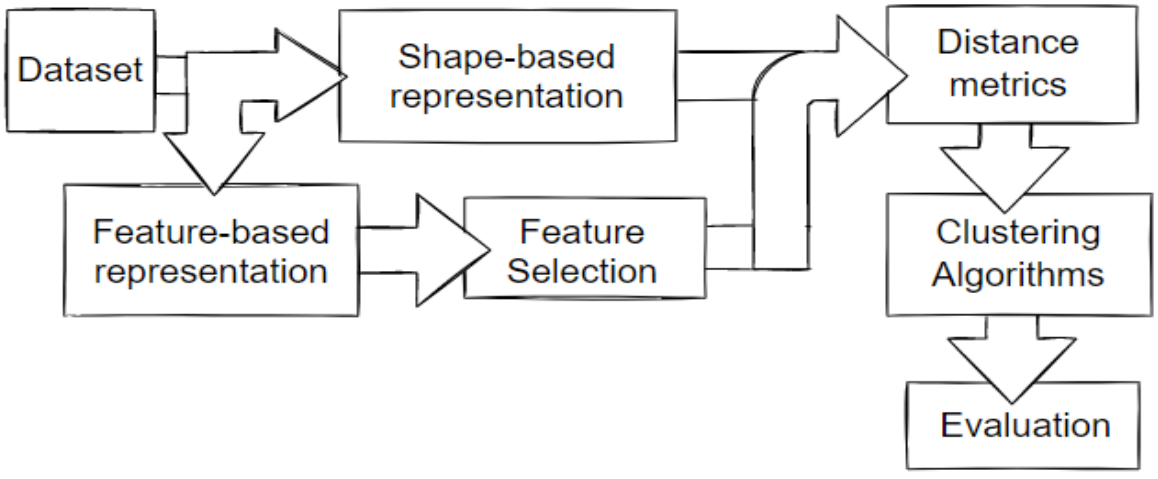

Figure 1. Summarized workflow of this paper. The raw data is used from the time series dataset of Dengue cases in Paraguay, and the features are extracted. Both data (raw data and features) are used in clustering algorithms (agglomerative hierarchical clustering, DBscan, and k-means), and their performance was evaluated.

3. 1. Time series distances

The distance (or dissimilarity degree) between two-time series: T1 and T2, should be considered. The distance metrics can obtain projections of the time series in a certain space in a functional form considering the proximity between the series. The metrics considered in this work are the Euclidean distance, Pearson's correlation, Spearman's correlation, and Dynamic Time Warping.

- *Euclidean Distance:* Let $T_1 = (u_1, u_2, \dots, u_n)$ and $T_2 = (v_1, v_2, \dots, v_n)$ be time series, observed at time instants $(t_1, t_2, \dots, t_n)$. We define the Euclidean distance, between T_1 and T_2, as $Dist\,(T_1, T_2) = (\sum_{i=1}^{n} (u_i - v_i)^2)^{\frac{1}{2}}$. It can be verified that it depends exclusively on the proximity of the observed values at the corresponding time instants.
- *Pearson's correlation:* evaluates the linear relationship, which represents the proportionality between the changes of one series and another.
- *Spearman's correlation*: evaluates the monotonic relationship, which represents the constant variation between two series, which do not necessarily occur at the same rate. A comparison of two functions is made in Figures 2 (A) and Figure 2 (B) (Minitab Express Support 2022).

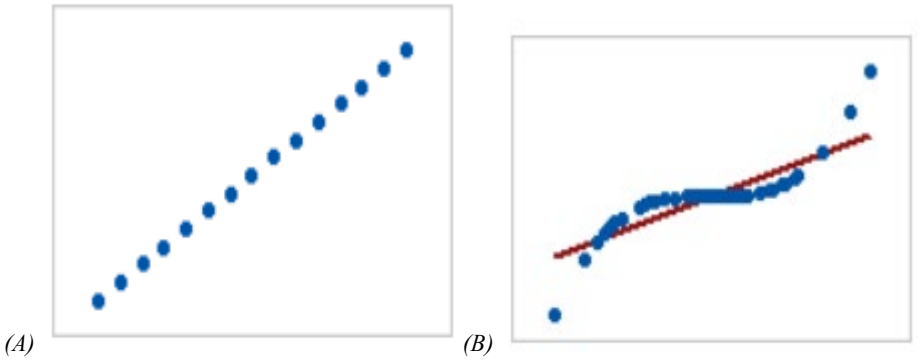

Figure 2. (A) Representation of a function with Pearson's correlation = 1 and Spearman's correlation = 1. (B) Representation of a function with Pearson's correlation = 0.851 and Spearman's correlation = 1.

