Journal of Engineering and Technology for Industrial Applications

ISSN 2447-0228

October 2020 Volume 06 / No 25

Editor-in-Chief: J. C. Leite

www.itegam-jetia.org

ITEGAM - JETIA ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)

O **ITEGAM-JETIA: Journal of Engineering and Technology for Industrial Applications** is a publication of the Galileo Institute of Technology and Education of the Amazon (ITEGAM), located in the city of Manaus since 2008. JETIA publishes original scientific articles covering all aspects of engineering. Our goal is the dissemination of research original, useful and relevant presenting new knowledge on theoretical or practical aspects of methodologies and methods used in engineering or leading to improvements in professional practice. All the conclusions presented in the articles It should be state-of-the-art and supported by current rigorous analysis and balanced assessment. Public magazine scientific and technological research articles, review articles and case studies.
JETIA will address topics from the following areas of knowledge: Mechanical Engineering, Civil Engineering, Materials and Mineralogy, Geosciences, Environment, Information and Decision Systems, Processes and Energy, Electrical and Automation, Mechatronics, Biotechnology and other Engineering related areas.

Publication Information:
ITEGAM-JETIA (ISSN 2447-0228), (online) is published by Galileo Institute of Technology and Education of the Amazon on a every two months (February, April, June, August, October and December).

Contact information:

Web page: www.itegam-jetia.org
Email: editor@itegam-jetia.org
Galileo Institute of Technology and Education of the Amazon (ITEGAM).

Joaquim Nabuco Avenue, No. 1950. Center. Manaus, Amazonas. Brazil.
Zip Code: 69020-031. Phone: (92) 3584-6145.

Previous Notice
All statements, methods, instructions and ideas are the sole responsibility of the authors and do not necessarily represent the view of ITEGAM -JETIA. The publisher is not responsible for any damage and / or damage to the use of the contents of this journal. The concepts and opinions expressed in the articles are the sole responsibility of the authors.

Directory

ITEGAM - JETIA ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)

ITEGAM-JETIA. v.6, n.25. October of 2020. Manaus - Amazonas, Brazil. ISSN 2447-0228 (ONLINE)
https://www.itegam-jetia.org

SUMMARY

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.6 n.25, p. 4-15, Sep/Oct, 2020

DOI: https://doi.org/10.5935/jetia.v6i25.685

RESEARCH ARTICLE OPEN ACCESS

A SYSTEM PROPOSAL FOR AUTOMATED DATA CLEANING ENVIRONMENT

Carlos Roberto Valêncio[1], Toni Jardini[2], Victor Hugo Penhalves Martins[3], Angelo Cesar Colombini[4] and Márcio Zamboti Fortes*[5]

[1, 2, 3] São Paulo State University - UNESP. São Jose do Rio Preto–São Paulo, Brazil.
[4, 5] Federal University of Fluminense - UFF. Niterói–Rio de Janeiro, Brazil.

[1] http://orcid.org/0000-0002-9325-3159 [2] http://orcid.org/0000-0002-0686-6791 [3] http://orcid.org/0000-0003-4507-0128
[4] http://orcid.org/0000-0002-8906-4128 [5] http://orcid.org/0000-0003-4040-8126

Email: *mzamboti@id.uff.br

ARTICLE INFO

Article History
Received: August 12th, 2020
Accepted: October 19th, 2020
Published: October 30th, 2020

Keywords:
Data cleaning,
Semantic,
Knowledge Discovery in Databases (KDD),
Data analysis.

ABSTRACT

One of the great challenges to obtaining knowledge from data sources is to ensure consistency and non-duplication of stored information. Many techniques have been proposed to minimize the work cost and to allow data to be analyzed and properly corrected. However, there are still other essential aspects for the success of data cleaning process that involve many technological areas: performance, semantic and autonomy of the process. Against this backdrop, we developed an automated configurable data cleaning environment based on training and physical-semantic data similarity, aiming to provide a more efficient and extensible tool for performing information correction which covers problems not yet explored such as semantic and autonomy of the cleaning implementation process. The developed work has, among its objectives, the reduction of user interaction in the process of analyzing and correcting data inconsistencies and duplications. With a properly calibrated environment, the efficiency is significant, covering approximately 90% of inconsistencies in the database, with a 0% percentage of false-positive cases. Approaches were also demonstrated to show that besides detecting and treating information inconsistencies and duplication of positive cases, they also addressed cases of detected false-positives and the negative impacts they may have on the data cleaning process, whether manual or automated, which is not yet widely discussed in literature. The most significant contribution of this work refers to the developed tool that, without user interaction, is automatically able to analyze and eliminate 90% of the inconsistencies and duplications of information contained in a database, with no occurrence of false-positives. The results of the tests proved the effectiveness of all the developed features, relevant to each module of the proposed architecture. In several scenarios the experiments demonstrated the effectiveness of the tool.

I. INTRODUCTION

Storage of information has become increasingly larger and more frequent, since data may contain valuable information [1]. In order to obtain correct information from all possible or convenient sources, it is necessary to ensure that the analyzed data are integrated consistently and correspond to the reality of the information. This warranty is only possible if there is a data pre-processing; a step where the information stored in different sources is integrated and analyzed to detect inconsistencies, physical or semantic duplications, standardization and corrections, etc. One of the stages of data pre-processing is called data cleaning, which develops computational techniques for handling data issues and inconsistencies from the same source or even distributed sources. Only after that process is the data suitable for analysis and correlation [2].

Although there are several suggestions and techniques in literature to solve the problem of inconsistency and data duplication, enough tools that contemplate efficiency and effectiveness cannot yet be found, and their cost of processing is quite high and results do not always guarantee that information is properly treated and corrected.

This paper presents a data cleaning environment that addresses needs that are extremely important but previously poorly explored, to provide a tool which is effective and independent as possible of human interaction to implement a large data sources cleaning process, providing an extensible tool which considers possible extensions of new algorithms, techniques, rules and settings, besides being portable to different relational database management systems.

II. BACKGROUND

A data cleaning process should detect and remove errors and inconsistencies from one or more information sources to improve data quality. This process is needed in order to analyze and extract useful knowledge from a repository, a file, a data set of databases, as well as is essential for consistent integration of heterogeneous data sources and one of the data pre-processing steps for data mining. Problems and inconsistencies that are found mainly arise due to spelling errors, invalid data or lack of data during information entry in a file or data repository [3].

Systems of heterogeneous data sources such as data warehouses require a comprehensive treatment of information, especially to support data cleaning. In this type of architecture, data is loaded, huge amounts of information are continuously updated from several repositories and the probability of some sources containing "dirty data" is high. As example of this complexity data analysis, we can indicate the research of a herbarium in [4]. Due to the wide variety of possible inconsistencies and the large volume of data, the data cleaning process is considered one of the most complex processes in order to obtain knowledge. Even nowadays, computer systems provide only limited support for data cleaning because they focus on data transformation for translation and integration schemes [5].

A data cleaning process must satisfy several requirements. First of all, it should detect and remove any mistakes and inconsistencies in the information stored in both single and integrated multiple sources. The approach should be supported by tools that minimize the manual effort and are generic and easily extensible to cover new sources of information. Furthermore, data cleaning should not be done in isolation, but together with related data schema transformations based on comprehensive metadata. Mapping functions for cleaning and other data transformations are specified declaratively and can be reusable for other data sources, as well as for query processing. An infrastructure workflow must reliably and efficiently support all stages of processing data from multiple sources and large data sets. According to [3], data cleaning process may involve the following phases: data analysis, conflict treatment, conflict treatment, transformation, verification, transformation and duplicate elimination tasks.

II.1 SIMILARITY AND DUPLICATE DETECTION ALGORITHMS AND TECHNIQUES

Determining matching cases is typically a high cost operation for large databases. Algorithms for character-based similarity detection aim to treat mainly misspellings or data entry. The main algorithms cited in literature include the Smith Waterman Distance, Affine Gap Distance [5], Edit Distance [6-7], Jaro Distance Metric [8], Q-Gram Distance [9-11], Hamming Distance [12-13] and Jaccard Coefficient [14-15].

Algorithms for token-based similarity detection aim to treat mainly misspellings or typing data as character-based, but with the difference of treating compound words or set of words. The main algorithms that address these cases are: Cosine Similarity [16-18], Atomic Strings [19], WHIRL [20] and Q-Grams with tf.idf [21].

Two techniques previously discussed deal similarities representative of characters composing one or a set of words. However, depending on each language, a set of different characters can be phonetically similar, even if they are physically similar, i.e. similarity of characters or tokens. The main algorithms with this approach are: Soundex [22] Metaphone and Double Metaphone [23], NYSIIS - New Your State Identification and Intelligence System [24], ONCA - Oxford Name Compression Algorithm [25] and PSI - Phonetic Similarity Identification [2].

II.2 TOOLS

Over the last years, some data cleaning tools have been presented and research groups have contributed presenting techniques and free software packages which can be used for duplicate records detection. The main tools presented in literature include WHIRL [26], Flamingo Project5 [27], BigMatch [28], Database Cleaner [29], WizSame [5], Febrl Syste3 - Freely Extensible Biomedical Record Linkage [30-31], TAILOR [32] and SmartClean [33].

II.3 REVIEW OF DATA CLEANING RESEARCH

Published researches in recent years in the data cleaning area have covered some important perspectives such as new algorithms, frameworks and tools that include a combination of several techniques, involving detection of duplicate and statistics in order to try to improve information quality with low cost and minimal user interaction in data cleaning process. Some samples work such as [34-41], propose frameworks in order to offer extensibility and efficiency to the data cleaning process.

Some other researches propose new techniques and algorithms for duplicates detection. Ciszak [42] proposes an algorithm based on data correlation methodology (data mining) to identify and correct physical or semantic duplicate information. Qian [43] introduces the need for VGI data cleaning. Wang [44] presents an algorithm for data cleaning using outlier data detection technique. Okita [45] presents an algorithm for automated data translators cleaning. Bertossi et al. [46] presents an example of data cleaning process using the concept of corresponding dependencies as a procedure for detecting duplicates. Chaturvedi [47] proposes a method that selects a set of data records which, when used to create the data cleaning-based rule model, may include the maximum number of records. Prasad [48] present a tool to improve data quality that identifies variations and synonyms of a given entity present in the data.

II.4 CONTRIBUTIONS OF THIS WORK

The most significant contributions of this work refer to the developed tool that, without user interaction, is automatically capable of analyzing and eliminating 90% of the inconsistencies

and duplications of information contained in a database, with no occurrence of false-positives. Moreover, the data cleaning environment is extensible and portable, allowing it to be easily upgraded and enhanced with new algorithms, techniques, dictionaries, language, etc.

Features, techniques and approaches for the data cleaning process with various tools and frameworks available in literature and the market are summarized in Table 1, where the character "x" indicates the present of the corresponding feature. It is worth noting that the developed data cleaning environment addresses many aspects still little explored by these works in the state of art, thus confirming its contribution.

Table 1: Comparison of features contemplated by diverse data cleaning tools.

	This work	[26]	[27]	[31]	[2]	[4]	[48]	[47]	[46]	[29]	[39]	[40]	[44]
Duplicate Detection	x	x	x	x	x	x	x	x	x	x	x	x	x
Detection and Transformation	x	x	x			x		x		x	x	x	
3 or more Ad-hoc Algorithms	x			x		x		x	x	x	x	x	
Special Character Treatment	x	x	x	x	x	x	x	x		x	x	x	
Phonetics	x			x	x					x			
Multi-Language Phonetics	x				x								
Stopwords	x												
Multi-Language Stopwords	x												
Semantic	x						x			x			
Multi-Language Semantic	x												
Intelligent Approach and Training	x							x	x				
Manual Data Cleaning	x	x	x			x		x		x	x	x	
Semi-Automated Data Cleaning	x												
Automated Data Cleaning	x												
Extensibility	x					x				x	x	x	
Portability	x					x				x	x		

Source: Authors, (2020).

III. COMPARISON OF FEATURES CONTEMPLATED BY DIVERSE DATA CLEANING TOOLS

The data cleaning process can be basically divided into two main activities: analysis and transformation. For both cases, it needs a large computational effort to scan whole database, find duplicate information and correct them.

In the analysis stage, physical and semantic approaches are needed, since in many cases, words with similar duplication spellings do not match, or otherwise, different words physically may represent the same real-world entity, which would represent duplicity or dirt information in the database.

There are many techniques presented in literature for detecting duplicate information based on its physical resemblance (spelling) and phonetics, but it has still been little explored from a semantic perspective. To cover semantic term variations is difficult, since the semantics within a language vary not only in its formal construction, with synonyms, but vary in the same language meanings in different regions, culture and even in the context in which they are used.

Another important point to be explored corresponds to the frequency that the database needs to be cleaned. Most databases have real-world data constantly inserted, resulting in daily duplications and dirt. This means that a cleaning process has to be performed again on the following day, often a few hours later. It shows that a significant portion of the problems are recurrent, i.e., instances of dirt that had previously been detected and treated.

Available algorithms and tools offer partial solutions, attacking some of the problems, with advantages and disadvantages, but there is still work to be done in order to solve problems in analysis, processing and providing a truly effective and efficient data cleaning solution.

The developed data cleaning environment offers efficiency in the whole data cleaning process. The proposed architecture includes modules for specific treatment of analysis and transformation stages, which will be described in detail in the following sections of this paper.

III.1 DATA CLEANING ENVIRONMENT OVERVIEW

The work was developed using C++ development framework supported by Qt with modules for SQL data manipulation language. One of the advantages of the proposed tool is related to its portability and expandability, since it was modularly constructed and can be portable to different DBMS.

Some of its features are directly dependent on the used DBMS, but possibilities are offered to adapt to new database mechanisms that support foreign keys and referential integrity. It is possible to use a simpler database manager system, but some of environment features can still be used. Repositories as auxiliary files and embedded database [49] for training database, stop words database and synonyms database are also used.

Besides all the offered techniques organized into modules, the environment offers several useful features aimed at the infrastructure of the cleaning process such as connecting database through the GUI – Graphical User Interface framework, database cloning and database backup restore. An application of the GUI Software is presented by [50].

