Journal of Applied Research and Technology

www.jart.icat.unam.mx

Journal of Applied Research and Technology 21 (2023) 484-495

Original

Predictive models in pandemic times and their impact on the analysis of crime

G. Silva Atencio[a]* • M. Umaña Ramírez[b]

[a]*Universidad Latinoamericana de Ciencia y Tecnología (ULACIT), San José, Costa Rica*
[b]*Universidad Católica de El Salvador (UNICAES), Santa Ana, El Salvador*

Received 03 16 2022; accepted 11 25 2022
Available 06 30 2023

Abstract: Through the descriptive analysis of the open data of the Poder Judicial de Costa Rica, alarming results are reflected in the number of complaints imposed in the Organismo de Investigación Judicial (OIJ), exceeding fifty thousand complaints in 2019. Based on those numbers, the objective of this research is to generate a data analysis model that allows to potentiate of these statistics and to indicate in advance the regions with the most remarkable propensity to suffer crimes in the next five years, to promote the proactivity of both the citizen and the police to be alerted and to avoid upcoming crimes. Statistical prediction models are used to prove mathematical methods applicable to the data obtained and their behavior during 2015-2019. The analysis reflects the need to apply the simple linear regression algorithm to the developed solution available to all Costa Ricans on the Tableau public website. The results show pessimistic predictions for the country, especially in the Gran Area Metropolitana (GAM); the behavior of crimes will significantly impact this area, which indicates the need to establish police strengthening programs and improvements in education and employment to counter the potential crimes projected for the next five years.

Keywords: Crime Prediction, Preventive Patrolling, Police Statistics, Crime-Fighting

*Corresponding author.
E-mail address: gsilvaa468@ulacit.ed.cr (G. Silva Atencio).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

1. Introduction

In the 2010s, citizens' perception of security was a high-profile issue on the political agenda and in the Costa Rican electoral framework (Madrigal, 2012). However, today, citizens raise other matters of a more urgent nature. This research indicates whether security has improved in the last five years, using police statistics openly published on the Poder Judicial de Costa Rica website as a data source. In addition, a technological solution is proposed that strengthens society and transforms the shortcomings identified in 2021.

According to the Estado-de-la-Nación (2021), the net number of complaints received in the justice system between 2015 and 2019 a significant amount is shown; for example, in 2019, there were more than fifty-seven thousand complaints annually; however, in 2016, more than fifty-eight thousand. In addition to this, the same source shows the rate of intentional crimes against life per hundred thousand inhabitants between 2000 and 2009, with an average of 223 cases per year, while between 2010 and 2019, the same rate reflects an average of 249 crimes per year, showing a growth of approximately 11.7% in the 2010s.

One of the main factors to evaluate this issue is the work carried out by public organizations to fulfill their moral duty of accountability through opening data and information and the civic responsibility of supervising government institutions in their actions. Every day, an example of how open data provides citizens with tools to qualify the work of these instances.

The selected data set allows for geographical statistical analysis, by gender, nationality, and history, among others, making it possible to correlate various qualitative factors to accompany the quantitative analysis and thus illustrate the differences in data segmentation. However, analyzing crime trends in the country's regions is possible by triangulating other sources and applying statistical prediction models and data analysis.

Currently, technology has been allied to statistical models for data analysis and, in this way, automates tasks that would take a lot of time and resources. This work identifies correlational analysis models and technological tools to create a solution that supports proactive police and citizen action decision-making through predictive projections regarding crime in Costa Rica.

To do this, open data from the Poder Judicial de Costa Rica is used to forecast potential crimes in the areas with the highest crime rates in the country over five years. The first hypothesis to be analyzed in this research is whether the crime rate in Costa Rica between 2015-2019 is higher in urban cantons than in rural ones, concentrating a more significant number of crimes in the Gran Area Metropolitana (GAM). Two other hypotheses to be tested are whether crime in Costa Rica has remained the same or increased in the last five years, even though the perception of citizen security remains unchanged. On the other hand, the third hypothesis is that citizen security will be affected in the next five years, sustained by crime.

This article aims to answer the hypotheses mentioned above by extracting and analyzing relevant data that allow civil society to collaborate with the institutions in charge of citizen security. A data analysis model is proposed that potentiates the police statistics of the open data website of the Poder Judicial (OIJ, 2021) and indicates in advance the areas with the most remarkable propensity to suffer crimes.

As mentioned, the data from police statistics make it possible to obtain historical information on crimes and sub-crimes in Costa Rica by gender, canton, and type of victim. The information analysis was conducted in 2015-2019 based on the latest statistics published on the platform. The canton de Río Cuarto was taken from its foundation in April 2018 (Gaceta, 2021); there was no data for the area before this date.

2. Materials and methods

This article approaches the research from the quantitative perspective, generating a data analysis model that allows triangulating the metrics present in the open data site of the Poder Judicial de Costa Rica, (OIJ, 2021) and the Ministerio de Planificación (MIDEPLAN) regionalization data to predict the propensity of crime in a specific area of the country, to help citizens to visualize and understand the evolution of crimes in the regions of the country.

The information used to analyze and develop the article is available to all citizens on the Judiciary's website (OIJ, 2021). Public data can be found in various formats, including Excel, JavaScript Object Notation (JSON), and .txt format. The information pertinent to the regionalization, cantons, districts, and the number of inhabitants per square kilometer is documented by the MIDEPLAN. This data, unlike that offered by the Judiciary, is found in maps in Portable Document Format (PDF) and Excel documents as finished tables, so a process of collecting and cleaning them is required, with the involvement of the manual part or processes more complex automated.

A trend longitudinal non-experimental investigative design uses Tableau 2020.4 tools for data visualization, Tableau public to provide the results to the public, and the Python programming language in version 3.9.7, for predictive analysis through the linear regression algorithm.