- *Dynamic Time Warping (DTW)*: this distance measure seeks a pairing of the time indices such that it minimizes the Euclidean distance between two aligned series. Figure 3 represents how the Euclidean distance and the Dynamic Time Warping of two time series are measured (Tavenard 2020).

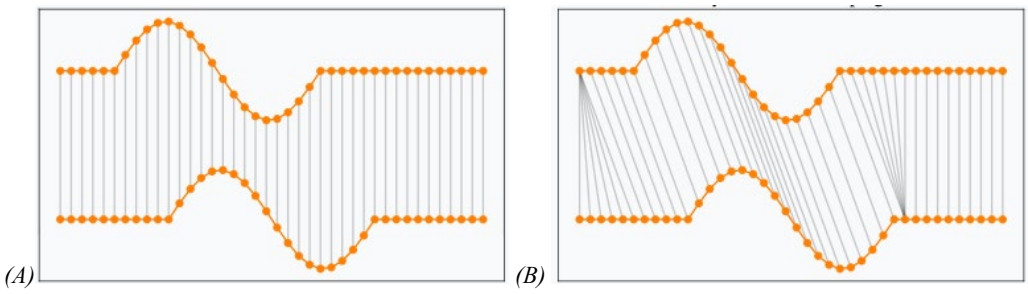

Figure 3. (A) Representation of the Euclidean Distance between two-time series. (B) Representation of Dynamic Time Warping between two-time series. Note that the distance between T1 and T2 is measured to the closest point.

3. 2. Clustering algorithms

According to Keogh and Kasetty (2003), clustering algorithms can be divided into five groups: hierarchical, partitional, grid-based, model-based, and density-based.

- **Hierarchical clustering** creates a series of nested clusters, which can be represented graphically as dendrograms. A dendrogram is a tree-like diagram that shows the similarity between clusters. The shape of the clusters can be seen at the edge of the dendrogram. By setting a threshold in the dendrogram at a certain level, you can determine a set number of clusters for each level (Özkoç 2021). A representation of it is observed in Figure 3 (Pradeep and Singh 2010).

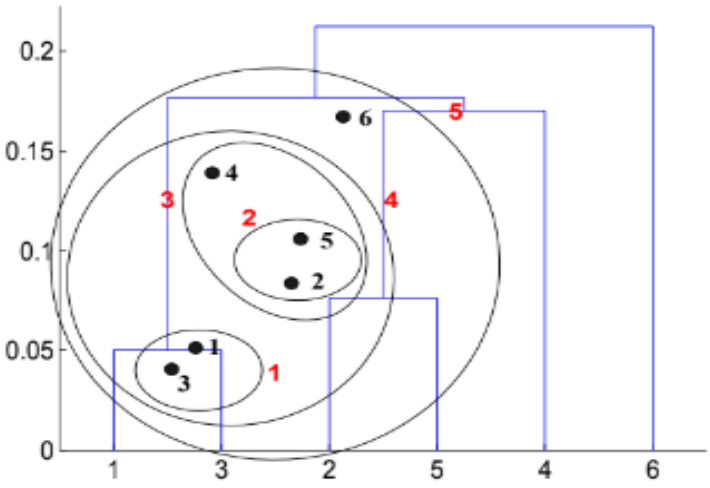

Figure 3. Nested cluster Diagram.

- **Partitional clustering** decomposes a data set into a set of disjoint clusters. Given a N-points data set, a partitioning method performs K ($N \geq K$) division of the data, with each division representing a cluster.

- **Grid-based clustering** uses square cells to explore data places. A representation of it is shown in Figure 4 (Wang and Li, 2017).