III.2 DATA CLEANING ENVIRONMENT ARCHICTETURE

The proposed environment is organized into modules that support the processes of analysis and data transformation. Global auxiliary modules can be used in each step together with the specific modules, supporting both analysis and processing. Its architecture, illustrated in Figure 1, provides structured and independent modules, allowing the user the possibility to work

flexibly and perform tasks ranging from standardization, detection and processing data manually using all the features offered with the fully automated process.

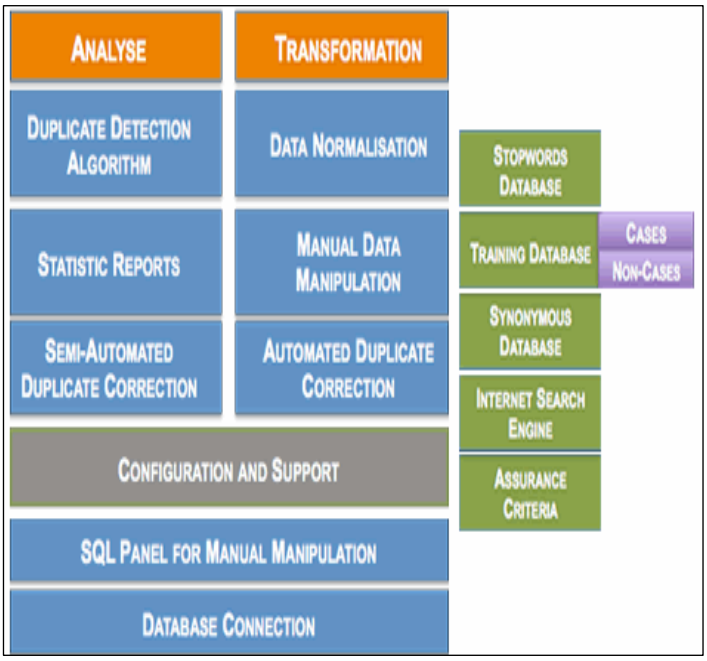

Figure 1: Architecture of developed data cleaning environment.
Source: Authors, (2020).

To support the analysis process, there are modules that include duplicate detection algorithms, statistical reports and semi-automated correction of duplicates. For the transformation process, specific modules are contemplated: data normalization, data manipulation and automation of duplicates correction.

The proposed architecture also includes several global modules that are used in support of both the analysis and transformation processes: SQL panel for manual data manipulation; DBMS connection settings and support, which involves the following sub-modules: Stop words database; Training database (divided into duplicate cases and non-duplicate cases); Synonyms database; Internet Search Engines; Assurance Criteria Rule Settings.

III.3 DATA CLEANING ENVIRONMENT ARCHICTETURE: DATA ANALYSIS

There are three modules parts of the data analysis process whose goals include providing analysis and detection of problems, duplicates and data inconsistencies.

The first module provides several algorithms for detecting duplicate and inconsistent data. In the second module there are statistical reports available generated by automated analysis performed by the tool, allowing the user to have knowledge of how dirty or inconsistent the database is in accordance to defined and configured criteria. The third module, named "semi-automated duplicates correction", contains techniques and algorithms that suggest to the user which of the tuples identified as duplicity is correct, based on database historical references criteria, featuring the environment as an intelligent and trainable tool and thus assisting the user in making the decision to perform the correction and data cleaning.

III.4 ALGORITHMS FOR DETECTING DUPLICATE AND INCONSISTENT DATA

There are several variations and algorithms to detect duplicates in databases, each aiming to attack a specific problem. In order to make the developed tool as applicable as possible, main algorithms presented in literature that involve techniques of similarity of characters, and phonetic tokens similarity were implemented.

The environment also provides the user with the option to create new similarity detection algorithms like attributes through the system interface, which support JavaScript language, on a panel that validates syntax of the language, illustrated in Figure 2, and allows total flexibility for the user to apply specific techniques beyond those already offered, according to user's needs. This is one of the ways that the data cleaning environment includes expandability, but it is also possible that other algorithms be included in C++ language, as the natives of the environment. When the analysis and detection of duplicity and inconsistency process is started, one of the algorithms implemented in the environment must be chosen.

III.5 SYNONYMS DATABASE

Synonyms functionality was developed in order to cover the lack of semantics in the process of detecting inconsistencies and duplicate information in the database. With a synonyms database configured by the user, a scan is performed for similarity detection and it is done not only through the techniques of the chosen algorithm, but also includes checks of semantically equivalent terms, such as "helper", "help" and "collaborator".

Depending on the context, words can have the same meaning, but would not be easily detected unless there was a dictionary of equivalent terms. Another example would be a database of accidents that diagnoses animal bites. Generally it reports that the victim was bitten or stung by a specific animal, like a dog, snake, etc., but statistically it is only important to obtain accidents that were caused by animals. In this case, a list of animals can be included in the synonyms database and when any of them is detected, the system automatically converts them to "animal", and the cleaning will be performed in an automated process.

III.6 MULTI-LANGUAGE DATABASE

A multi-language database allows the user to load translation dictionaries for multiple languages in order to transform the environment into a tool that can detect duplicates of information regardless of the language in which it is. Once it is properly configured, similar data attributes or semantically identical attributes will be detected even if they are described in different languages.

For example, a table containing data of animals would be detected as tuples with similar attributes, dog, "cachorro", "perro", "cão", etc. With this functionality, the language barrier between data repositories is broken, making it feasible to prepare an integrated base of different languages for knowledge extraction, after applying the cleaning process in which all information would be cleaned (standarded) to a language key.

III.7 MULTI-LANGUAGE PHONETIC DETECTION

For each of the techniques and algorithms contemplated by the environment, an option that considers the phonetics of words in the data analysis process is also offered. Thus, each technique considers the phonetic differences among the possibilities of writing the same word to scan each tuple of the table in similarity searches and overlaps, caused either by an incorrect entry or misspelling which are highly recurrent on systems with data entry by users, e.g. "S" instead of "SS", "CH" instead of "X", etc. The implemented algorithm includes the following Portuguese language phonetic conversions, represented in Table 2.

For data cleaning environment scalability, the phonetic transcriptions are stored into the embedded database tool and can be improved with new rules, allowing phonetic rules insertion of any languages besides Portuguese and English.

Table 2: Portuguese language phonetic transcriptions.

Letter/Syllable	Phoneme	Condition
X	Z	If not in the beginning or at the end and is surrounded by vowels, have sound Z
Ç	S	
CH	X	
CE	SE	
CI	SI	
CA	KA	
CO	KO	
CU	KU	
GE	JE	
GI	JI	
PH	F	
QU	K	
RR	R	
SS	S	
S	Z	If not in the beginning or at the end and is surrounded by vowels, have sound Z
Z	S	If is the last position, have sound S
W	V	

Source: Authors, (2020).

III.8 SEMI-AUTOMATED DUPLICATES CORRECTION

One of the biggest difficulties for the user in a data cleaning process is to detect two or more similar attributes that actually correspond to the same entity. Often, similar words do not match semantically and the conversion of one to another may generate more database inconsistency. This leads to some limiting factors for cleaning process implementation, for example, the need for a data specialist to make decisions and great efforts in manual searches to check which detected information is the correct one and if indeed these are the same entity. It is important to emphasize that this situation is constant and many data cleanings must be repeatedly performed since the input is inconsistent in most cases and impossible to be completely avoided.

With the constant realization of data cleaning processes, the cost based on historical cleaning can be minimized. An automated analysis was developed based on the quantity of times each tuple is referenced by another if two or more tuples are detected as similar records according to the chosen similarity detection algorithm. The system checks the quantity of times each one is referenced by other tables and, therefore, indicates to the user the tuple most likely to be correct. It is worth thinking that if a tuple is more referenced than another one, the chances of it containing the correct information is greater, since after several cleanings have been done and where other tuples have also been converted to it, their referenced quantity is higher.

The process called semi-automated duplicates correction can be described with the following pseudo-algorithm in Figure 2.

```
Semi-automated duplicates correction

1. Check similarity based on chosen algorithm;
2. For each group of detected similar tuples:
        a. Verify if the tuples had been detected earlier
              i. If yes, check the history of what was
                 done,i.e., whether cleaning was done, keep
                 the correct tuple and go to step 4;
        b. Check each tuple for how referenced it was by other
           tables in the database;
        c. Keep the total amount;
3. Check between tuples of the group for the one that has the
   biggest amount;
4. Show the user which is the most referenced or historically
correct tuple.
```

Figure 2: Semi-automated duplicates correction algorithm.
Source: Authors, (2020).

This functionality is useful for the manual cleaning process, mainly because it assists the user in making decisions based on historical cleanings performed previously by the amount of tuples and referenced in the current landscape of the database. Even in cases where two attributes are substantially similar, the user has a tool to aid in deciding whether, for example, a company name is written in one way or another, according to the quantity in which they are referenced as it is expected that, in a database constantly cleaned, the correct name is referenced more than a new inconsistent data that has been recently inserted, or a duplicity which has been treated and corrected earlier.

IV. DATA CLEANING ENVIRONMENT ARCHITECTURE: DATA TRANSFORMATION

The data transformation step basically consists in applying the necessary corrections and changes to the database. The data cleaning environment provides a set of functionalities that offer many possibilities of user interaction to make changes and corrections.

IV.1 AUTOMATION OF DUPLICATES CORRECTION

The main objective for developing this work consisted in the need of avoiding as much user interaction in the data cleaning process as possible, since large relational databases contain millions of records organized in tables and columns, and a human analysis for duplicity and inconsistency detection is impossible without the appropriate computational support.

Using all the analysis techniques previously described, it is possible that some, if not all the database is cleaned and handled automatically. To do that, a module called Automation of Duplicates Correction was developed that consists of an automatic application analysis and database table transformation, without the need for user interaction during the data cleaning process.

The user needs only to indicate which table should be cleaned, choose the type attribute, the similarity detection algorithm that will be used and check the option Automatic Cleaning. Once the mechanism is triggered, the environment will do the work of scanning, analyzing and correcting duplicate and inconsistent data.

In order to avoid the automated correction processes converting tuples that are not actually duplicates, despite the high physical similarity or semantic, a set of rules and conditions that guarantee a right automatic correction was developed, called Assurance Criteria. Their terms and conditions are described by pseudo-following algorithm in Figure 3.

```
Assurance Criteria

1. Begin the duplicate detection process
2. If two or more possible duplicates detected:
   a. Checks if the attributes are identical
      1. If YES:
         i.  Verify which tuple is more referenced in database
         ii. Perform clean-up and returns to STEP 2
      2. If NO:
         i. Checks whether this case has been previously cleaned in cleaning
            historical database
            If YES:
               a. Perform clean-up according to the history
               b. Return to STEP 2
            If NO:
               a. Select tuple which is the most referenced in the database
                  a. Perform clean-up
                  b. Return to STEP 2
               b. In case of a tie, i.e., the tuples have the same amount that
                  are referenced by other tables in the database
                  a. There is a tuple that has more attributes not null or not
                     empty
                     i. Perform clean-up
                     ii.Return to STEP 2
                  b. If there is a tie, the cleaning is not performed
3. END
```

Figure 3: Assurance criteria algorithm for automated duplicates correction.
Source: Authors, (2020).

The conditions contemplated by Assurance Criteria contribute towards corrections not performing automatic errors. Obviously, the process is highly dependent upon the chosen algorithm and that settings for the tuple identified as potential duplicates are indeed equivalent. The environment also provides an additional check before all the fixes are actually performed.

IV.2 DATA CLEANING ENVIRONMENT TRAINING

In every completed data cleaning process, be it manual, semi-automated or automated, the data cleaning environment takes care of storing cases in a training database where clean-ups were done. In the manual and semi-automated options, all converted tuples are stored as historical and a positive-case, even when the event is recurring, the system will automatically suggest the correct tuple. For the automatic cleaning option, all tuples will be automatically converted in accordance with the historical database.

Likewise, the training is performed for negative cases that are stored in the negative historical database in which the algorithm, according to the criteria and parameters set by the user, indicates possible replication but which, although similar, do not represent duplicates. In these cases, it is possible to indicate that this is a false-positive and, once appointed, the recurrent cases are automatically ignored by the tool.

In the case of false-positives, interaction is required from the user to train the tool, indicating which cases should be treated as non-duplicates. With the support of automated clean-up, this analysis can be quick, since the environment generates a report of all possible cases found so that the user can mark them as false-positives. In all labeled cases, history will be recorded and in future clean-up recurrences will be automatically discarded by the environment.

The data cleaning environment training process of false-positives is illustrated in Figure 4. The selected lamp icon indicates that the event identified by the tool is correct, i.e. it is a duplicate. In cases of non-duplicates, the user must click on the lamp, which will be unmarked, and the case will be automatically stored in the database history as a non-duplicate. As the tool gets trained, the cleaning of a specific database becomes increasingly reliable and rapid, besides making the automated process very convenient because the cases are handled automatically to be more reliable.

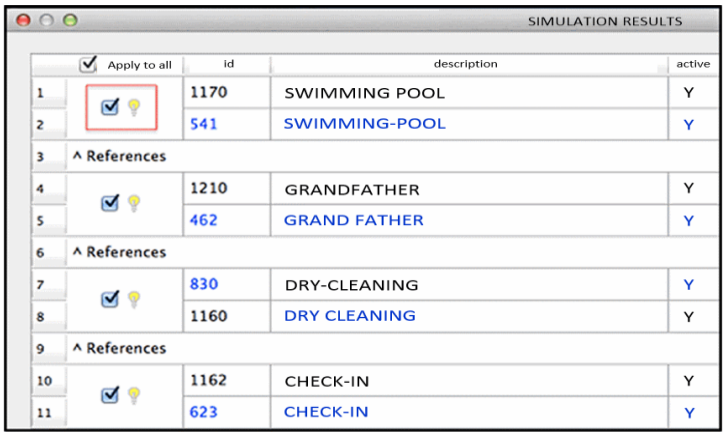

Figure 4: Training for cases of false-positive.
Source: Authors, (2020).