The analysis follows the stages mentioned by Liebowitz (2020), starting with data collection; after obtaining and cleaning them, they were used in the Tableau 2020.4 tool to unite the information and show the data analysis visually and dynamically, these being published on the Tableau public site as a tool for the open consumption of the public. The predictive analysis is carried out using Excel and the

programming of an instrument with the Python language; these yield results that allow us to answer the prediction questions raised in the article. To generate these predictions, the data's behavior is analyzed to propose the most suitable prediction algorithm and thus obtain results with a confidence level greater than 65%. It is concluded that the linear regression algorithm is the one that best adapts, according to the behavior of the data, continuously over time to the number of crimes perpetrated.

3. Theoretical framework

This research uses open data as the basis of its analysis, understood as available data "that can be used, reused and redistributed freely by anyone, and that is subject, at most, to the requirement of attribution and sharing in the same way in which they appear" (ODH, 2021).

Open data has become a tool that allows citizens to monitor the government's actions. In 2015, a group of governments around the world, together with citizen organizations, coined this concept in the ODC (2021), where six principles are defined that correspond to this philosophy:

They are open by default; they are timely and complete, accessible and usable, comparable and interoperable, improve governance and citizen participation, and, finally, are used for inclusive innovation and development.

Costa Rica is part of the 85 nations that have adopted these principles since October 2016 (Zúñiga, 2016); however, according to the most recent report of the Economic Commission for Latin America and the Caribbean (CEPAL, 2018), the country is one of the few in the isthmus without legislation on access to public information, in the same group as Bolivia, Cuba, Haiti, and Venezuela. Thanks to the constant action of civil society organizations, Executive Decree 40199, published in the Gazette, San José, Costa Rica, Friday, May 12, 2017, indicates that the central government must implement the access to information section on their digital platforms.

Since then, institutions such as the Poder Judicial de Costa Rica have tried to provide citizens with transparent and easy-to-use information, creating an Open Data portal that offers multiple statistics ranging from institutional budgets and salaries to information related to domestic violence and femicides and other relevant statistics. In this way, the institution transfers to civil society the responsibility of using and consuming this information to monitor and assist the government with the analysis of institutional data and provide tools that enable the solution of problems faced by Costa Rican society.

Among the options provided by the open data portal of this entity, many police statistics stand out, providing information on crimes committed throughout the country since 2015, including crimes such as thefts and cross-outs on vehicles, assaults, homicides, etc. These statistics indicate the date on which the crime was committed, the gender and nationality, and whom the victim was, all confidentially, without including names, identity cards, license plates, or any other information that allows deducing the parties' identity.

Other pertinent data includes the division by province and canton so that the geographical point where the crime was perpetrated is determined. Like any data source, it has limitations; for example, in this case, the hours at which the crime was committed are omitted, and the age of the parties; additionally, the Police Statistics User Manual (OIJ, 2018). indicates that: "The data exposed in the consultation take as a source the complaints filed directly with the OIJ; in addition, they do not include complaints from prosecutors or other police; and, finally, the date that is taken as a reference for the computation of the crime (with few exceptions), is the date of the act and not the date of the complaint." Another limitation is not the available age range; instead, the study uses categories in the data: legal age, older adult, minor and unknown.

It is also important to define three key concepts that this data source uses: The first concept is a crime, a category of police statistics used to indicate the criminal offense perpetrated at the macro level. They are divided into assault, theft, robbery, vehicle strikeout, vehicle theft, and homicide. According to the Police Statistics User Manual, "a police crime category (for example, assault) can become a homicide time later (hours, days, weeks or months inclusive), due to the death of the person, as a result of the injuries inflicted" (OIJ, 2018).

The second is the sub-crime, a subcategory of crimes that can be expanded in detail to a predetermined description of the crime committed. For example, for assaults, there are sub-crimes: knife, firearm, outburst, blows, and immobilization, among others." All are accompanied by common characteristics that group the reported crime.

Finally, the third and last concept is the victim, which indicates that "it can be a natural or legal person, among others. However, it is attractive to the police to know who or what the offender is directing his activity towards; for example, for the query, the victim can be either a person or a material asset, whether movable or immovable.

The next step is to use a regionalization technique that makes it possible to illustrate a national strategy, as is often used by statistical studies by entities such as the Instituto Nacional de Estadística y Censo (INEC) in Costa Rica. The entity in charge of defining this regionalization is the MIDEPLAN, elaborated with the objective of grouping districts that meet similar topographic and socioeconomic characteristics, which allows focusing action plans according to each zone. This division will be one of the pillars of the study; below is the list of the six regions to be analyzed: Central, Brunca, Chorotega, Huetar Norte, Pacífico Central, and finally, the Región Huetar Caribe.

This issue is not unrelated to the work of multiple government institutions, and although it sometimes goes unnoticed, as shown by the latest report of the Security Perception Survey in Costa Rica (Mora et al., 2020), "68.9% of the population considers that they do not live safely in the country, and 31.1% consider that they do. In other words, according to this source, the number of people who feel insecure doubles that of those who feel safe", as pointed out by studies such as the one carried out by the Escuela de Estadística de la Universidad de Costa Rica (UCR) (Madrigal, 2012), where 59% of those interviewed perceived an increase in crime in the nation.

Regarding demographic issues of the urban or rural location, the difference is barely one percentage point between the two, with the perception of insecurity being more prominent in the metropolitan area, but without significant differences that motivate us to focus the study on a comparison of urban crime-rural, as Mora et al. (2020) details specific differences in the perception of security according to gender. Women show a greater sense of insecurity, as is usually expected in this study, due to street harassment.

The current affairs survey carried out in a pre-election setting seeks to understand, in addition to security, the feelings that moved the population before the 2014 elections. In 2010, the security issue was an essential point of discussion among politicians. Based on Mora et al. (2020), "Despite the recent decrease in insecurity in the country, the perception of citizens is that during the last three years, it continues to increase."