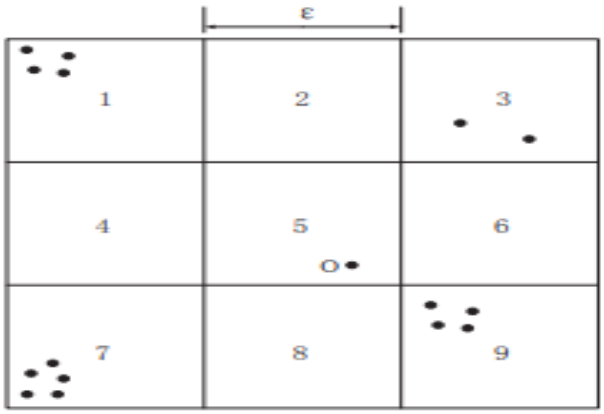

Figure 4. Distribution of data points in grid cells.

- **Model-based clustering** models the cluster structure of time series data. It is assumed that the underlying probability distribution of the data comes from the final mixture. Model-based algorithms usually try to estimate the likelihood of model parameters using some statistical technique such as Expectation Maximization (EM) (Özkoç, 2020).

- **Density-based clustering** defines clusters as dense regions in the data space, separated by regions of lower object density. A cluster is defined as a maximal set of density-connected points and may have an arbitrary shape.

Table 1 lists the three groups of algorithms considered in this work regarding the baseline (Bogado et al., 2021).

Table 1. Groups and clustering algorithms

Group	**Algorithm**
Hierarchical	Agglomerative
Partitional	K-Means
Density-based	DBSCAN

The number of clusters defined for the Agglomerative Hierarchical and K-means algorithms is obtained by the Elbow method, i.e., six groups. The minimum distance and the minimum number of points parameters of the DBSCAN algorithm are determined by stochastic methods. The performance of each algorithm, with their respective metrics, has been calculated using the Silhouette score (Bogado et al. 2021).

4. Data Collection

The data used consists of weekly reports of Dengue fever in 238 districts of Paraguay, corresponding to the period 2009 to 2013, which was obtained from the COMIDENCO project (Gomez-Guerrero 2017) of the *CIMA.*

5. Results and Discussion

5.1 Numerical Results

Table 2 presents the Silhouette score values for the agglomerative hierarchical clustering. Columns indicate that the feature-based approach gives better clustering results. Rows indicate the distance metric considered for each case; the Euclidean distance scores are higher than other metrics.

Table 2. Performance of the agglomerative hierarchical algorithm according to its representation with different distance metrics. Bold values indicate the best ones.

Agglomerative Hierarchical		
Distance Metrics	**Silhouette Score**	
	Shape-based	**Feature-based**
Euclidean Distance	0.909203	**0.931435**

Pearson Correlation	0.332346	0.854293
Spearman Correlation	0.479232	0.509250
Dynamic Time Warping	0.810251	0.923852

Table 3 presents the Silhouette score values for the K-Means algorithm. Columns also indicate that the feature-based approach gives better clustering results than the shape-based approach. Rows indicate the distance metric considered for each case; the Euclidean distance scores followed by the Dynamic Time Warping are higher than other metrics.

Table 3. Performance of the K-Means algorithm according to its representation with different distance metrics. Bold values indicate the best ones.

K-Means		
Distance Metrics	**Silhouette Score**	
	Shape-based	**Feature-based**
Euclidean Distance	0.909203	**0.931101**
Pearson Correlation	0.345943	0.842254
Spearman Correlation	0.537405	0.529607
Dynamic Time Warping	0.814692	0.923985

Table 4 presents the Silhouette score values for the DBSCAN algorithm. Columns also indicate that the feature-based approach gives better clustering results than the shape-based approach. Rows indicate the distance metric considered for each case; the Pearson Correlation scores are higher than other metrics.

Table 4. Performance of the DBSCAN algorithm according to its representation with different distance metrics. Bold values indicate the best ones.

DBSCAN		
Distance Metrics	**Silhouette Score**	
	Shape- Based	**Feature-based**
Euclidean Distance	-0.377393	-0.508293

Pearson Correlation	0.233523	**0.852580**
Spearman Correlation	0.434410	0.571939
Dynamic Time Warping	-0.352043	-0.500988

Table 5 presents the two highest values of each algorithm and the best-performing distance measure from the first experiment.

Table 5. Ranking of attribute extraction according to algorithm.

Algorithm	Rank	Excluded Feature	Silhouette Score
Agglomerative Hierarchical	1	Cpoints	0.931544
	2	Mean	0.931438
K-Means	1	Cpoints	0.931201
	2	Mean	0.931103
DBScan	1	Var	0.891829
	2	Curvature	0.867007

5.2 Graphical Results

Figure 5 shows the six clusters obtained when the districts have outbreaks in the same year, and with similar incidences. The clustering result was visualized by applying Principal Component Analysis to reduce the dimensionality and verify the clusters (Jolliffe and Cadima 2016) (see Figure 6).