V. EXPERIMENTS AND RESULTS

The SIVAT – Sistema de Informação e Vigilância de Acidentes de Trabalho (Labour Accidents Vigilance System) system database to manage information about labour accidents in São José do Rio Preto city, state of Sao Paulo, Brazil and the region, totaling more than one hundred cities was used in all performed experiments in this work. It currently has about 90 thousand registered notifications of labour accidents collected by CEREST – Centro de Referência em Saúde do Trabalhador (Labourer Health Reference Centre), responsible for managing and apply actions involving the prevention and investigation of labour accidents.

V.1 COMPARATIVE EXPERIMENTS

This section presents and discusses the results obtained from experiments conducted in the data cleaning process, in order to prove the efficacy of cleaning using the data cleaning environment developed in this work. The table that was used was the "Machine Involved in the Labour Accident" which had a total of 639 records. A manual analysis was performed on all table data and it was found that there were 473 non-duplicate records, or approximately 27% of the table showed inconsistencies or duplicate data.

It is important to highlight that, unlike the results reported by most studies in literature, whose main purpose is to highlight the effectiveness of techniques to detect duplicates of valid cases, this work also presented the percentage of false-positives, i.e., the rate of incorrect duplicates that were detected. Data cleaning process is costly and if the amount of false-positives is high, the difficulty for the user to analyze and make the necessary corrections increases significantly.

It is also noteworthy that the proposal to automate the cleaning process must provide for the treatment of false-positives,

since incorrect changes could generate more inconsistencies in the database instead of providing an adequate treatment of corrections and cleaning.

Experiments were carried out by applying manual / semi-automated and automated data cleaning, using the duplicate detection algorithm Q-Gram, with parameter q=3, with the percentages of similarity 90%, 80%, 75%, 70% and 65%. From some preliminary tests it was observed that above 90% similarity, false-positives are not detected and, below 65%, the results presented by the environment are not very relevant.

The application process for cleaning experiments was organized into 6 stages. Functionality or technique to verify the improvement of inconsistencies detection in the database was added at each step.

In the first step, only duplicate detection Q-Gram was applied to the algorithm to verify the percentage of detections that were obtained. Additions to the subsequent steps were: (2) treatment of special characters, normalization and stemming, (3) treatment of word phonetics, (4) stop words treatment and removal, (5) synonyms and multi-languages databases and to the last step (6) the cleaning was done with the previously trained environment. In each step, the percentage of detected false-positives was also measured.

V.2 MANUAL AND SEMI-AUTOMATED DATA CLEANING APPLICATION

As it can be seen from the results shown in Table 3, with a rate of similarity of 90%, 37 inconsistencies were detected with the application of the Q-gram algorithm, corresponding to 22% of all inconsistencies and duplicates present in the database. As refinement techniques are added, the number of presented inconsistencies significantly increases. Interestingly, with the use of synonyms in the cleaning process, the efficiency increased 71% and 75% of these inconsistencies were detected. With the trained environment, the percentages rose to 72% and 78%.

Table 3: Results obtained with data cleaning process using the algorithm Q-Grams, with q=3 and 90% similarity.

Q-Gram, q=3, 90%	Inconsistencies	False Positive	+Effectiveness	Effectiveness
Ad-hoc Algorithm	37	0	0%	22%
Special Characters	37	0	0%	22%
Phonetic	42	0	12%	25%
Stop words	58	0	36%	35%
Synonymous	127	2	71%	75%
Training	130	0	72%	78%

Source: Authors, (2020).

With a similarity factor of 80%, as shown in Table 4, 53 inconsistencies were detected with the application of Q-gram algorithm, but 6 of these were false-positives. The difference, namely 47 cases corresponded to 28% of all inconsistencies and duplicates present in the database. On adding refinement techniques, the number of displayed inconsistencies significantly increased, representing a total of 139 cases detected and the number of false-positives decreased. With the use of synonyms in the cleaning process, the efficiency increased by 62%, and 80% of all the inconsistencies were detected. With the trained environment, the efficiency reached 83%. It is emphasized that, with the trained tool, no false-positives were presented.

Table 4: Results obtained with data cleaning process using the algorithm Q-Grams, with q=3 and 80% similarity.

Q-Gram, q=3, 80%	Inconsistencies	False Positive	+Effectiveness	Effectiveness
Ad-hoc Algorithm	53	6	0%	28%
Special Characters	54	5	2%	29%
Phonetic	58	4	9%	32%
Stop words	73	3	27%	42%
Synonymous	139	5	62%	80%
Training	139	0	62%	83%

Source: Authors, (2020).

With similarity factor of 75%, as shown in Table 5, 65 inconsistencies were detected with the application of the algorithm Q-Gram, but 8 of them were false-positives. The difference, namely 57 cases correspond to 34% of all inconsistencies and duplicates present in the database. On adding refinement techniques, the number of presented inconsistencies further increased, for a total of 143 cases detected and the number of false-positives decreased. With the use of synonyms in the cleaning process, the efficiency increases by 56%, and 83% of these inconsistencies were detected. With the trained environment, the efficiency reached 86%. Same as with the experiment with similarity factor 80% with the trained tool, no false-positives were presented.

Table 5: Results obtained with data cleaning process using the algorithm Q-Grams, with q=3 and 75% similarity.

Q-Gram, q=3, 75%	Inconsistencies	False Positive	+Effectiveness	Effectiveness
Ad-hoc Algorithm	65	8	0%	34%
Special Characters	67	9	0%	35%
Phonetic	70	9	4%	37%
Stop words	83	8	19%	45%
Synonymous	147	9	54%	83%
Training	143	0	53%	86%

Source: Authors, (2020).

With a percentage similarity of 70% as shown in Table 6, 74 inconsistencies were detected with the application of Q-gram algorithm, but 11 of them were false-positives. The difference, namely 63 cases corresponded to 38% of all inconsistencies and duplicates present in the database. On adding refinement techniques, the number of presented inconsistencies rose, representing a total of 143 cases detected and the number of false-positives also decreased. With the use of synonyms in the cleaning process, the efficiency increased by 50% and 83% of all the inconsistencies was detected. With the trained environment, the efficiency reached 86%. Same as with the experiments with similarity factor 90%, 80% and 75%, with the trained tool, no false-positives were presented.

Table 6: Results obtained with data cleaning process using the algorithm Q-Grams, with q=3 and 70% similarity.

Q-Gram, q=3, 70%	Inconsistencies	False Positive	+Effectiveness	Effectiveness
Ad-hoc Algorithm	74	11	0%	38%
Special Characters	75	11	0%	38%
Phonetic	78	10	4%	41%
Stop words	85	11	12%	44%
Synonymous	148	10	49%	83%
Training	143	0	48%	86%

Source: Authors, (2020).

Finally, with a percentage similarity of 65% as seen in Table 7, 94 inconsistencies were detected in the application of the algorithm Q-Gram, but 14 of them were false-positives. The difference, namely 80 cases corresponded to 48% of all inconsistencies and duplicates present in the database. On adding refinement techniques, the number of presented inconsistencies rose, representing a total of 153 cases detected, but the amount of false-positives nevertheless increased until the environment was trained. With the use of synonyms in the cleaning process, the efficiency increased by 42%, and 84% of these inconsistencies were detected. With the trained environment, the efficiency reached 92%.

Table 7: Results obtained with data cleaning process using the algorithm Q-Grams, with q=3 and 65% similarity.

Q-Gram, q=3, 65%	Inconsistencies	False Positive	+Effectiveness	Effectiveness
Ad-hoc Algorithm	94	14	0%	48%
Special Characters	97	21	0%	46%
Phonetic	96	17	-1%	47%
Stop words	103	19	6%	50%
Synonymous	161	20	40%	84%
Training	153	0	37%	92%

Source: Authors, (2020).

What stands out in these first experiments is the high degree of effectiveness that developed environment presented in the manual/semi-automated cleaning process and the treatment of false-positives that are not presented to the user for analysis.

V.3 AUTOMATED DATA CLEANING APPLICATION

The next set of experiments approach the automated data cleaning process, that is, where was no user interaction in both the data analysis and processing stages.

As it can be seen from the results shown in Table 8, similarly rate at 90%, 34 inconsistencies were detected with the application of the algorithm Q-gram, corresponding to 20% of all inconsistencies and duplicates present in the database. On adding refinement techniques, the number of presented inconsistencies significantly increased. It is interesting to note that, same as with the manual process with the use of synonyms in the automated cleaning process, efficiency increased 58% and 54% of these inconsistencies were detected. With the trained environment, the percentages rose to 69% and 74%. With a similarity of 90%, false-positives are virtually undetected, but not all inconsistencies present in the database are found by data cleaning environment.

Table 8: Results obtained with data cleaning process using the algorithm Q-Grams, with q=3 and 90% similarity.

Q-Gram, q=3, 90%	Inconsistencies	False Positive	+Effectiveness	Effectiveness
Ad-hoc Algorithm	34	0	0%	20%
Special Characters	39	0	13%	23%
Phonetic	39	0	13%	23%
Stop words	55	0	38%	33%
Synonymous	92	1	58%	54%
Training	124	0	69%	74%

Source: Authors, (2020).

With similarity factor of 80%, as shown in Table 9 were detected 51 inconsistencies with the application of the algorithm Q-Gram, but 5 of them were false-positives. The difference, namely 46 cases corresponded to 28% of all inconsistencies and duplicates present in the database. On adding refinement techniques, the number of displayed inconsistencies significantly increased, representing a total of 133 cases detected and the number of false-positives decreased. With the use of synonyms in the cleaning process, the efficiency increased 47% and 60% of these inconsistencies were detected. With the trained environment, efficiency reached 80%. It is emphasized that, with the trained tool, no false-positives were presented.

Table 9: Results obtained with data cleaning process using the algorithm Q-Grams, with q=3 and 80% similarity.

Q-Gram, q=3, 80%	Inconsistencies	False Positive	+Effectiveness	Effectiveness
Ad-hoc Algorithm	51	5	0%	28%
Special Characters	55	4	7%	31%
Phonetic	55	4	7%	31%
Stop words	70	3	27%	40%
Synonymous	104	3	47%	60%
Training	133	0	59%	80%

Source: Authors, (2020).

With a similarity factor of 75%, as shown in Table 10, 61 inconsistencies were detected with the application of the algorithm Q-Gram, but 9 of them were false-positives. The difference, namely 52 cases corresponds to 31% of all inconsistencies and duplicates present in the database. On adding refinement techniques, the number of inconsistencies presented further increases, for a total of 136 cases detected and the number of false-positives decreased. With the use of synonyms in the cleaning process, the efficiency increased by 42%, and 62% of these inconsistencies were detected. With the trained environment, the efficiency reached 81%. Same as the experiment with similarity factor 80% with the trained tool, no false-positives were presented.

Table 10: Results obtained with data cleaning process using the algorithm Q-Grams, with q=3 and 75% similarity.

Q-Gram, q=3, 75%	Inconsistencies	False Positive	+Effectiveness	Effectiveness
Ad-hoc Algorithm	61	9	0%	31%
Special Characters	64	9	5%	33%
Phonetic	64	9	5%	33%
Stop words	77	8	21%	41%
Synonymous	111	7	42%	62%
Training	136	0	53%	81%

Source: Authors, (2020).

With a percentage similarity of 70%, as shown in Table 11, 66 inconsistencies were detected with the application of Q-gram algorithm, but these 9 false-positives. The difference, namely 57 cases correspond to 34% of all inconsistencies and duplicates present in the database. On adding refinement techniques, the amount of inconsistencies presented rose, representing a total of 136 detected cases and the number of false-positives also decreased. With the use of synonyms in the cleaning process, the efficiency increased by 39%, and 60% of these inconsistencies were detected. With the trained environment, the efficiency reached 81%. Same as the experiments with similarity factor 90%, 80% and 75% with the trained tool, no false-positives were presented.

Table 11: Results obtained with data cleaning process using the algorithm Q-Grams, with q=3 and 70% similarity.

Q-Gram, q=3, 70%	Inconsistencies	False Positive	+Effectiveness	Effectiveness
Ad-hoc Algorithm	66	9	0%	34%
Special Characters	68	10	3%	35%
Phonetic	68	10	3%	35%
Stop words	77	9	14%	41%
Synonymous	111	10	39%	60%
Training	136	0	50%	81%

Source: Authors, (2020).

Finally, with percentage of similarity of 65% seen in Table 12, 79 inconsistencies were detected in the application of the algorithm Q-Gram, but 16 of them were false-positives. The difference, namely 63 cases correspond to 38% of all inconsistencies and duplicates present in the database. On adding refinement techniques, the number of inconsistencies presented rose, representing a total of 147 detected cases, but the amount of false-positives nevertheless increased until the tool was trained. With the use of synonyms in the cleaning process, the efficiency increased 33% and 62% of these inconsistencies were detected. With the trained environment, the efficiency reached 88%.

Table 12: Results obtained with data cleaning process using the algorithm Q-Grams, with q=3 and 65% similarity.

Q-Gram, q=3, 65%	Inconsistencies	False Positive	+Effectiveness	Effectiveness
Ad-hoc Algorithm	79	16	0%	38%
Special Characters	80	14	1%	40%
Phonetic	80	14	1%	40%
Stop words	87	14	9%	44%
Synonymous	119	15	33%	62%
Training	147	0	46%	88%

Source: Authors, (2020).

Even with the cleaning process running automatically without user interaction in the data analysis, we highlight the high degree of effectiveness that the environment has developed, covering approximately 90% of total inconsistencies in the database, and an appropriate treatment of false-positives, as no cleaning was done incorrectly when using the trained environment.

V.4 ANALYSIS AND COMPARISON OF OBTAINED RESULTS

The results obtained by applying data cleaning process using manual and automated approaches are displayed in the next comparative graphs. Due to the rules established by Assurance Criteria used in automated approaches, some cases were not cleaned automatically, and therefore the efficiency of coverage of treated inconsistent cases is somewhat lower.

In Figure 5 is shown the percentage of effectiveness since the application of the algorithm, on the left, until the process with all available resources was executed, such as processing of special characters, phonetics, stop words, multi-language synonyms and training. The automated cleaning efficiency was only 5% lower than the automated when the applied algorithm similarity was 90%.