Madrigal (2012) expresses from the analysis; that it can be inferred that regardless of the work carried out by the Ministerio Público, Poder Judicial, and Ministerio de Justicia y Paz, in terms of prevention programs and enforcement policies, the perception of insecurity is due to other factors such as media, personal experiences and acquaintances, and recreation environment where it unfolds; As the Instituto de Estudios Sociales en Población (IDESPO) report mentions, "perception is a process of individual construction, but elements of a group nature also have an influence" (Mora et al., 2020).

This study uses the information provided by Poder Judicial based on the reports of crimes perpetrated in the last five years. It thus has an objective perspective the next time that national and international media refer to concepts such as the perception of insecurity compared to the number of crimes committed.

To introduce the context on which the research is based, a theoretical approach to the concepts part of the analysis is established. Álvarez (2017) highlights drugs as one of the primary motivators to carry out crimes and the possession of weapons as a tool that has gained popularity in recent decades for criminals.

As Sánchez and Muñoz (2017) expose, starting in the 1980s, Costa Rica took its first steps to strengthen society through different strategies to prevent crime. A National Development Plan was established during Oscar Arias Sánchez's (1986-1990) government, where two objectives stand out. The first aims to strengthen and consolidate preventive programs in security. The second focuses on formulating comprehensive prevention projects, significantly impacting the youth in areas where certain pathological behaviors are identified. The following governments propose improvements to this National Plan based on this initiative, always considering factors that favor proactive actions against crime.

On the other hand, according to Fonseca (2020), there are some references to using technological tools to address situations such as the one exposed; such is the case of the InterAmerican Development Bank (IADB). This institution has tested Geographic Information Systems, Big Data Systems, and Closed-Circuit Monitoring Systems (CCTV) to support police strategies for planning, crime prevention and investigation, and prosecution of suspected criminals.

Emerging technologies are leading the way to the fourth industrial revolution. Arteaga (2018) indicates that despite its implementation in countries, organizations, or people, it requires effort, time, and resources; it also offers an excellent opportunity for growth and improvements due to the innovative impact in strategies and processes.

Data analytics, machine learning, and cloud computing are among the technologies that support this research; as Liebowitz (2020) mentions, data analytics and artificial intelligence work hand in hand towards benefiting science, health, the economy, manufacturing, and citizen security, among others.

Regarding data analytics as a fundamental pillar towards generating predictions, Rábade-Roca (2018) exposes a concept implemented in a city such as Los Angeles, California, known as "prediction patrolling." This is based on massive data collection and analysis, where crucial information is generated for the police, indicating what crimes are likely, where, and when, in an area of approximately 45 square meters, in such a way that police surveillance is intensified in these areas. As a result of these actions, crime has been reduced by 13%, which is significant in a territory of more than 1.3 million people.

There are two essential concepts that Rábade-Roca (2018) also highlights; one of these is "problem-oriented policing," and the other is "intelligence-based policing." The first concept is oriented towards strategies where the police attack recurring problems through community policing cooperation with other institutions such as schools, civic organizations, and community centers. The second concept highlights information's importance in generating knowledge that supports the authorities' decisions. It should be noted that both problem-oriented policing and intelligence-led policing

are not mutually exclusive; the latter supports community and surveillance activities to target the most problematic areas with greater ownership and foundations.

In the field of data analytics is essential to explain other concepts. Data science comprises specific tasks for data manipulation, leading to a final result, with data preparation for further analysis. According to (Liebowitz, 2020), the steps of this concept are generation, collection, processing, management, recovery, and research. These actions aim to generate clean data that solidifies the analysis and generation of results. Clean data is a process where missing, repeated, or conflicting data with other available data is identified.

When the cleanup process is complete, the data scientist profile appears. This collaborator extracts everything valuable for the analysis; the scientist also identifies unavailable elements that could affect it. These activities are not easy; fortunately, artificial intelligence is a tool that has advanced on a large scale in recent decades and provides crucial support for this work. It can analyze and process standardized and non-standardized data, extract features and patterns, categorize them, and store them for later retrieval when necessary.

Going deeper into artificial intelligence, Liebowitz (2020) refers to the first ideas of mechanizing human thought, promoted by Aristotle and Euclid. Of course, the technology that supports such statements turns them into a reality. Between 1956 and 1970, computers helped implement rule-based managers and expert systems. These computers generated what is known as symbolic reasoning, based on the supervision of engineers who provided knowledge and validated the results produced. This period marks the first wave of artificial intelligence, turning into the reality of the idea of Aristotle and Euclid of mechanizing human reasoning.

One branch of artificial intelligence is known as automatic or machine learning. Shalev-Shwartz and Ben-David (2014) clarify their purpose and importance. The complexity of the problems and the need to adapt to a particular environment or situation are two aspects that contribute to a system or machine learning from previous experience. This experience or prior knowledge is a fundamental input for decision-making. It is necessary to highlight the importance of this learning through systems and machines since there are tasks that would take a long time for human beings to ensure accurate results without errors. Some examples to consider are analyzing large amounts of data, predictions, searches, and extracting valuable information automatically.

In the bowels of machine learning, there is a vast amount of technology and mathematical concepts in statistics; these collaborate to make predictions a reality. Clarke and Clarke (2018) rightly expose the need to create forecasts. They explain that its purpose is due to the certain need-to-know circumstances that could happen and make decisions based on solid data that indicate with a high probability that they can be avoided or, on the contrary, execute an action in advance that favors a specific objective or strategy, whether personal, group or organizational.

Behind predictive analysis, as exposed by McCarthy et al. (2022), there is a set of advanced statistical models and machine learning techniques based on historical data. For predictions to be compelling, this historical data input must be representative. To achieve this, it is necessary to clean, analyze and prepare the data so that the noise that can reduce the effectiveness of the developed solution is eliminated.

Algorithms are a set of steps that solve a problem. This concept also applies to predictive analytics, divided into supervised and unsupervised learning algorithms (McCarthy et al., 2022). The first techniques are used where the automated system's data are correct, labeled, and validated by an observer. In contrast, the data is not marked in the case of the second algorithm. Therefore, the system must find patterns, categories, and characteristics that generate that learning without someone or something that corrects it. The information produced is new to the observer.