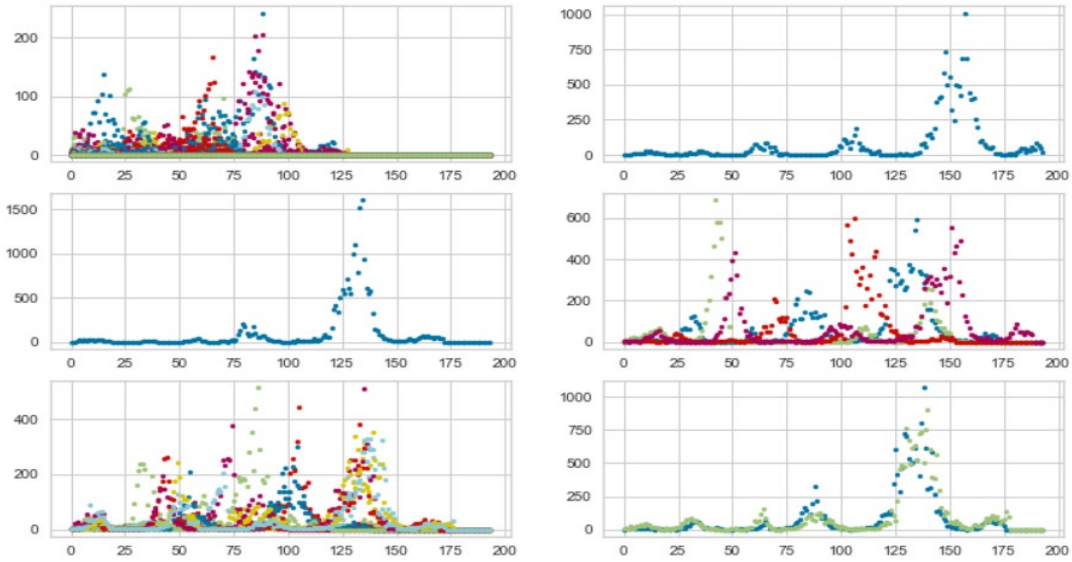

Figure 5. Clusters are formed by Agglomerative Hierarchical Clustering with Euclidean distance and using a Feature-based approach. The time series includes Dengue cases from 2009 to 2013.

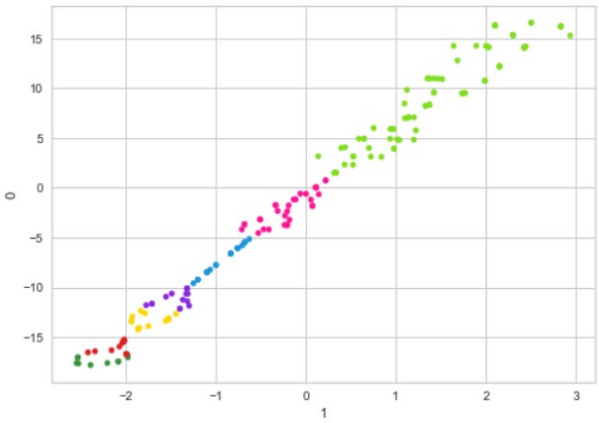

Figure 6. Clusters visualization with PCA shows the separation of the clusters.

6. Conclusion

In the first experiment, it can be observed that using a feature-based representation; better results have been achieved in all three clustering algorithms, standing out the Euclidean distance and DTW metrics. In the second experiment, by eliminating features, such as Cpoints and Mean, higher Silhouette Scores were obtained. When DBScan was tested, the more features were considered, the more the silhouette score deteriorated. These results indicate that feature selection can improve or worsen Feature-based clustering. So, feature selection is important to improve clustering. Hence, at this moment, motivated by the results, the authors are working to improve the DBScan to be more competitive.

7. References

Aghabozorgi, A., Shirkhorshidi, A. S. and Wah, T. Y., Time-series clustering – A decade review, *Information Systems 53*, pp. 16–38, 2015.