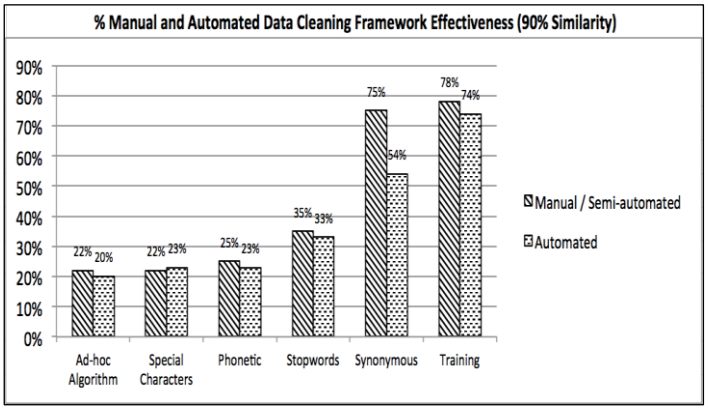

Figure 5: Manual and automated data cleaning environment effectiveness with 90% similarity.
Source: Authors, (2020).

As illustrated in Figure 6, with an 80% similarity algorithm, the automated cleaning with the trained environment had an efficacy of only 3.5% lower than manual.

As can be observed in Figures 7 and 8, with 70% and 65% similarity algorithms, the cleaning automated with the trained environment had an efficiency of approximately 6% less than manual.

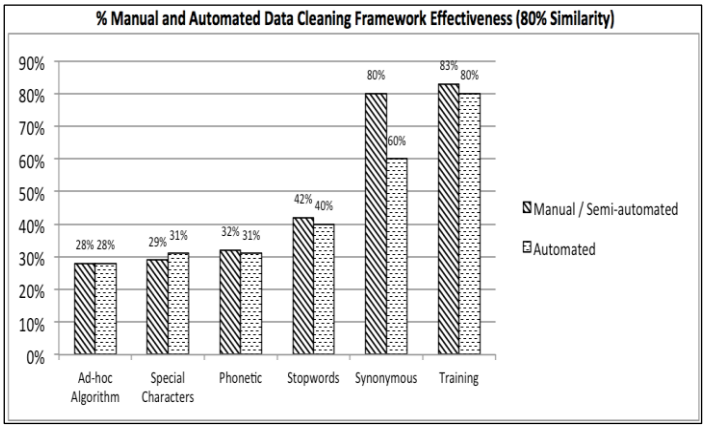

Figure 6: Manual and automated data cleaning environment effectiveness with 80% similarity.
Source: Authors, (2020).

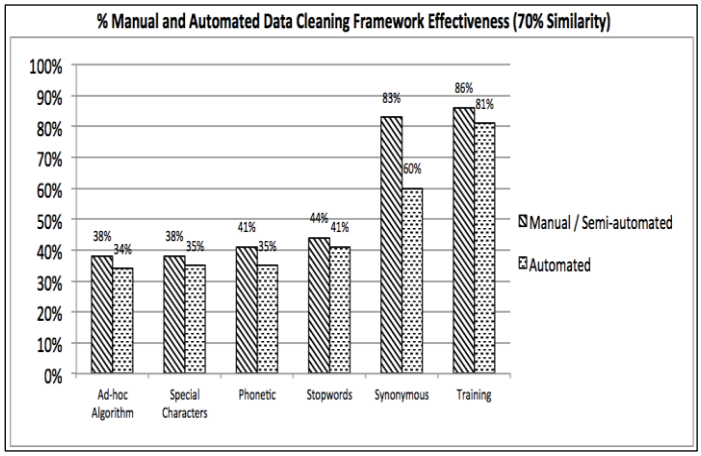

Figure 7: Manual and automated data cleaning environment effectiveness with 70% similarity.
Source: Authors, (2020).

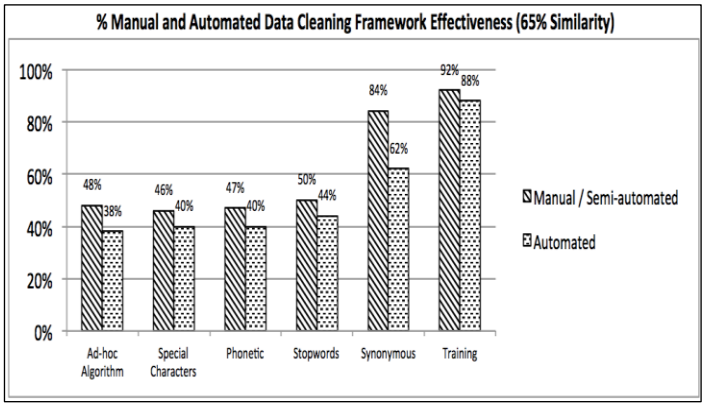

Figure 8: Manual and automated data cleaning environment effectiveness with 65% similarity.
Source: Authors, (2020).

The automated cleaning process appeared to be more efficient when compared to the manual process in analysis, detection and treatment of inconsistency and duplicate problems in databases, but it is important to note that the process can only be automated when there is an environment that supports the treatment of false-positive cases. The percentages of false-positive results shown in Figures 9 and 10. The charts show that only with an intelligent approach is it possible to treat such cases and thus avoid inconsistencies that are presented to the user in case of a manual approach or improperly transformed in the case of automated approach.

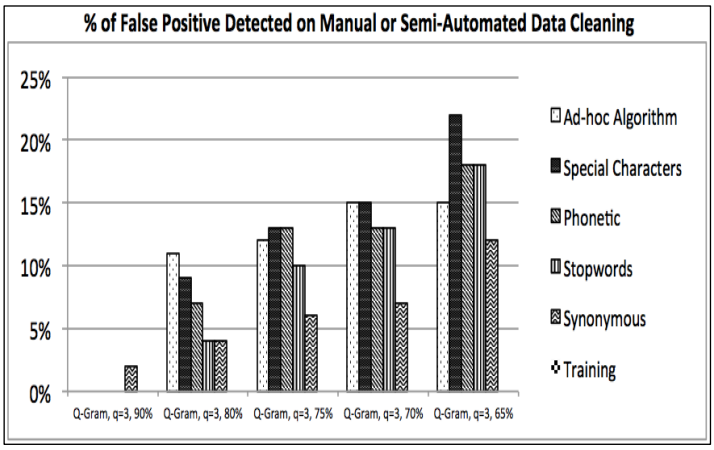

Figure 9: Percentage of false-positives detected in manual data cleaning process.
Source: Authors, (2020).

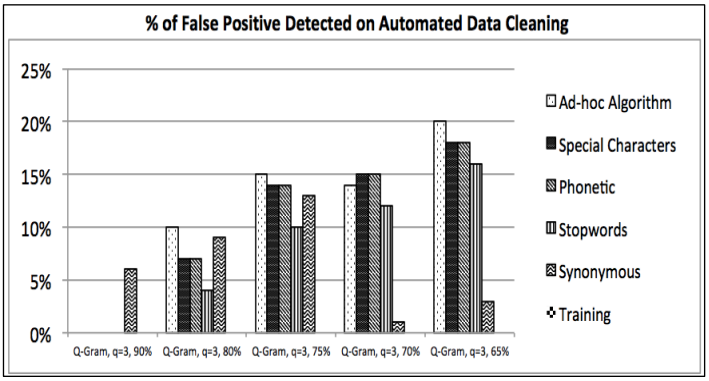

Figure 10: Percentage of false-positives detected in automated data cleaning process.
Source: Authors, (2020).

Depending on the similarity factor used in the application of the algorithms to detect duplicates, the percentage of false-positives can exceed 20%. However, in all experiments in which all these techniques in the developed environment modules were applied simultaneously, including the intelligent approach in which the tool was trained, a false-positive rate of 0% was delivered, ensuring user safety for transformations of inaccurate information present in databases where the environment was applied.

With a proper adequacy, with the proposed algorithm and developed modules being used together, the efficiency of the tool was significant and covered approximately 90% of all inconsistencies in the database, with a 0% percentage of false-positive cases. A survey of the works in literature shows that they only have a 50% efficiency and only part of the functionality and treatments of the developed environment.

Also noteworthy is that the objective of the experiments was to demonstrate technologies and approaches not only to detect and treat positive cases of information inconsistencies and duplication, but also address cases of false-positives detection and analyses the negative impacts that they represent to the data cleaning process, whether manual or automated. It is important to emphasize that this analysis perspective is still largely discussed in literature and can be regarded as being essential in order to obtain increasingly effective results.

VI. CONCLUSIONS

Data cleaning process is shown as an essential and most important step to obtain knowledge. There are few studies that compare the effectiveness of different techniques of data cleaning and the few studies that focus on these analyses are insufficient, since most of the techniques and solutions proposed so far require a great interaction with the user to analyse and decide detections made and, with regard to large databases, it is humanly impossible to direct interaction in the process of cleaning that easily contains millions of records. This interaction is mainly due to the power of semantic decision, because the techniques and work done so far focused specifically on detection of physical similarity between texts and bodies.

This work has, among its objectives, to reduce user interaction in the process of analyzing and correcting inconsistencies and duplications. A data cleaning environment was built for configurable automatic data cleaning based on training and physical-semantic data similarity that provides a more effective and extensible solution, addressing little or unexplored gaps in literature such as training, semantic support, multi-language support and partial or total automation of the cleaning process.

The results of the tests proved the effectiveness of all the developed features, relevant to each module of the proposed architecture. In several scenarios the experiments demonstrated the effectiveness of the tool. With a proper configuration, and with the proposed algorithm and developed modules used together, the efficiency is significant and covers approximately 90% of database inconsistencies, with a 0% percentage of false-positive cases. The work area in literature has presented only part of the functionality and treatments when compared to the developed environment, and their effectiveness does not exceed 50%.

Approaches have also demonstrated that in addition to detecting and treating positive cases of information inconsistencies and duplication, they address cases of detected false-positives and consider the negative impacts that they

represent to the data cleaning process, whether manual or automated, not yet discussed in literature.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Carlos Roberto Valêncio, Toni Jardini and Victor Hugo Penhalves Martins.
Methodology: Carlos Roberto Valêncio and Toni Jardini.
Investigation: Carlos Roberto Valêncio and Toni Jardini.
Discussion of results: Carlos Roberto Valêncio, Toni Jardini, Victor Hugo Penhalves Martins and Angelo Cesar Colombini.
Writing – Original Draft: Toni Jardini.
Writing – Review and Editing: Carlos Roberto Valêncio and Angelo Cesar Colombini.
Resources: Toni Jardini.
Supervision: Carlos Roberto Valêncio and Márcio Zamboti Fortes.
Approval of the final text: Carlos Roberto Valêncio and Márcio Zamboti Fortes.

VIII. REFERENCES

[1] Bharat, K. et al.. Special Issue on Data Cleaning, Bulletin of the Technical Committee on Data Engineering, 2000.

[2] Andrade, T.L.; Souza, R.C.G.; Balbini, M.; Valêncio, C.R.. Optimization of Algorithm to Identification of Duplicate Tuples through Similarity Phonetic Based on Multithreading, in Proc. of 2011. 12th International Conference on Parallel and Distributed Computing, Applications and Technologies - PDCAT, 299–304, 2011. DOI: 10.1109/PDCAT.2011.58.

[3] Rahm, E.; Do, H.H.. Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin 23, pp. 3–13, 2000.

[4] Silva, L.A.E..A Data Mining Approach for Standardization of Collectors Names in Herbarium Database. IEEE Latin America Transactions, vol.14, no.2, pp.805-810, 2016. DOI: 10.1109/TLA.2016.7437226.

[5] Elmagarmid, A.K.; Ipeirotis, P.G.; Verykios, V.S.. Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering, vol.19, no.1, p.1–16, 2007. DOI: 10.1109/TKDE.2007.250581.

[6] Ayad, L.A.K.; Barton, C.; Pissis, S.P.; A faster and more accurate heuristic for cyclic edit distance computation. Patter Recognition Letters, vol.88, p.81-87, 2017. DOI: 10.1016/j.patrec.2017.01.018.

[7] Su, Z. et al.. Plagiarism Detection Using the Levenshtein Distance and Smith-Waterman Algorithm, in Proc. of 3rd International Conference on Innovative Computing, Information and Control - ICICIC '08, p.569-573, 2008. DOI: 10.1109/ICICIC.2008.422.

[8] Gueddah, H.; Yousfi, A.; Belkasmi, M.. The filtered combination of the weighted edit distance and the Jaro-Winkler distance to improve spellchecking Arabic texts, in Proc. of 12th International Conference of Computer Systems and Applications – AICCSA, 2015. DOI: 10.1109/AICCSA.2015.7507128.

[9] Hanada, H.; Kudo, M.; Nakamura, A.. Average-case linear-time similar substring searching by the q-gram distance. Theoretical Computer Science. vol. 530, p.23-41, 2014. DOI: 10.1016/j.tcs.2014.02.022.

[10] Wang, G.; Wang, B.; Yang, X.; Yu, G.. Efficiently Indexing Large Space Graphs for Similarity Search. IEEE Transactions on Knowledge and Data Engineering. vol.24, no.3, p.440-451, 2010. DOI: 10.1109/TKDE.2010.28.

[11] Petrovic, S.; Bakke, S.. Improving the Efficiency of Misuse Detection by Means of the q-gram Distance, in Proc. of Fourth International Conference on Information Assurance and Security - ISIAS '08, p.205–208, 2008. DOI: 10.1109/IAS.2008.39.

[12] Khristodulo, O.I.; Makhmutov, A.A.; Sazonova, T.V.. Use Algorithm based at Hamming Neural Network Method for Natural Objects Classification. Procedia Computer Science, vol. 103, p.388-395, 2017. DOI: 10.1016/j.procs.2017.01.126.

[13] Boutalis, Y.S.. A new method for constructing kernel vectors in morphological associative memories of binary patterns. Computer Science and Information Systems, vol.8, no.41, p.141-166, 2011. DOI: 10.2298/CSIS091114026B.

[14] Chen, S.Y. et al.. Concept Extraction and Clustering for Search Result Organization and Virtual Community Construction. Computer Science and Information Systems, vol. 9, no.1, p.323-355, 2012. DOI: 10.2298/CSIS101124020C.

[15] Jiang, Z.; Evans, M.; Oliver, D; Shekkar, S.. Identifying K Primary Conditions from urban bicycle GPS trajectories on a road network. Information Systems, vol.57, p.142-159, 2016. DOI: 10.1016/j.is.2015.10.009.