Another relevant concept for this work is cloud computing, which, according to AWS (2021), one of the pioneers in this technology, consists of the distribution of technological resources that are used on-demand, or that is, whenever the user requires it at the time and place that they need it. To use it, the user must have an internet connection and pay for its use according to the frequency established by the provider. Data storage capacity, computing capacity, and computer equipment availability can be mentioned in this distribution of technological resources.

One of the reasons cloud computing is mentioned in this work is the benefits it brings to analyzing data, preparing it, projecting predictions, and publishing the solution easily and quickly. Some of these benefits are the agility with which technological tools can be easily used to facilitate working, cost savings by not buying high-technology equipment, or hiring collaborators responsible for the solution's maintenance, control, and support. In addition, the elasticity of the cloud solution is also highlighted since, according to demand, its capacity and availability behave in this way.

4. Analysis of results

Once the theoretical research was carried out, the data from the Judiciary Department was analyzed to understand the behavior of crimes and, in this way, establish a predictive mathematical model that projects information on potential crimes in the next five years in each region, as well as the monthly increase or decrease during this same period.

The data for the analysis goes from January 2015 to December 2019. Due to the atypical behavior that hit the world at the beginning of 2020, this year is excluded from the study

so as not to generate anomalies in the results since the total and partial closures that Costa Rica has had both in places of recreation and transportation since March 2020, may have influenced a significant decrease in the number of crimes committed in the nation, reducing them by almost half, in contrast to the previous year. In the same way, this research seeks to generate a model that can be replicated in conditions like those that occurred before the pandemic; therefore, the results obtained are based on what is projected for a way of life without health restrictions, pandemics, or other significant events that affect the daily lives of Costa Ricans.

Following the steps, Liebowitz (2020) explained that the field of data analytics, particularly data science, began with the collection of data. In the case of this research, the data from the website of the OIJ (2021) was downloaded. Process and management steps involved cleaning and changing activities and eliminating incomplete, repeated data that did not provide relevant information to the study.

A benefit of data analysis is the possibility of generating predictions based on historical information; in addition, regardless of the forecasting methodology selected, it is always possible to carry out the following steps: project definition, including guidelines and objectives; exploration, where the method to collect information and its range is determined; data preparation, in which data is processed for the study; model building, where metrics are created and evaluated; implementation, where the results are applied to the models; and finally, management, where the necessary improvements for the continuous evolution of the predictive model are evaluated.

Taking advantage of the open data provided by the Judiciary Department, the Police Statistics between the years 2015-2019 were used to project how the actions of crimes in different segmentations, such as region, canton, age, and gender, as well as evaluating the current situation using descriptive analytics. As a final product, Tableau public provides an available tool for a free consultation by the person visiting the website. This analysis can materialize in preventive actions carried out by the different entities in charge of security, from the Municipal Police or the Public Force, carrying out restraints or deploying their units in places of greater risk, making better use of resources and police personnel.

An essential part of the research has data that allows proper analysis. For this study, information was used from the open data crime section of the Poder Judicial de Costa Rica on criminal acts that occurred in the country between 2015 and 2019. Concerning segmentation, the variables contained in these data are used to determine the province, canton, and district where the crime was committed, the type, the gender of the person who committed it, the age range of the alleged offender, and the kind of victim. This information is enriched with the segmentation used by MIDEPLAN to regionalize the country's districts and categorize them into urban-rural, to perform more relevant macro analysis; the primary source for taking this information is the geographical classification manual for statistical purposes of Costa Rica, developed by the Instituto Nacional de Estadística y Censos (INEC, 2016).

The information obtained by downloading data online from the website is used to create the analysis matrix and apply the study correctly. The data goes through cleaning and preparation to be joined appropriately. For example, while the information provided by the INEC names a district as "La Fortuna," the Poder Judicial mentions it as "Fortuna." Other problems were canton names marked in some lines and others without keeping; incorrectly described pronouns in some cases. Some words were also found extensive in their abbreviated form, among others. This same pattern was present when reviewing data for the districts. The most significant inconvenience occurs because the Judiciary Department loads data from various places and needs to apply filters or cleaning before executing it. This process is carried out in the five Comma Separated Values (CSV) files of the Police Statistics of the Poder Judicial, one per year. These are consolidated in a final master file to unite it with the INEC information. After downloading, cleaning, and reducing the data sources, the postal codes are added to facilitate this homologation concerning the cantons and districts' names to send to the authorities with minor errors.

Regarding the geographic data provided by the INEC, there is a data source in PDF, so the information must be extracted to a tabular file, preferably CSV. This work was carried out with a web tool called iLovePDF (2023), which allows converting files of this format to Excel. As a result, the Excel file has several spreadsheets that must be cleaned and consolidated into a single source. The most common errors in this part are duplicate spaces or before beginning or ending a name; there are also rows with misaligned column names, mainly affecting each spreadsheet's initial and final line. Once this process is done, it is consolidated into a single sheet without formatting to be stored in CSV format.

The databases are joined in the tools selected for analysis, using the default functions for this task. On the market, there is a considerable number of ways to process, analyze and visualize data; for this analysis, several tools were selected that, when complemented, allow us to provide a solution more in line with today's world, where most of the effort is to create a visually attractive solution for the consumer: Microsoft Excel for cleaning and data consolidation, Tableau Desktop, was used for descriptive analysis and data visualization, Python for predictive analysis.

Descriptive analysis is used in statistics to describe what has happened; the categorical variables are taken, and studies are carried out according to the different segments. This analysis allows us to answer questions in this research. Figure

1 shows that the crime rate in Costa Rica between 2015-2019 is higher in urban cantons than in rural ones, concentrating a more significant number of crimes in the GAM. Two other questions to be tested are whether crime in Costa Rica has remained the same or increased in the last five years, even though the perception of citizen security remains unchanged. On the other hand, the third question is whether citizen security will be affected in the next five years, sustained by crime. This study is essential to understand the information with which one works deeply; it also allows an understanding of the current situation and indicates which areas are desired to be investigated further.