Bogado, JV; Stalder, D. H.; Schaerer, C. E.; Gómez-Guerrero, S, Time Series Clustering to Improve Dengue Cases Forecasting with Deep Learning,*2021 XLVII Latin American Computing Conference (CLEI)*, 2021, pp. 1-10, doi: 10.1109/CLEI53233.2021.9640130.

Bogado, JV; Stalder, D. H.; Gómez, S.; Schaerer, C. E. Deep learning-based dengue cases forecasting with synthetic data- Proceeding Series of the Brazilian Society of Computational and Applied Mathematics. v. 7, n. 1 (2020), 2020

Chen, J. R., Making subsequence time series clustering meaningful, *Fifth IEEE International Conference on Data Mining*, pp. 8 pp.-, 2005.

Gomez-Guerrero, S., et al. Construcción de un modelo de incidencia de dengue aplicado a comunidades de Paraguay, *Segundo Encuentro de investigadores*. Sociedad Científica del Paraguay. 2017.

Guha, S., Mishra, N., Motwani, R. and O'Callaghan, L., Clustering data streams. Symposium on Foundations of Computer Science, pages 359–366, RedondoBeach, CA, USA, 2020.

Hyndman, R. J., Wang, E. and Laptev, N., Large-Scale Unusual Time Series Detection, *2015 IEEE International Conference on Data Mining Workshop*, pp. 1616–1619, 2015.

Jolliffe, I. T. and Cadima, J., Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, 2016.

Keogh, E. and Kasetty, S., On the need for time series data mining benchmarks: A survey and empirical demonstration. data mining and knowledge discovery, *Data Mining and Knowledge Discovery,* vol. 7, pp. 349–371, 2003.

Minitab Express Support. Available: https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/, Accessed February 10, 2022.

Oates, T., Identifying distinctive subsequences in multivariate time series by clustering. *International Conference on Knowledge Discovery and Data Minin*g, pages 322–326, San Diego, CA, USA, Aug 15-18., 1999.

Özkoç, E. E., Clustering of Time-Series Data. *Data Mining - Methods, Applications and Systems*, 2020.

Pradeep, R. and Singh, S., A Survey of Clustering Techniques. *International Journal of Computer Applications,* 2010.

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa L. F., et al., Clustering algorithms: A comparative approach. *PLoS ONE 14(1)*, 2019.

Tavenard, R. An introduction to Dynamic Time Warping. Available: https://rtavenar.github.io/blog/dtw.html, June 10, 2022.

Wang, L., and Li, H., Clustering algorithm based on grid and density for data stream, *AIP Conference Proceedings 1839*, 2017.

Biographies

M. Giohanna Martínez is attending Asuncion National University, studying Mechatronics Engineering. In 2020, was chosen by the Erasmus+ program for academic mobility at the Complutense University of Madrid, Spain. In 2021, was selected as Paraguayan representative for the "Nobel Prize Dialogue: Latin America and the Caribbean". Currently, she is collaborating with the Computer Science and Mathematics Research group, applying clustering algorithms in order to find a hidden structure in dengue cases data.

Diego H. Stalder received the BS degree in Electronic Engineering (2010) from the Engineering Faculty of Asuncion National University (FIUNA), the Master (2013), and the PhD degree (2017) in Applied Computing from the National Institute for Space Research, Brazil. Since 2019, is a full-time researcher at FIUNA, Paraguay. His research interests include time series analysis, deep learning models, bio-metric signal processing, and instrumentation.

Christian E. Schaerer has graduated from the National University of Asuncion - UNA (1995), PhD (2002) from the Federal University of Rio de Janeiro. Post-doctorate in Applied Mathematics (2003) and Associate Researcher at the Institute of Pure and Applied Mathematics - IMPA (2007) - Brazil. From 2008 to date, he has been DITCODE Research Professor at UNA. He is interested in the simulation and control of large systems (scientific computing), differential equations and domain decomposition (parallel computing), and mathematical models for problems based on conservation laws (Petroleum, Biology). He is the coordinator of the Research Group in Scientific Computing and Applied Mathematics of the Nucleus of Research and Technological Development (NIDTEC) and the Laboratory of Scientific and Applied Computing (LCCA) at the Polytechnic School at UNA.

Juan V. Bogado is a Computer Systems Engineer from the National University of Caaguazú and master's in computer science from the UNA. He is pursuing a PhD in Computer Science from the UNA. University professor and researcher in the areas of time series analysis, clustering, deep learning models and epidemiological models.