[16] Hernandez, A.F.R.; Garcia, N.Y.H.; Distributed processing using cosine similarity for mapping Big Data in Hadoop. IEEE Transactions on Latin America, vol.14, no.6, p.2857-2861, 2016. DOI: 10.1109/TLA.2016.7555265.

[17] Jimenez, S.; Gonzalez, F.A.; Gelbukh, A.. Mathematical properties of soft cardinality: Enhancing Jaccard, Dice and cosine similarity measures with element-wise distance. Information Sciences, vol.367-368, p.373-389, 2016. DOI: 10.1016/j.ins.2016.06.012.

[18] Zhu, S.; Wu, J.; Xia, G.. TOP-K Cosine Similarity Interesting Pairs Search, in Proc. of Seventh International Conference on Fuzzy Systems and Knowledge Discovery - FSKD 2010, 1479–1483, 2010. DOI: 10.1109/FSKD.2010.5569212.

[19] Monge, A.E.; Elkan, C.P.. The Field Matching Problem: Algorithms and Applications, in KDD-96 Proceedings, p.267–270, 1996.

[20] Cohen, W.W.. Integration of heterogeneous databases without common domains using queries based on textual similarity, in Proc. 1998 ACM SIGMOD International Conference on Management of data, 201–212, 1998. DOI: 10.1145/276304.276323.

[21] Gravano, L. et al.. Text joins in an RDBMS for web data integration, in Proc. of WWW '03 – 12th International Conference on World Wide Web, 90-101, 2003. DOI: 10.1145/775152.775166.

[22] Holmes, D.; McCabe, M.C.. Improving precision and recall for Soundex retrieval, in Proc. of International Conference on Information Technology: Coding and Computing – ITCC, p.22–26, 2002. DOI: 10.1109/ITCC.2002.1000354.

[23] Mandal, A.K.; Hossain, M.D.; Nadim, M.. Developing an efficient search suggestion generator, ignoring spelling error for high speed data retrieval using Double Metaphone Algorithm, in Proc. of 13th International Conference on Computer and Information Technology – ICCIT, p.317–320, 2010. DOI: 10.1109/ICCITECHN.2010.5723876.

[24] Taft, R.L.. Project SEARCH - Name Search Techniques – Special Report No. 1, New York State Identification and Intelligence System, 1-118, 1970.

[25] Gill, L.E.. OX-LINK: The Oxford Medical Record Linkage System, in International Workshop and Exposition, p.15–45, 1997.

[26] Cohen, W.W.. Data Integration using Similarity joins and a Word-Based Information Representation Language. ACM Transactions on Information Systems, vol.18, no.3, p.288-321, 2000.

[27] Flamingo. Flamingo Project on Data Cleaning, Department of Computer Science. UC Irvine, 2017. http://flamingo.ics.uci.edu.

[28] Yancey, W.E.. Big Match: A program for extracting probably matches from a large file for record linkage. U.S. Census Bureau, 2004. <http:// http://ww2.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000592.pdf>.

[29] Valêncio, C.R. et al.. Otimização de técnicas e algoritmos aplicados à limpeza de Banco de Dados utilizando Multithreading, in Proc. of Conferencia Ibero Americana WWW/Internet 2010 - IADIS, p.414-418, 2010.

[30] ANU Data Mining Group.: Febrl - FREELY Extensible Biomedical Record Linkage, 2012. <http://datamining.anu.edu.au/projects/linkage.html>.

[31] Christen, P.. A Freely Available Record Linkage System with a Graphical User Interface, in Australasian Workshop on Health Data and Knowledge Management - HDKM 2008. p.17–25, 2008.

[32] Elfeky, M.G.; Verykios, V.S.; Elmagarmid, A.K.. TAILOR: a record linkage toolbox, in Proc.of 18th International Conference on Data Engineering, 17–28, 2002. DOI: 10.1109/ICDE.2002.994694.

[33] Oliveira, P.; Rodrigues, F.; Henriques, P.. SmartClean: An Incremental Data Cleaning Tool, in Proc. of 9th International Conference on Quality Software - QSIC 2009, 452–457, 2009. DOI: 10.1109/QSIC.2009.67.

[34] Yan, H.; Diao, X.C.; Li, K.Q.. Research on Information Quality Driven Data Cleaning Framework, in Proc. Of International Seminar on Future Information Technology and Management Engineering - FITME´08, p.537–539, 2008. DOI: 10.1109/FITME.2008.126.

[35] Yu, H.; Yi, Z.X.; Zhen, Y.; Guo-quan, J.. A Universal Data Cleaning Framework Based on User Model, in Proc.of International Colloquium on Computing, Communication, Control and Management - CCCM 2009, p. 200–202, 2009. DOI: 10.1109/CCCM.2009.5267946.

[36] Bao, Y.; Song, J.; Shi, J.; Yu, G.. Case Study on Modeling Approaches and Framework of Scientific Data Cleaning, in Proc. of 9th IEEE International Conference on Computer and Information Technology - CIT'09, p.266–271, 2009. DOI: 10.1109/CIT.2009.88.

[37] Arasu, A.; Kaushik, R.. A grammar-based entity representation framework for data cleaning, in Proc. of ACM SIGMOD International Conference on Management of data - SIGMOD'09, p.233-244, 2009. DOI: 10.1145/1559845.1559871.

[38] Suriadi, S. et al.. Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. Information Systems, vol. 64, p.132-150, 2017. DOI: 10.1016/j.is.2016.07.011.

[39] Shi, Y.; Li, S.. The design and implementation of dynamic data cleaning modeling, in Proc. of 2010 International Conference on Computer Application and System Modeling – ICCASM, 2010. DOI: 10.1109/ICCASM.2010.5619213.

[40] Ali, K.; Warraich, M.A.. A framework to implement data cleaning in enterprise data warehouse for robust data quality, in Proc. of 2010 International Conference on Information and Emerging Technologies - ICIET 2010, p.1–6, 2010. DOI: 10.1109/ICIET.2010.5625701.

[41] Berti-Equille, L.; Dasu, T.; Srivastava, D.. Discovery of complex glitch patterns: A novel approach to Quantitative Data Cleaning, in Proc. of 27th International Conference on Data Engineering – ICDE, p.733 –744, 2011. DOI: 10.1109/ICDE.2011.5767864.

[42] Ciszak, L.. Application of clustering and association methods in data cleaning, in Proc. of International Multiconference on Computer Science and Information Technology - IMCSIT 2008, p. 97–103, 2008. DOI: 10.1109/IMCSIT.2008.4747224.

[43] Qian, X. et al.. Data cleaning approaches in Web2.0 VGI application, in Proc. of 17th International Conference on Geoinformatics, p. 1–4, 2009. DOI: 10.1109/GEOINFORMATICS.2009.5293442.

[44] Wang, H.. The Research of Outlier Data Cleaning through Relevance Comparison, in Proc. of 2nd International Conference on e-Business and Information System Security – EBISS, 1–3, 2010. DOI: 10.1109/EBISS.2010.5473717.

[45] Okita, T.. Data cleaning for word alignment, in Proc. of ACL-IJCNLP Student Research Workshop, p.72–80, 2009.

[46] Bertossi, L.; Kolahi, S.; Lakshmanan, L.V.S.. Data cleaning and query answering with matching dependencies and matching functions, in Proc. of 14th International Conference on Extending Database Technology - ICDT'11, p.268-279, 2010.

[47] Chaturvedi, S. et al.. Optimal Training Data Selection for Rule-Based Data Cleansing Models, in Proc. of 2011 Annual SRII Global Conference, p.126–134, 2011. DOI: 10.1109/SRII.2011.25.

[48] Prasad, K.H. et al.. Data Cleansing Techniques for Large Enterprise Datasets, in Proc. of 2011 Annual SRII Global Conference, p.135–144, 2011. DOI: 10.1109/SRII.2011.26.

[49] SQLite: SQLite. http://www.sqlite.org.

[50] Yu, Z., Bai, C.; Cai, K.Y.. Mutation-oriented test data argumentation for GUI software fault localization. Information and Software Technology, vol.55, no.12, p.2076-2098, 2013. DOI: 10.1016/j.infsof.2013.07.004

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.6 n.25, p. 16-20, Sep/Oct, 2020

DOI: https://doi.org/10.5935/jetia.v6i25.686

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

PREVENTIVE DIAGNOSIS OF PHOTOVOLTAIC FACILITIES BASED ON THE INTERPRETATION OF INFRARED IMAGES

Sandy Morales Galvez[1] and Alain Martinez*[2]

[1] Photovoltaic Department, Hydroenergy Company, Santa Clara, Cuba.
[2] Central University Las Villas - UCLV, Santa Clara, Cuba.

[1] http://orcid.org/0000-0003-4746-2192, [2] http://orcid.org/0000-0002-6873-126X

Email: smgalvez@outlook.es, *alainmlaguardia@gmail.com

ARTICLE INFO

Article History
Received: August 21th, 2020
Accepted: October 19th, 2020
Published: October 30th, 2020

Keywords:
Photovoltaic parks,
Diagnostics,
Infrared images.

ABSTRACT

Currently, Cuba is struggling to increase the presence of renewable energies in its energy matrix, this in order to reduce dependence on fossil fuels and help curb global climate change. One of the missions of the National Electric Union is to expand the installation of photovoltaic parks throughout the national territory and achieve efficient use of them. The present work aims to present examples of the diagnosis of different types of failures in such facilities, based on the interpretation of infrared images by means of the thermal contrast method It seeks the analysis of non-destructive infrared images of photovoltaic installations, with the goal of detecting failures due to heating of the parts that compose them. The execution of this type of preventive diagnostics, helps to plan maintenance or to carry out immediate repair actions that allow maintaining the efficiency and generation capacity of the photovoltaic parks.

I. INTRODUCTION

Renewable energy sources are part of the solutions for the "sustainable development" that humanity seeks, a development that responds to the needs of today without compromising the ability of future generations to respond to their own [1]. The use of renewable energy also contributes to the global effort to curb climate change, materialized at the political level by the implementation of the Kyoto protocol in 2005 and the signing of the Paris Agreement in 2015 [2]. The Latin American region, despite having a relatively low level of CO2 emissions and a high presence of the use of renewable energy (approximately 28% of the total used), has low diversification of sources, focusing on hydroelectric power [2]. Other unconventional sources such as solar, wind and biomass are little diversified in this geographic region, despite having extensive research on their use in recent years [3].

Cuba, unlike other countries in the region, does not have big rivers that allow boosting the use of hydroelectric energy, for this reason for years the energy matrix relied on thermoelectric plants that use fossil fuels [4]. The 21st century saw an increase in oil prices and with it an increase in the cost of producing electricity. Added to this situation was a greater awareness of the effects of climate change and the fact that improvements in the efficiency of non-conventional renewable energy sources appeared [5]. This environment caused the National Electricity Union (UNE), the entity responsible for operating and managing power generation in the country, began an assessment and gradual introduction of non-conventional energy sources in the energy matrix of the country [4, 6].

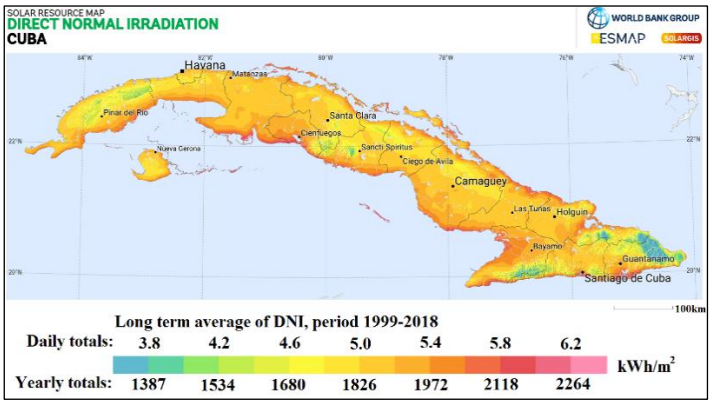

Figure 1: Cuba direct normal irradiation map.
Source: [7].

Within renewable energies, photovoltaics, has had wide national attention in recent years [6, 8-17]. This is due to the ease of its deployment in a tropical country like Cuba, with high levels of solar radiation (Figure 1). Small size photovoltaic systems, interconnected to the grid, have been installed throughout the country [6, 8, 13]. In the same way, isolated micro systems have allowed reaching remote communities that for years were isolated from the national network due to geographic complexities [18], with the consequent improvement in their quality of life. Economic and social objectives have also been enhanced with the installation of these systems to allow them energy autonomy in contingency situations [15].

The deployment of photovoltaic technology has required strategies to be drawn up for its efficient management [12, 17, 19]. Among them, the study of new preventive diagnostic techniques that allow achieving high efficiency rates and a reliable operation. Among the preventive diagnostic techniques, infrared thermography is a very useful tool. It is supported by the images acquired with a thermal camera. In these images, each pixel of the photographed object makes radiometric reference to its temperature. The images, once corrected according to the conditions of the environment (humidity and air temperature, distance to the analyzed object, reflected temperature, incident radiation), allow obtaining a map of the temperature of the object of interest at a distance, accurately and without the need for physical contact with it.

An accurate thermography provides maintenance personnel with the information required to avoid or lessen failures in photovoltaic installations, as a result of the most common faults in their elements. It is noteworthy that thermography, unlike other diagnostic techniques, does not require the stop of the photovoltaic solar park (PVSP) and is carried out with it fully operational. Only simple procedures should be followed that involve the environmental conditions at the time the images are taken and the way the images are acquired.

II. MATERIALS AND METHODS

Photovoltaic systems are made up of photovoltaic panels, electrical elements and electronic devices that make it possible to transform solar radiation into useful electrical energy for consumers. In them, the interconnected photovoltaic panels generate a direct current voltage with which an electronic inverter is supplied that supplies the alternating current normally used in distribution. Photovoltaic systems have a long useful life and are capable (according to their technology) to integrate with different types of architectural structures [20, 21]. In the same way, they are capable of operating with both direct and diffuse solar radiation.