We also have prescriptive and predictive analytics. This research aims to develop predictive analytics to answer questions about the propensity to commit a crime in each region. The predictive analysis takes historical information and applies statistical models. The most suitable is linear regression, a mathematical model used to approximate the dependency of the relationship between a dependent variable and independent variables. The key to using predictive models is understanding an estimate and knowing what is not exact since there is always a margin of error (Triola, 2004).

Tableau allows a union of data sources, with a simple pull and drag of the files and selecting the variable (or set of them) joint between the two databases; in this case, the chosen field is the district code or zip code. The descriptive analysis was carried out in this tool, generating an interactive analytical dashboard that allows evaluating the different segmentation types and quickly seeing the historical changes by simply applying filters or dragging new dimensions into view. This dynamic dashboard is accessible to anyone who can consult it on the Tableau public site (Aviles, & Coto, 2021).

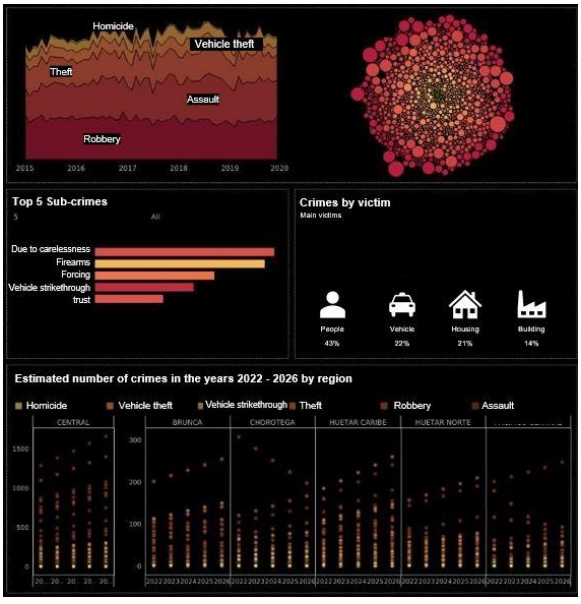

Figure 1. Crimes in Costa Rica 2015 – 2019.

Other study data was processed using Python, one of the most used languages by analysts and data scientists to study large amounts of information. A sweep of the Excel data is performed using a program created specifically for prediction. For the creation of the analysis program in Python, there is a library called sklearn (Scikit-learn, 2021), which contains many statistical tools: regression and prediction (it is not recommended to manipulate data or obtain summaries).

The data analysis process through Python is moved to Tableau, creating graphs that can be filtered more efficiently, allowing the information to be taken to more granular levels and generating a more visually attractive design.

In addition, two functions are used for prediction in the Excel tool: slope and intercept. The slope function, which stands for pitch, calculates the linear regression slope when given two data points on the X-Y axes. The intercept function is then used to fit the prediction functions; it could be said that it is an extension of the diagram to a future point where the variables join, which predicts continuous data until its intercept.

A 0.65 or 65% is required to make projections with greater mathematical robustness in predictive analysis. The equation for linear regression given by $y = mx + b$ generates more efficient predictions. Where y is the number of crimes to predict, m; is the value of the slope that shows whether the relationship of the data is positive or negative, and b; is the value of the coefficient of determination that indicates the effectiveness of the mathematical expression in predicting future outcomes of the variable y.

The first results show 292,315 different crimes over five years, distributed almost evenly each year, with an approximate average of 58,000 events per year, except for 2018, when there were more than 61,000 events. Divided into six main categories, these almost 300,000 crimes are distributed: 32% theft, 30% robbery with violence, 21% robbery without violence, 9% vehicle tag, 8% vehicle theft, and 1% homicide. Regarding provincial geographical division, 38% of the events occurred in San José, Alajuela with 16%, followed by Puntarenas with 12%, Heredia and Limón with 9% each, Guanacaste behind with 8%, and Cartago with 6%.

The victims are primarily people, representing 43% of all incidents, while vehicles and homes account for around 21% and 14% correspond to buildings. The perpetrators are primarily male (67%), adults (86%), minors who committed 4% of the crimes, and 5% by older adults.

Understanding the data distribution is necessary to decide which ones to leverage for deeper predictive analytics. For example, a predictive analysis for the next homicide will not return a reliable figure due to its high dispersion.

On the other hand, enriching this information with data such as regionalization makes it possible to point out specific behaviors in certain country areas. Taking Heredia as an example, where more than half of its territory is in the Región

Huetar Norte (Sarapiquí canton) and the rest in the Región Central, it can be inferred that, due to their socioeconomic conditions, both spaces have very different behaviors, despite belonging to the same province.

A granular level by canton or district could hide these differences. Except for the canton Central de San José and Alajuela, sufficient information is lacking for a meaningful predictive analysis. These phenomena are known as outliers and can lead to an erroneous interpretation if taken as usual within the data distribution. In the descriptive analysis by region, thefts are the most common crimes, except in the Región Central and Huetar Caribe, where assaults predominate. Even in the other four areas, before assaults, robberies occupy second place. Additional noteworthy information is that the place that occupies the highest number of vehicles: Pacífico Central, is significantly higher than its coastal peers; in fact, the available total of vehicle strikeouts is 15%, while in the other regions, it ranges between 4% and 9%.

Figure 2 visually summarizes the above information, where the bars show the number of crimes ordered from more significant to lesser. At the same time, the colors represent each of the regions and the proportional interference they have in each of these incidents.