The basis of a photovoltaic system are photovoltaic panels or modules, these contain solar cells (photovoltaic cells) based on semiconductors sensitive to solar radiation, which they transform into electricity. The materials used in the construction of solar cells can vary depending on the application for which they are intended, the most common being silicon, either in a single crystalline, polycrystalline or amorphous structure [20-22]. The solar cells are grouped in series or parallel to achieve the voltage and output power required by the photovoltaic module to which they are integrated.

However, when a cell is damaged or does not generate energy because it does not receive enough solar radiation, it can be polarized in reverse, becoming a charge instead of a generator, which can imply a high dissipation of energy in the form of heat, the so-called "hot spots" [23]. This situation is easily detectable if an infrared thermography is performed, in the same way faulty splices or welds with false contact can be detected [24-26].

In order to achieve quality thermographs, which provide information for decision-making, certain measurement procedures and conditions must be taken into account:

- A thermal imaging camera with the proper accessories and the correct resolution must be used.
- Sufficient solar radiation is required (at least 500 W/m^2, preferably more than 700 W/m^2).
- The viewing angle must be within the safety margins (between 5° and 60°).
- Shading and reflections should be avoided.

The UNE has Testo 875-i infrared cameras for the inspection of the photovoltaic parks currently installed in the country. This camera offers the following features: resolution of 160x120 pixels, a viewing angle of 32x23 degrees, thermal sensitivity of less than 50mK and side-by-side display of the image acquired in the visible and infrared spectrum. The manufacturer has included in its firmware an application for automatic detection of hot and cold spots, along with other adjustment options to speed up the diagnostic process. Likewise, the "Testo IRSoft" software is included for the processing of images acquired on a PC. The software allows to compensate factors such as the ambient temperature, the relative humidity and the distance between the camera and the photographed object.

Starting from the images obtained with the Testo 875-i camera, the operation specialist´s proceed to apply the thermal contrast method to evaluate the detected incidents in it. The thermal contrast method in its simplest mode can be defined as the difference between the temperature in an area under normal conditions (Tsa), and a region with hot spots (Td), as presented in Equation 1. This method, also known as "absolute thermal contrast" and several variants of it, are widely used by professionals in the energy sector and specialists in preventive diagnosis of photovoltaic systems [27-29].

$$\Delta T = Td\ (t) - Tsa(t) \tag{1}$$

This method and its variants suffer from a common lack, that is, the need to find what will be considered the healthy zone of the solar module. In addition, for the correct analysis of the thermal images, the nominal performance of the panel at the time of inspection and the maximum operating temperature must be taken into account. An adequate assessment of all these factors allows the operation specialist´s to conduct an estimated diagnosis of the safety risks related to module's operating situation. The safety risks, based on the thermal contrast information are divided into several scales:

Td – Tsa ≤ 10ºC → Relevance **Normal**
10ºC < Td – Tsa ≤ 20ºC → Relevance **Low**
20ºC < Td – Tsa ≤ 40ºC → Relevance **Moderate**
40ºC < Td – Tsa ≤ 70ºC → Relevance **Severe**
Td – Tsa > 70ºC → Relevance **Critical** (**Safety hazard**)

Normal: No action is required until the next predictive study.

Low: Monitor to see the evolution of the hot spot using the most appropriate methodology.

Moderate: Act as soon as possible taking into account the dynamics of the power plant and its work shifts, wait for a planned shutdown to correct the problem.

Severe: Urgently study the possibility of stopping the process to correct the problem.

Critical: Stop the process immediately to correct the problem.

In the last two cases, the inspection report will be notified in advance to the power plant administration, in order to fix the problem before it's too late.

In a similar way to the risk analysis, the detection and location of the existing degradations in the solar modules, allows the realization of diagnoses to determine the state of efficiency of the same.

In the next section the inspection of a photovoltaic power plant in operation is carried out by means of the analysis of infrared images.

III. THERMOGRAPHY BASED DIAGNOSTIC EXAMPLES

The "Troncoso 1" PVSP, in the Province of Pinar del Rio, was selected to perform the diagnostic and extract examples of typical faults. This PVSP is based on the DSM-240-C photovoltaic modules of national manufacture [30]. The modules are composed of 60 polycrystalline silicon cells in 156mm X 156mm format, connected in series to provide 29.8V and a current of 8.2A at the maximum power point.

The images were taken manually, with an average solar radiation between 800 W/m^2 and 1000 W/m^2, the software IR Soft V3.3 was used for the analysis of the images. It is valid to emphasize that since each material has a different emissivity, it is necessary to compare temperatures of elements formed by the same material with nominal temperature (denoted as PF1 in the analyzed thermal images).

In Figure 2, on the left side, the captured thermal image is observed (taken from the back of the photovoltaic module), while on the right the same image is shown in the visible spectrum. The IR soft software highlights areas with temperature variations. The thermal inspection carried out shows that there are several cells with evidence of being affected with temperature differences "hot spots" (Table 1).

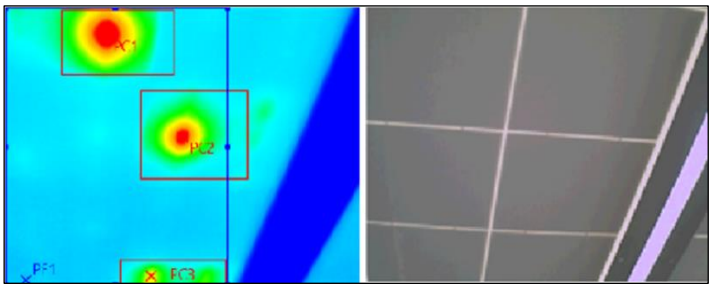

Figure 2: Burns on solar module caused by severe hot spots.
Source: Author, (2020).

Table 1: Markers referring to the analysis of Figure 2.

Name	Temperature	Emissivity	Reflected temperature
PF1	58.9	0.93	20.0
PC1	111.3	0.93	20.0
PC2	95.5	0.93	20.0
PC3	76.3	0.93	20.0

Source: Author, (2020).

The analysis of the thermal image and its visible counterpart shows that the high temperatures detected are associated with two welding tapes on the positive side of the solar cell. It should be noted that the image in the visible spectrum does not show visible damage. The nominal temperature of the module was 58.9 ºC, so the safety risks is considered: "severe", being necessary to plan corrective actions.

Figure 3 shows a similar case where no visible image damage is observed (right site of image), however, there are hot spots with temperature differences that mark a "severe" safety risk (Table 2). The high temperatures are located in three solder tapes on the positive side of the solar cell.

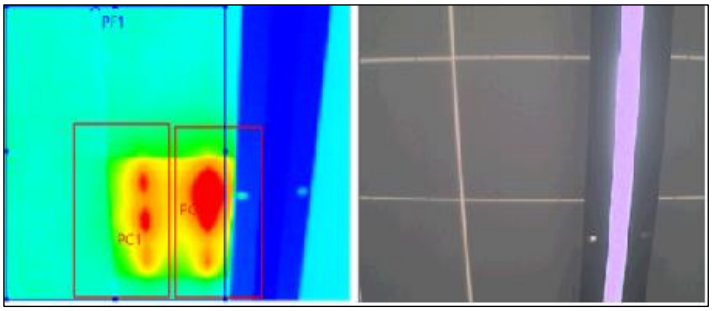

Figure 3: Burns on solar module caused by severe hot spots.
Source: Author, (2020).

Table 2: Markers referring to the analysis of Figure 3.

Name	Temperature	Emissivity	Reflected temperature
PF1	64.6	0.95	20.0
PC1	93.2	0.95	20.0
PC2	123.3	0.95	20.0

Source: Author, (2020).

Figure 4 and 5 shows modules with cells in a state of severe security risk (Tables 3 and 4), the damage can be perceived with the naked eye on the right side of Figure 4 and 5.

Figure 4: Cells with critical damage.
Source: Author, (2020).

Table 3: Markers referring to the analysis of Figure 4.

Name	Temperature	Emissivity	Reflected temperature
PF1	50.4	0.95	20.0
PC1	130.0	0.95	20.0
PC2	82.0	0.95	20.0

Source: Author, (2020).

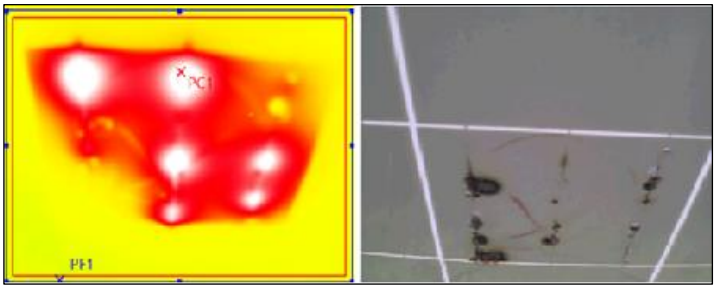

Figure 5: Cells with critical damage.
Source: Author, (2020).

Table 4: Markers referring to the analysis of Figure 5.

Name	Temperature	Emissivity	Reflected temperature
PF1	50.4	0.95	20.0
PC1	130.0	0.95	20.0

Source: Author, (2020).

Figures 6, 7 and 8, unlike the previous ones, were taken from the front of the panel, in them is possible to observe other affectations such as fractures in the glass covert, dirt and oxidation.

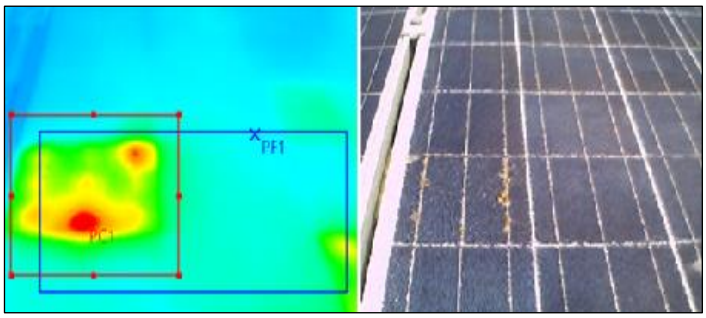

Figure 6: Fractures in the glass covert of the PV module.
Source: Author, (2020).

Table 5: Markers referring to the analysis of Figure 6.

Name	Temperature	Emissivity	Reflected temperature
PF1	52.1	0,93	24.0
PC1	95.4	0,93	24.0

Source: Author, (2020).

The cell damaged by the fractures in the glass covert is clearly identified in the visible spectrum image and its thermal counterpart shows severe damage to it (Table 5).

Figure 7 shows a PV module covered partially by dirt, in it, a loss of efficiency occurs and hot spots appear on its surface, with danger to the integrity of the cells. The impact on the module is considered moderate (Table 6) and a simple cleaning procedure could correct the situation.

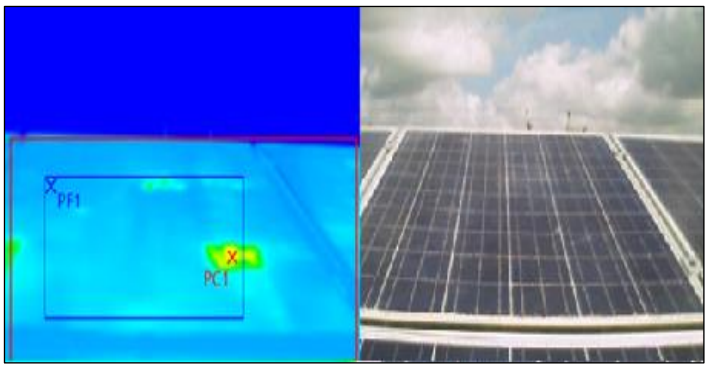

Figure 7: PV module covered partially by dirt.
Source: Author, (2020).

Table 6: Markers referring to the analysis of Figure 7.

Name	Temperature	Emissivity	Reflected temperature
PF1	43.7	0.95	20.0
PC1	69.8	0.95	20.0

Source: Author, (2020).

Figure 8 shows the oxidation of a cell on the edge of a photovoltaic module. The modules are designed to be waterproof, but damage to the sealing system could lead to the entry of rainwater and the consequent damage to the cells.

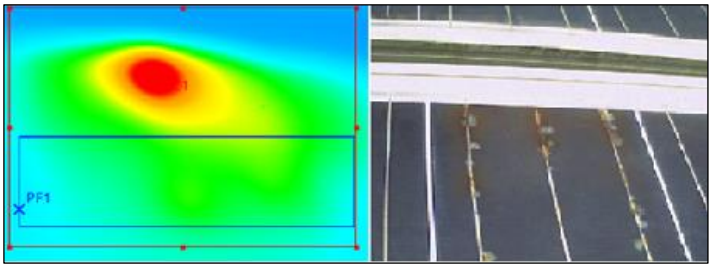

Figure 8: Oxidation of a cell on the edge of a photovoltaic module.
Source: Author, (2020).

Table 7 shows how a hot spot of severe magnitude has emerged in the area in question.

Table 7: Markers referring to the analysis of Figure 8.

Name	Temperature	Emissivity	Reflected temperature
PF1	54.5	0.95	20.0
PC1	100.1	0.95	20.0

Source: Author, (2020).

As can be seen throughout the exposed examples, thermography is a very useful tool when making a preventive diagnosis of photovoltaic installations. At present, its extensive and frequent use is limited by the manual acquisition and processing of the images. A qualified operator normally uses four days for the acquisition of the images associated with a PVSP of 1MW of power, to this must be added the time for image processing that far exceeds the acquisition time. The use of unmanned aerial vehicles (UAV) equipped with thermal imaging cameras and automated image processing software could be a solution to this limitation by reducing the acquisition and processing time [31-34].

IV. CONCLUSIONS

The use of infrared imaging helps to identify damaged areas on photovoltaic modules that are not detectable with the naked eye. With the help of the information obtained from them, the operating personnel can plan actions to avoid serious failures to the facilities.

The thermal contrast method, despite its simplicity, is valid to highlight the defects that may occur in a photovoltaic module, facilitating the diagnostic work. Expanding the use of preventive diagnosis based on thermal images in solar photovoltaic parks installed in our country, can result in better efficiency and use of installed technology.

V. FUTURE WORK

The Automatic Control Department of the Universidad Central “Marta Abreu” de Las Villas features an infrared camera NEC F30, which meets the weight requirements to be integrated into an X8+ UAV also available at the department. Future works will be dedicated to the integration process in order to be able to carry out the inspection of photovoltaic parks using aerial images, reducing exploration times.

VI. AUTHOR’S CONTRIBUTION

Conceptualization: Alain Martinez.
Methodology: Alain Martinez.
Investigation: Sandy Morales Galvez.

Discussion of results: Sandy Morales Galvez.
Writing – Original Draft: Sandy Morales Galvez.
Writing – Review and Editing: Alain Martinez.
Resources: Sandy Morales Galvez.
Supervision: Alain Martinez.
Approval of the final text: Alain Martinez.

VII. REFERENCES

[1] W. H. Organization, "Tracking SDG 7," 2019.

[2] C. Washburn and M. Pablo-Romero, "Measures to promote renewable energies for electricity generation in Latin American countries," Energy policy, vol. 128, pp. 212-222, 2019.

[3] G. Valencia Ochoa, J. Nunez Alvarez, and C. Acevedo, "Research evolution on renewable energies resources from 2007 to 2017: a comparative study on solar, geothermal, wind and biomass energy," International Journal of Energy Economics and Policy, vol. 9, pp. 242-253, 2019.

[4] L. Vazquez, Y. Majanne, M. Castro, J. Luukkanen, O. Hohmeyer, M. Vilaragut, and D. Diaz, "Energy System Planning towards Renewable Power System: Energy Matrix Change in Cuba by 2030," IFAC-PapersOnLine, vol. 51, pp. 522-527, 2018.

[5] D. Zhu, S. M. Mortazavi, A. Maleki, A. Aslani, and H. Yousefi, "Analysis of the robustness of energy supply in Japan: Role of renewable energy," Energy Reports, vol. 6, pp. 378-391, 2020.

[6] D. D. Guerra, E. V. Iakovleva, and A. Y. Shklyarskiy, "Alternative Measures to Reduce Carbon Dioxide Emissions in the Republic of Cuba," Journal of Ecological Engineering, vol. 21, p. 4, 2020.

[7] SOLARGIS. "Solar resource maps of Cuba". The World Bank, Global Solar Atlas 2.0, Solar resource data: Solargis. 2019. Available in: https://solargis.com/maps-and-gis-data/download/cuba

[8] J. E. Alonso Trimiño, "Análisis de los parques fotovoltaicos en las redes eléctricas de Santa Clara," Ingeniería Eléctrica, Universidad Central" Marta Abreu" de Las Villas, Facultad de Ingeniería Eléctrica, 2018.

[9] R. M. Arias García and I. Pérez Abril, "Nueva metodología para determinar los parámetros de un módulo fotovoltaico," Ingeniería Energética, vol. 39, pp. 38-47, 2018.

[10] O. C. Castillo and A. R. S. Sera, "Influencia combinada del espaciamiento y la inclinación de módulos en generación fotovoltaica," Revista Cubana de Ingeniería, vol. 8, pp. 29-34, 2017.

[11] J. R. Chantres Borges, "Comparación de algoritmos MPPT aplicados a sistemas fotovoltaicos," Control Automático, Universidad Central" Marta Abreu" de Las Villas, Facultad de Ingeniería Eléctrica, 2018.

[12] R. Díaz Santos, M. Castro Fernández, A. Santos Fuentefría, and M. Vilaragut Llanes, "Análisis de la influencia del ángulo de inclinación en la generación de una central fotovoltaica," Ingeniería Energética, vol. 39, pp. 146-156, 2018.

[13] Y. A. Gallego Landera, R. Arias García, L. Casas Fernández, and R. Sosa Plasencia, "Analisis de la implementacion de un parque fotovoltaico en la Universidad Central de las Villas," Ingeniería Energética, vol. 39, pp. 82-90, 2018.

[14] J. A. G. Gutiérrez, J. P. Bejerano, and Á. R. D. Deulofeu, "Riesgos ambientales en parques solares fotovoltaicos del occidente de Cuba," Revista ECOVIDA, vol. 9, pp. 195-211, 2020.

[15] M. Guzmán Villavicencio, C. R. Soto Castellón, I. Águila Bernal, and J. M. Torres Águila, "Procedimiento para instalación de un sistema fotovoltaico sobre techos en la corporación cuba ron sa," Centro Azúcar, vol. 44, pp. 70-81, 2017.

[16] A. López, V. Parnás, and J. Cataldo, "Experimentos en túnel de viento sobre paneles fotovoltaicos montados en el suelo," Revista ingeniería de construcción, vol. 34, pp. 15-24, 2019.

[17] J. R. Núñez, I. Benítez, R. Proenza, L. Vázquez, and D. Díaz, "Metodología de diagnóstico de fallos para sistemas fotovoltaicos de conexión a red," Revista Iberoamericana de Automática e Informática industrial, vol. 17, pp. 94-105, 2020.

[18] J. A. Cherni and Y. Hill, "Energy and policy providing for sustainable rural livelihoods in remote locations–The case of Cuba," Geoforum, vol. 40, pp. 645-654, 2009.

[19] B. Pedroso Mestre, "Análisis de la degradación de módulos fotovoltaicos tras 7 años de operación en Santiago de Cuba," Departamento de Eléctrica, 2019.

[20] M. Alaaeddin, S. Sapuan, M. Zuhri, E. Zainudin, and F. M. Al-Oqla, "Photovoltaic applications: Status and manufacturing prospects," Renewable and Sustainable Energy Reviews, vol. 102, pp. 318-332, 2019.

[21] S. Mughal, Y. R. Sood, and R. Jarial, "A review on solar photovoltaic technology and future trends," Int J Sci Res Comput Sci Eng Inform Technol (IJSRCSEIT-2018), vol. 4, 2018.

[22] M. Boulaid, A. Tihane, R. Oaddi, A. Elfanaoui, K. Bouabid, and A. Ihlal, "Comparative performance assessment of mono crystalline, multi crystalline, and amorphous silicon grid-connected photovoltaic systems under actual climatic conditions of Agadir, Morocco," International Journal of Green Energy, vol. 14, pp. 1182-1191, 2017.

[23] S. Deng, Z. Zhang, C. Ju, J. Dong, Z. Xia, X. Yan, T. Xu, and G. Xing, "Research on hot spot risk for high-efficiency solar module," Energy Procedia, vol. 130, pp. 77-86, 2017.

[24] M. Cubukcu and A. Akanalci, "Real-time inspection and determination methods of faults on photovoltaic power systems by thermal imaging in Turkey," Renewable Energy, vol. 147, pp. 1231-1238, 2020.

[25] S. Gallardo-Saavedra, L. Hernádez-Callejo, and Ó. Duque-Pérez, "Analysis and characterization of PV module defects by thermographic inspection," Revista Facultad de Ingeniería Universidad de Antioquia, pp. 92-104, 2019.

[26] Á. H. Herraiz, A. P. Marugán, and F. P. G. Márquez, "A review on condition monitoring system for solar plants based on thermography," in Non-Destructive Testing and Condition Monitoring Techniques for Renewable Energy Industrial Assets, ed: Elsevier, 2020, pp. 103-118.

[27] J. A. Tsanakas, L. Ha, and C. Buerhop, "Faults and infrared thermographic diagnosis in operating c-Si photovoltaic modules: A review of research and future challenges," Renewable and Sustainable Energy Reviews, vol. 62, pp. 695-709, 2016.

[28] G. Cipriani, V. Boscaino, V. Di Dio, F. Cardona, G. Zizzo, and S. Di Caro, "Application of Thermographic Techniques for the Detection of Failures on Photovoltaic Modules," in 2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), 2019, pp. 1-5.

[29] A. Sinha, O. Sastry, and R. Gupta, "Detection and characterisation of delamination in PV modules by active infrared thermography," Nondestructive Testing and Evaluation, vol. 31, pp. 1-16, 2016.

[30] G. d. I. CCE. (2020, 10-08-20). Empresa de Componentes Electrónicos "Ernesto Che Guevara". Available: https://www.cce.cu/

[31] Y. Zefri, A. ElKettani, I. Sebari, and S. Ait Lamallam, "Thermal infrared and visual inspection of photovoltaic installations by UAV photogrammetry—application case: morocco," Drones, vol. 2, p. 41, 2018.

[32] P. Addabbo, A. Angrisano, M. L. Bernardi, G. Gagliarde, A. Mennella, M. Nisi, and S. L. Ullo, "UAV system for photovoltaic plant inspection," IEEE Aerospace and Electronic Systems Magazine, vol. 33, pp. 58-67, 2018.

[33] A. Niccolai, A. Gandelli, F. Grimaccia, R. Zich, and S. Leva, "Overview on photovoltaic inspections procedure by means of unmanned aerial vehicles," in 2019 IEEE Milan PowerTech, 2019, pp. 1-6.

[34] G. Francesco, L. Sonia, and N. Alessandro, "A semi-automated method for defect identification in large photovoltaic power plants using unmanned aerial vehicles," in 2018 IEEE Power & Energy Society General Meeting (PESGM), 2018, pp. 1-5.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.6 n.25, p. 21-26, Sep/Oct, 2020

DOI: https://doi.org/10.5935/jetia.v6i25.692

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

A COOPERATIVE GAME THEORY APPLICATION IN CHICKS BROOD FOOD ALLOCATION BY USING SHAPLEY VALUE METHOD IN GOOD YEARS DATA

Obaidul Haque*[1], Anwar Hossen[2] and Sharmin Akter[3]

[1, 3] Department of Mathematics, Comilla University - UNIC, Bangladesh.
[2] Associate Professor, Department of Mathematics, Comilla University - UNIC, Bangladesh.

[1] https://orcid.org/0000-0003-4455-6708 [2] http://orcid.org/0000-0002-5822-8628 [3] http://orcid.org/0000-0001-9293-0552

Email: *obaidul.cou@yahoo.com, mdahossen@yahoo.com, sharmincou@gmail.com

ARTICLE INFO

Article History
Received: September 04th, 2020
Accepted: October 19th, 2020
Published: October 30th, 2020

Keywords:
Cooperative Game,
Shapley Value,
Player,
Coalitions,
Good Years.

ABSTRACT

Game theory as well as cooperative game theory has played a vital role in many fields of research since its introduction within the early twentieth century. During this paper we study food allocation in chick broods from the attitude of cooperative theory of games. We would like to explore whether or not food distribution data fit into the known solution concepts of cooperative game theory to create an economic temperature within the biological field. A primary approach to be handled is that the incontrovertible fact that within the chick brood data we only see the solutions, while the starting position, the game, isn't immediately clear. In and of itself we'd like to reconstruct the game from the solutions given. A second approach is that by using the answer concepts Shapley value we would like to investigate which of those fits best. Most interesting is to specifically address the properties that cause these solutions because these would be most useful to find motivation for the particular solution concept found in nature. The goal is to anticipate moves to create, which is able to cause ultimate victory.

I. INTRODUCTION

Game theory is an autonomous discipline that is used in applied mathematics, social sciences, most remarkably in mathematics, economics, as well as in computer science, biology, engineering, international relations, philosophy, and political science [1]. It was displayed in economics and mathematics to measure economic behaviors including behaviors of firms, markets, and consumers. The cooperative game theory can be contemplated as a modeling procedure that is used to analyze and explain the actions of all players joined in competitive situations and to compare and determine the relative optimality of distinct strategies. The first inspect of games in terms of economics was by Cournot on pricing and production but Neumann (1944) is considered as the founder of the modem game theory [1]. Coined by Shapley (1953), this one-point solution concept introduced in his paper has some desirable properties called efficiency, anonymity, dummy, and additivity [4]. A solution is efficient if it assigns to every game an allocation in such a way that the sum of every marginal contribution of each player will be equal to the value of the grand coalition [1, 2, 3]. Recent works by Forbes (2005, 2007) is employing a financial tool to the study of parental investment in chick, as normally the foremost important investment any organism makes is in its offspring [5]. In cooperative game theory, it can be said that the set of bounded computational capacity of equilibrium payoffs carries only one valuation, that the valuation of the game with penalty approaches the valuation of the one-shot game as the penalty goes to zero [1, 6].

I.1 NOTATIONS

eggs: Shows how many eggs are firstly available in total before the hatching period.
c: In day 1, denotes the number of core eggs (eggs that are hatched in day 1) inside one brood.
m: The number of marginal eggs (eggs that are left/not yet hatched in day 1) in one brood.

brood: Shows how many broods that are available in total.
$\boldsymbol{m_1}$**:** The number of marginal that are hatched on the first day after the core.
$\boldsymbol{m_2}$**:** The number of marginal that are hatched on the second day after the core.
$\boldsymbol{m_3}$**:** The number of marginal that are hatched on the third day after the core.
$\boldsymbol{total}$: In day 8, shows how many chicks that are continue living after one week of feeding.
$\boldsymbol{d8\ val}$: The average number of all chick (both core and marginal) which survive until one week of feeding (i.e. day 8).
$\boldsymbol{m\ av}$: The average of the marginal chick' survival rate which we will use to fill the coalition value for the marginal later in our method.
$\boldsymbol{c\ val}$: The survival rate of each chick.
$\boldsymbol{m_1}$ **val**: The survival rate of starting from the core chick.
$\boldsymbol{m_2}$ **val**: The survival rate of the first marginal chick.
$\boldsymbol{m_3}$ **val**: The survival rate of until the third marginal chick.
$\boldsymbol{v}(\boldsymbol{S})$: The value of the coalition.
$\boldsymbol{A}$: A matrix with real entries a_{ij}.
$\boldsymbol{x}$: A column vector with real entries x_i.
$\boldsymbol{x^S}$: The payoff vectors.
$\boldsymbol{x^T}$: The transpose of vector **x.**
$\boldsymbol{\{x_1, x_2\}}$: The sequence of vectors x_1, x_2
$\boldsymbol{x^*}$: The optimal value of vector **x**
∅: Shapley value

II. STUDY ON CHICK BROOD FAIR ALLOCATION PROBLEM

Recent works by Forbes is employing a financial tool to the study of parental investment in chick, as normally the foremost important investment any organism makes is in its offspring [5]. A key dimension of any investment decision including what proportion to speculate in offspring is a way to balance risk and reward that portfolio theory offers a broad set of analytical tools. An initial complexity for the biologist introduces in a way to translate the economic models into a biological patent. The tool he used is named financial beta and is well-known to the study of parental investment, derived from the capital asset pricing model of recent portfolio theory. Beta provides a measure of the volatility within the price of an asset for a wide market or index. Forbes suggested that the reproductive returns from individual brood structures (e.g., mean edging success in an exceedingly given year) may well be usefully equated to a private asset, which means population reproductive success may well be equated to the market as an entire.