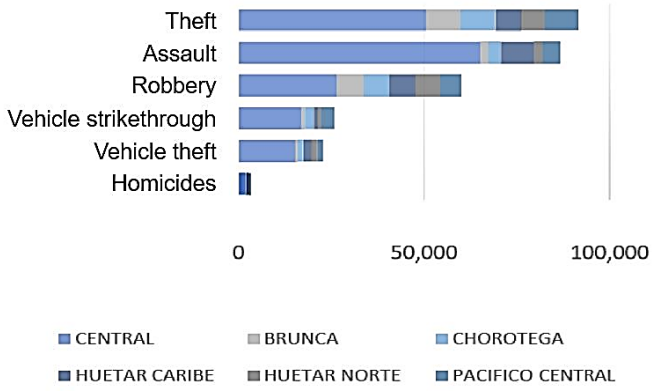

Figure 2. Crimes by region, 2021.

When the temporal analysis is carried out to understand the evolution of crimes over time and highlight any difference in behavior, it stands out by region: in the Región Central, assaults of five percentage points between 2015 and 2019, rising from 34% to 39%. In Brunca, in 2015, thefts represented half of the crimes, decreasing to 42% at the end of the five years, while robberies increased. In Huetar Norte, thefts predominate, followed by thefts, a behavior that remains stable in all periods, except in 2016, where thefts decreased by 5%, leaving thefts in first place with 38%. However, in 2017 it returned to its normal state. The other detail that can escape the eye is an increase in vehicle theft, from 5% to a gradual rise, reaching 12% in 2018 and 11% in 2019.

The Región Chorotega shows increased assaults, from 12% to 17% at the end of the study period, while thefts decreased from 46% to 38%.

The Región Huetar Caribe is one of the most particular cases since assault, robbery, and theft tied at 30% at the beginning of the analysis period. Little by little, the percentage of assaults increased to 38%, and robberies and thefts decreased by 23%. Finally, in the Pacífico Central, there is a downward trend in thefts, which start at 42% and end at 33%; assaults and vehicle robberies increase in opposition to this distribution.

The predictive analysis carried out in each region aims to follow the behavior trend of each region, supported mainly by the notable characteristics of the above descriptive analysis. The years for which this analysis is projected correspond to the following five-year period: from 2022 to 2026.

Table 1 shows the confidence coefficients or percentages obtained from the projection model, crime, and region. This value is obtained from the linear regression of the model executed in Python. As mentioned above, a minimum coefficient of 0.65 is used to trust the model. The higher the percentage, the greater the probability that the model is correct in its propensity.

Table 1. Prediction Confidence Coefficient by Region and Crime, 2021.

Crime	Central	Brunca	Chorotega	Huetar Caribe	Huetar Norte	Pacífico Central
Assault	84.2%	82.0%	88.0%	82.5%	84.2%	86.6%
Homicides	96.2%	96.0%	98.5%	92.9%	92.2%	93.9%
Theft	83.2%	77.7%	80.8%	86.0%	83.1%	86.8%
Robbery	82.4%	84.1%	76.5%	89.4%	85.0%	85.1%
Vehicle theft	86.0%	86.2%	93.5%	82.9%	77.4%	84.8%
Vehicle strikethrough	89.0%	95.5%	87.0%	97.0%	94.8%	89.0%

From the projections, the trends that attract attention are highlighted, based on the data from 2015 to 2019 show us and information that is noteworthy to consider. We can point out that we developed a study based on the trend to complement the model and linear regression analysis. As mentioned in the previous table, the focus remains on socioeconomic regions and types of crime.

In the same line of analysis, we have that the Región Central projects an increase in assaults from 9,000 in 2022 to 11,000 in 2026; this means a slight downward trend, compared to the historical, with an approximate average of 12,000 assaults per year, but going back up over time; It should be noted that there is a projection of a slight increase in homicides, from 114 to 163 from 2022 to 2026. This example of homicides is not an alarm that historical data shows us; however, the projection and its high coefficient highlight it as a point of future attention.

The Región Brunca confirmed a downward trend in thefts, with a confidence percentage of 78%, which is low compared to the other estimates; however, it falls within the defined margin. The crossing of vehicles is shown as one of the issues to be highlighted, with an increase from 124 events to 168 in the annual ranges mentioned.

In the Región Huetar Norte, an upward trend in vehicle theft is projected, confirmed by linear regression, going from approximately 400 to almost 500 in the following five years. Assaults tend to increase, while thefts tend to decrease; generally, the increase and decrease values do not generate a significant finding.

In the Región Chorotega, the trend shown in the previous five years is projected as follows: an increase in assaults and a decrease in thefts. Assaults increase from about 132 in 2022 to 181 in 2026, while thefts go from 639 to 440 in the same period. The projection has two exciting factors not precisely demonstrated in the historical data: a 30% increase in homicides in the following five-year period. The second corresponds to a 25% increase in vehicle theft.

In the Región Huetar Caribe, an increase in assaults is projected, but a decrease in robberies and thefts. Assaults show a rise of 30% in the next five years, while thefts decrease by the same amount; on the other hand, robberies are projected at 441 in 2022 to 392 in 2026, a gradual decrease. One of the crimes with the highest coefficient that this analysis showed is an increase in the cross out of vehicles of approximately 20% for the following five years.

Finally, the Región Pacífico Central continues to trend upward, projecting an increase from 12 to 17 cars affected by blemishes. Homicides are also a factor of concern where an increase of more than 40% is cast in the following period, being a confidence level above 0.9. Robberies and assaults also tend to go up. This region is generally outlined with increased danger for the next five years.

In addition to analyzing macro crimes by region, a temporal analysis was carried out to understand how monthly crimes may behave in the following years; according to the data recorded and using the same mathematical technique, a monthly increase in assault crimes in the country is established for the next five years in January, February, April, June, November, and December. On the contrary, a decrease in the number of crimes is projected for the months not included in the previous list. Figure 3 shows the predictions where the questions and details previously described are answered. It should be noted that this monthly analysis projects a predictive relationship between the number of future assaults with an accuracy percentage of 52.8%.

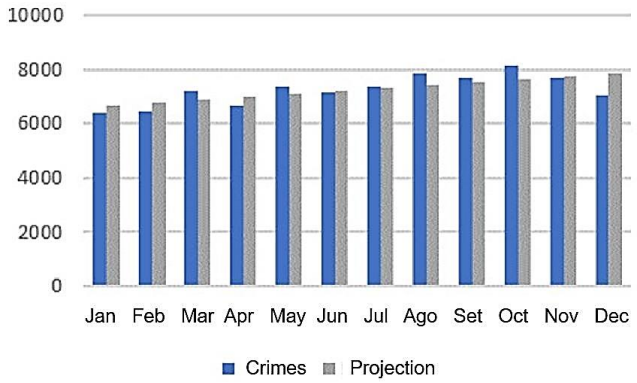

Figure 3. Projection of monthly crimes, 2021.

In this analysis, we highlight three months that are worth narrowing down. In December, an increase of close to 10% in the crimes reported in this period is projected, with the most significant increase in the violations expected for the following one-year period. However, not everything is bad news; in October and August, a reduction of crimes is expected by 6% compared to the period studied. This is an important point to highlight from the analysis resulting from the monthly projection model.

Regarding the study's limitations, one of the questions could not be answered with the available data: projecting the age range most likely to commit a crime in the next five years. However, the data only offers the following categories: legal age, older adult, minor and unknown. After the analysis made and making the respective correlation of these age categories concerning the number of crimes, it is concluded that the tendency of the victims is significantly high towards the elderly. In the case of thefts in 2015, there were 16,349 adult victims, while minors added up to 1,305 and older adults, 986. A similar behavior emerged in the following years analyzed.

Given this trend, the linear regression analysis generates a weak and scattered relationship, significantly affecting the prediction's effectiveness. Even so, predicting with a coefficient of 62.1% as obtained in the analysis, the contribution generated by the results is insignificant since it follows the trend of projecting the elderly as the most likely to suffer crimes. This prediction would yield a more precise value if the available data had ages in numbers or more detailed age ranges, for example, separated by years, five-year periods, or decades.

5. Conclusions and recommendations

Government institutions in Costa Rica must adopt best practices for transparency analysis, such as constantly publishing the data relevant to your institution. According to international agreements, citizens should have access to information that includes everything from the salaries of public officials to budget execution work plans, among others. In the case of the Judiciary, it attempts to comply with these guidelines by having this information and other relevant statistics available, such as data on femicides, domestic violence, and police statistics; however, without adequate communication, it is like shouting into the void. The first recommendation is to create the habit of government publishing entities, promote information, promote state transparency, and encourage civil society's participation in supervising institutions.

It is understood that data analysis goes through many stages of collection and cleaning; however, even though the products provided by the Judiciary and the INEC are helpful and complete, a rigorous cleaning is always necessary to connect the sources. This first step is essential in any analysis process because it also has as a side effect a closeness between the analyst and the data, understanding how each of the records interacts. After joining the data, verifying there is no duplication is imperative. Data analysis tools facilitate this task, which is why the ones selected for this exercise have been Excel, Tableau, and Python since their intuitive use speeds up this stage.

The predictive model uses a mixture of trend analysis with linear regression since using more than one propensity technique is recommended. To interpret the results more appropriately, the initial study of the five-year data is the basis for drawing conclusions that may be more relevant based on the behavior of past trends. However, it is not the only thing taken; data with a high margin of precision or confidence, or essential changes in the data, are also highlighted as part of the model results.

When the research is formulated, it seeks to take advantage of many of the segmentation data included in the database; however, it aims to expand the database, enriching the information using regionalization to provide deeper details on the behavior of localities with similar socioeconomic conditions, taking advantage of the excellent study that entities such as MIDEPLAN and INEC have carried out on the matter, since the georeferencing of traditional political division, composed by province, canton, and district, is not always the most accurate to analyze some social behavior in a zone, given the contrasts that the provinces and cantons have, depending on the area where they are located.

The descriptive analysis demonstrated the national reality regarding crime, according to the complaints filed in the OIJ. It is observed that Región Central, especially the cantons of San José and Alajuela, tops the list of insecure areas and that the behavior of this area differs significantly from the regions further away from the Central Valley. The socioeconomic phenomena of each region, the population, and housing determine factors when viewing this information, so it is not advisable to compare areas since each has its reality. One of the illustrated results calls the focus on the Región Pacífico Central and the cross-out of vehicles whose incidence exceeds the average; being clear about this prediction allows preventive action regarding car safety in this area. Another issue that should be noted is that homicides do not exceed 1% of crimes in Costa Rica, which is an excellent approach to the behavior of data; however, the Región Caribe is a particular case and doubles the number of national homicides from a national perspective percentage.

Therefore, looking at the historical data, it can be thought that there is a high probability of assaults in the Región Central, vehicle markings in the Pacífico, and homicides in Huetar Caribe, in percentage comparison of crimes between the regions. The Región Chorotega, together with Brunca, shows the lowest crime projections.

The predictive analysis sought to take each of the outstanding characteristics of each region and verify what the behavior will be in the next five years. In addition, other exciting data was found, highlighting an upward trend in homicides in various country regions, with high-reliability percentages. However, it is an estimate, and it is advisable to continue working with this model to improve.

During the analysis, some questions could not be answered due to the high dispersion of the data, such as information by the district. The segmentation does not generate a significant value, such as gender or age. One of the Judiciary's recommendations is collecting relevant information about the aggressor and the victim, such as the gender of both parties. An improvement that can generate much value is to include the exact age of the person committing the crime and the victim, if possible.

In general, during this investigation, it was decided to meet the proposed objectives since, in a short period, a predictive data model was developed that allows visibility of crimes by region and a descriptive data analysis product that illustrates the situation of crimes in Costa Rica, accessible to anyone who wants to research the topic.

6. Future research lines

Given the results obtained in the research and development of the practical work, it is necessary to implement a constant feed mechanism with quarterly updated data to keep the proposed solution active and updated with the model, providing valuable information to Costa Rican society. The automated tool must download the data generated by the Judiciary, clean it, and store it in a cloud repository such as Google Drive in CSV

format, where it is accessed by the predictive tool and, after the application of linear regression techniques, is projected through a public access link to people in Tableau.

In addition, it is suggested to involve data scientists and statisticians to develop a deeper mathematical analysis to evaluate improvements in the accuracy of the set and future predictions. These profiles help strengthen the predictive model through more rigorous tests based on more advanced mathematical foundations. This work will expose the effort required to answer questions such as the monthly increase or decrease of each crime by region of Costa Rica for the next five years.

Additionally, it is suggested to extract more information from the data publicly available by the different government institutions to show the actions that other institutions in Costa Rica are carrying out regarding security and possible integrations of institutional data that strengthen the solutions—prevention through technological tools with a predictive approach.

Conflict of interest

The authors have no conflict of interest to declare.

Acknowledgements

The authors would like to thank all those involved in the work who made it possible to achieve the objectives of the research study.

Funding

The authors received no specific funding for this work.

References

Álvarez, P. (2017). Violencia en Centroamérica: reflexiones sobre causas y consecuencias. *Anuario Latinoamericano–Ciencias Políticas y Relaciones Internacionales, 4*, 21. Retrieved from https://journals.umcs.pl/al/article/view/5411

Arteaga, F. (2018). La cuarta revolución industrial (4RI): un enfoque de seguridad nacional. Real Instituto Elcano. Website: *Documento de trabajo, 12*, 2018. Retrieved from https://www.realinstitutoelcano.org/documento-de-trabajo/la-cuarta-revolucion-industrial-4ri-un-enfoque-de-seguridad-nacional/

Aviles, N., F., M., & Coto, D. (2021). Delitos en Costa Rica 2015 - 2019 y Estimación 2022 – 2026. *Tableau Public.* Retrieved from https://pjenlinea3.poder-judicial.go.cr/estadisticasoij/

AWS (2021). *¿Qué es la informática en la nube?. Amazon Web Services*, Retrieved from https://aws.amazon.com/es/what-is-cloud-computing

CEPAL (2018). América Latina y el Caribe: Países que cuentan con Ley de Acceso a la Información Pública y año de promulgación.Retrieved from https://observatoriop10.cepal.org/es/mapas/america-latina-caribe-paises-que-cuentan-ley-acceso-la-informacion-publica-ano-promulgacion

Clarke, B. S., & Clarke, J. L. (2018). Predictive statistics: Analysis and inference beyond models (Vol. 46). Cambridge University Press.

Estado-de-la-Nación. (2021). Visualizador de Datos. Retrieved from http://estadisticas.estadonacion.or.cr/visualizador

Fonseca, H. (2020). El desarrollo tecnológico en materia policial: una receta de éxito para la prevención del delito. Revista de Relaciones Internacionales, Estrategia y Seguridad, 15(1).

Gaceta. (2021). Creación del cantón XVI Río Cuarto, de la Provincia de Alajuela. *Gaceta Digital N° 9440* Retrieved from https://www.colegiotopografoscr.com/comunicados/2018/creacioncanton.pdf

iLovePDF. (2023). Convierte PDF a EXCEL. Retrieved from https://www.ilovepdf.com/es/pdf_a_excel

INEC (2016). Manual de Clasificación Geográfica con fines estadísticos de Costa Rica, Nº 41181 - Plan. Retrieved from http://www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_texto_completo.aspx?nValor1=1&nValor2=86891

Liebowitz, J. (Ed.). (2020). Data Analytics and AI. CRC Press.

Madrigal, J. (2012). Resulados de la encuesta actaulidades 2012. *Universidad De Costa Rica*, Retrieved from https://www.ucr.ac.cr/medios/documentos/2012/UCR-ESTADISTICA-ACTUALIDADES-2012.pdf

McCarthy, R. V., McCarthy, M. M., Ceccucci, W., Halawi, L., McCarthy, R. V., McCarthy, M. M., ... & Halawi, L. (2022). *Applying predictive analytics* (pp. 89-121). Springer International Publishing. https://doi.org/10.1007/978-3-030-83070-0

Mora, C., Solano, M., Hernández, J., Rodríguez, I., & Hernández, K. (2020). Informe de Encuesta: Percepcion sobre la seguridad en Costa Rica.

ODC. (2021). Open Data Charter.. International Open Data Charter. Retrieved from https://opendatacharter.net/principles/

ODH. (2021). Open Data Handbook. ¿Qué son los datos abiertos? Retrieved from http://opendatahandbook.org/guide/es/what-is-open-data/

OIJ. (2018). Organismo De Investigación Judicial. Manual de Usuario: Sistema de Estadísticas del OIJ.

OIJ. (2021). Organismo De Investigación Judicial. Datos Abiertos.

Rábade-Roca, J. (2018). La Innovación Policial en la Ciudad del Siglo XXI. Retrieved from.https://www.ciudades-creativas.com/proceedings/6ccc/proceedings-6ccc_040.pdf

Sánchez, M., & Muñoz, S. (2017). Reconstrucción de la política de prevención del delito en Costa Rica. Retrieved from http://www.joinpp.ufma.br/jornadas/joinpp2017/pdfs/eixo7/reconstrucciondelapoliticadeprevenciondeldelitoencostarica.pdf

Scikit-learn. (2021). Scikit-learn Tutorials. Retrieved from https://scikit-learn.org/stable/tutorial/index.html

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press. https://doi.org/10.1017/CBO9781107298019

Triola, M. F. (2004). Probabilidad y estadística. Pearson educación.

Zúñiga, A. G. (2016). Adopción de la Carta Internacional de Datos Abiertos. *Bitácoras Políticas.com* https://www.xn--bitacoraspolticas-ovb.com/2015/10/firma-americo-zuniga-convenio-de-datos.html