There is another study conducted by Alex Kacelnik, Peter A. Cotton, Liam Stirling, and Jonathan Wright which use the evolutionary theory of games to review Food Allocation among Nestling Starlings, drawing attention on Sibling Competition and also the Scope of Parental Choice. Chick feeding in chick is commonly viewed as a chief example of evolutionary conflict. this can be because the nestlings may benefit by inducing the parent to speculate more within the current brood compared to future ones. additionally, each nestling should benefit by obtaining a greater fraction of the full brood provision than would be optimal for the parent. Current theory suggests that at evolutionary equilibrium, the intensity of signalling (i.e. begging) by the chick should allow the fogeys to spot each chick's needs and to allocate more food to the one that gives the steepest marginal fitness gain per unit of parental resources [5].

III. THE CHICK BROOD DATA IN GOOD YEARS

The chick are puppies who have hatched together on the first day of nesting. Instead, chick hatch for more than a day and are identified as marginal chick. Parental choices for the number of eggs hatching on the first day may be based on their experiences during the early hatching period, or their instincts about the weather and food conditions near the nest. They hit their marginal according to their core every day. Therefore, if you have 2 core chick and 3 margins, the total hatch time is 4 days (1 day for all cores and 3 days for each margin). The raw data in Table 1 below will give you an idea of the number of chickens and chickens that one chick can have, and the number of chick available to chickens and chick. Be found. You can also see how many chicks died in a good week (from day 1 to day 8), so you can see if there are any differences in how parents are assigned.

Table 1: The good year data.

eggs	Dav 1			Dav 8					
	c	m	brood	c	m	m_1	m_2	m_3	total
18	1	0	15	15	0	0	0	0	15
59	1	1	29	27	27	27	0	0	53
120	1	2	40	38	62	33	29	0	95
218	1	3	55	54	124	52	48	24	176
30	2	0	15	28	0	0	0	0	28
236	2	1	78	134	56	56	0	0	186
430	2	2	108	208	140	85	55	0	345
75	2	3	15	29	26	14	9	3	55
81	3	0	27	74	0	0	0	0	74
292	3	1	73	196	33	33	0	0	224
100	3	2	20	55	16	12	3	0	71
32	4	0	8	27	0	0	0	0	27
25	4	1	5	19	2	2	0	0	21

Source: Authors, (2019).

As we've seen within the above data, one brood can contain at maximum 4 core chick, while on the opposite hand, it may also contain at maximum 3 marginal. While the whole number of chick is seven (i.e. there are seven players within the game), we don't have data with four core and three marginal at the identical time. The largest brood we've consisted of 5 chicks, either 2 core with 3 marginals, 3 core with 2 marginals, or 4 cores with 1 marginal. Note that from the story of chick the fogeys will feed the chick that beg louder, which usually are the core chick. Therefore, in a method to calculate the Shapley value, we fancy assuming that the feeding process will always start with the core chick, while the marginals 'fight' over the remaining food after the cores are being fed. This assumption is described later during calculations and experiments of our method. In theory, we also cannot have only marginals without having the core, or having the third and/or the second marginal without having the primary one. But this can be happening in a number of the brood data since there's a break that the egg is missing or being destroyed during the hatching period of the core chick, to not mention the chick that's directly dead after born, leaving only the marginals within the brood. The same case is additionally happening for the marginal. However, later we are going to see that our method excludes this sort of missing data from the calculations, and considers only the feasible coalitions. As a result of being born on different days where the marginals are hatched on a daily basis after the cores, we expect our brood data have a particular property: there exist different weights between core and marginal chick. this can be because the core and also the marginal chick may value their food in numerous ways. We predict that everyone the core chick's c will value their food within the same manner since they're hatched on an identical day (thus could also be as

strong as each other) and also the parents consider them to be equally important to continue the family legacy. As a result, the weights difference between the core chick is incredibly small or is ignored. In other words, we assume that competition between the core chick in one brood doesn't exist. However, there exist different weights between the core and also the marginals, in addition to between all the marginals, because the marginal m_i, i=1 to 3, are born consecutively on i days after the core. Note that this weight isn't a body-mass index but a further number that represents how the chick may value their food. To be ready to build a coalition model for the brood food allocation data during the great and also the bad years, we firstly we define an XY-brood game where X and Y denote the quantity of core and marginal children respectively.

As input for the chick brood food allocation model, we use the type of survival rate data A and also the importance weights data I which shows what percentage times a particular style of brood (i.e. core and marginal coalition) appears within the game.

Table 2: Average of the Survival Rate for XY-brood type during the Good Years.

A								I
c	**m**	**c val**	**m av**	**m_1 val**	**m_2 val**	**m_3 val**	**d8 val**	
1	0	1.000	0	0	0	0	1.000	15
1	1	0.931	0.931	0.931	0	0	1.862	29
1	2	0.950	0.838	0.892	0.784	0	2.626	40
1	3	0.982	0.765	0.963	0.889	0.453	3.287	55
2	0	0.933	0	0	0	0	1.867	15
2	1	0.859	0.747	0.747	0	0	2.465	78
2	2	0.963	0.688	0.794	0.514	0	3.234	108
2	3	0.967	0.578	0.933	0.600	0.200	3.667	15
3	0	0.914	0	0	0	0	2.741	27
3	1	0.895	0.465	0.465	0	0	3.150	73
3	2	0.917	4.000	0.632	0.158	0	3.539	20
4	0	0.844	0	0	0	0	3.375	8
4	1	0.950	4.000	4.000	0	0	4.200	5

Source: Authors, (2019).

Having the coalitions S, what value can we choose to be the value of the coalition v(S)? Since we have the average of the survival rate A for each off-springs in every XY-brood, taking into account its importance weight I (i.e. how many times the XY-brood data occur), we can take these values as the value of the coalitions.

Table 3: 21-brood game, Good years.

S	{1}	{2}	{5}	{1,2}	{1,5}	{2,5}	{1,2,5}
$v(S)$	1.000	1.000	0.931	1.867	1.862	1.862	2.465

Source: Authors, (2019).

As we are able to see within the brood data for the good years (Table 2), there are data of 10-brood, 11-brood, 12-brood, 13-brood, 20-brood, 21-brood, 22-brood, 23- brood, 30-brood, 31-brood, 32-brood, 40-brood, and 41-brood. to urge a sense of those data we are having, we are going to calculate the Shapley value ∅ by hand for a few of the smaller brood data, to determine if we are able to get something interesting as a result.

There are two different methods used to calculate the payment vector for the x^S for the coalition S, the key is the standard approach.

Remember that we've at maximum seven players (|N|=7), incorporates a maximum of 4 core players (i=1 to 4) and at maximum 3 marginal players (i=5 to 7).

Thus, we may consider seven different places for every different position of the players.

In this standard approach, the core chick is often placed in anywhere among the four first places while the marginals are placed consecutively within the three last places (in increasing order). Because the core may become the primary, the second, the third, or the fourth player, it is often placed in any of the four first places.

Therefore, we'd like to contemplate the identical survival rates for these four possible places of the core chick in each XY-brood game; Weight is important, but the ideal cover chicken is divided into four possible areas.

Note that whichever chick chooses the primary place is considered because of the first core, and so on. To be clear about this representation, we convert Table 2 of the good year's data into a replacement Table 4, by defining an allocation x^S because the average of each chick's survival rate within the corresponding XY-brood game, taking into consideration the possible coalitions that may be made by all the players involved within the game. For example, if we consider the good years 21-brood game using the quality approach, we are going to have a collection of possible coalitions that consists of a coalition. Since the 2 core chick can choose any of the four first places.

The Table 4 below will show what number players involved in each coalition of a selected XY-brood, what are the possible coalitions exist in a very specific XY- brood, and what are the survival rates of every player involves in those specific coalitions.

Note that the numbers i=1 to 7 in the table denote the players, where i=1 to 4 are core players and i=5 to 7 are marginal players.

Table 4: Survival rates of the players in the existing coalitions (Standard Approach, Good Years).

XY	**Possible Coalitions S**	X^s			
		1 to 4	**5**	**6**	**7**
10	$\{1\},\{2\},\{3\},\{4\}$	1.000	0	0	0
11	$\{1,5\},\{2,5\},\{3,5\},\{4,5\}$	0.931	0.931	0	0
12	$\{1,5,6\},\{2,5,6\},\{3,5,6\},\{4,5,6\}$	0.950	0.892	0.784	0
13	$\{i,5,6,7\}, \forall i = 1\ to\ 4$	0.982	0.963	0.889	0.453
20	$\{i,j\}, \forall i,j = 1\ to\ 4, i < j$	0.933	0	0	0
21	$\{i,j,5\}, \forall i,j = 1\ to\ 4, i < j$	0.859	0.747	0	0
22	$\{i,j,5,6\}, \forall i,j = 1\ to\ 4, i < j$	0.963	0.794	0.514	0
23	$\{i,j,5,6,7\}, \forall i,j = 1\ to\ 4, i < j$	0.967	0.933	0.600	0.200
30	$\{i,j,k\}, \forall i,j,k = 1\ to\ 4, i < j < k$	0.914	0	0	0
31	$\{i,j,k,5\}, \forall i,j,k = 1\ to\ 4, i < j < k$	0.895	0.465	0	0
32	$\{i,j,k,5,6\}, \forall i,j,k = 1\ to\ 4, i < j < k$	0.917	0.632	0.158	0
40	{1,2,3,4}	0.844	0	0	0
41	{1,2,3,4,5}	0.950	0.400	0	0

Source: Authors, (2019).

Note that in total we are going to have 54 coalitions if we consider the quality approach. Here we restrict n core players, n=1 to 4, to always be within the first n-places, the first n-places and the remaining 4-n places that seem to have been occupied by the core chick will have a zero value. as an example, if we have a 21-brood where there are two core players and one marginal, then the 2 core players will always fill the primary and also the second places, while the third and also the fourth places remain zero. that's we are going to have only 13 coalitions if we consider this approach. This number is the same because of the number of all existing XY-brood games.

Table 5: Survival rates of the players in the existing coalitions (Restricted Approach, Good Years).

X Y	Possible Coalition S	X^s						
		1	2	3	4	5	6	7
10	{1}	1.000	0	0	0	0	0	0
11	{1,5}	0.931	0	0	0	0.931	0	0
12	{1,5,6}	0.950	0	0	0	0.892	0.784	0
13	{1,5,6,7}	0.982	0	0	0	0.963	0.889	0.453
20	{1,2}	0.933	0.933	0	0	0	0	0
21	{1,2,5}	0.859	0.859	0	0	0.747	0	0
22	{1,2,5,6}	0.963	0.963	0	0	0.794	0.514	0
23	{1,2,5,6,7}	0.967	0.967	0	0	0.933	0.600	0.200
30	{1,2,3}	0.914	0.914	0.914	0	0	0	0
31	{1,2,3,5}	0.895	0.895	0.895	0	0.465	0	0
32	{1,2,3,5,6}	0.917	0.917	0.917	0	0.632	0.158	0
40	{1,2,3,4}	0.844	0.844	0.844	0.844	0	0	0
41	{1,2,3,4,5}	0.950	0.950	0.950	0.950	0.400	0	0

Source: Authors, (2019).

We will discuss two methods of calculation the Shapley value Ø for 12-brood by hand. the first one is finished for every possible ordering of the grand coalition, while the second concerns the elimination of orders that are considered unnecessary before starting calculating the value. to grasp how this calculation method works, we are visiting denote a group of possible orders P as any possible orders of the chick once they're being fed by the parents: starting from the firstly fed chick, until every chick within the corresponding XY-brood data is being fed. as an example, order 1-5-6 within the 12- brood game means the core chick is being fed within the primary place, followed by the first and also the second marginals consecutively. When chick parents come to the nest bringing the foods for his or her chick, whichever chick that begs harder are fed first with usually the most important amount of food, and vice versa; chick that's being fed last will get just the rest. As food given from parents is perhaps the sole source of the chick' nutrition's, a minimum of until they're ready to y and appearance for an additional source of food, this food is extremely important for them to survive. Therefore, we may logically assume that the chick's order of being fed will affect their survival rate. Using this assumption, it's possible to calculate the Shapley value by taking the common of the chick' survival rate within the corresponding brood data to be interpreted because the amount of food the chicks have gotten from their parents which can help them to survive. a mean of 1.000 for a chick's survival rate might be translated as: the chick is getting 100% of food that it must survive. so as to calculate the Shapley value by hand using the interpretation above, we define the subsequent allocation procedure:

1. Consider the XY-brood game during either the good or the bad year's period under the restricted approach. Make a coalition table for the XY- brood. As an example, now we consider the good years 12-brood game, looking at Table 2 for the chick' average survival rate data during the good years, notice that we use the total sum of all chick' survival rate in 12-brood to fill in the value of the grand coalition, while the marginals average m av of the 11-brood and 12-brood are used to fill in the value of the coalition {5} and {6}, respectively. To fill in the value for coalition {1,5} and {1,6} we use the sum of the survival rate for core chick in 11-brood with the m av of 11-brood and 12-brood respectively. Finally, the sum of m_1and m_2 survival rate of the 12-brood is used to fill in the value for coalition {5,6}. Thus, we have a coalition table for the 12-brood game as follows:

Table 6: 12-brood game, good years.

S	{1}	{2}	{5}	{1,2}	{1,5}	{2,5}	{1,2,5}
$v(S)$	1.000	0.931	0.838	1.862	1.769	1.676	2.626

Source: Authors, (2019).

2. List every possible orders of the grand coalition that correspond to this XY-brood data.

3. In order to be able to fill within the 'right' value that each player will get in keeping with their possible ordering, we adapt the identical Shapley procedure. This way, we are going to divide the worth of the grand coalition 'fairly' by considering the orders and also the value that are 'claimed' by each coalition. As an example, from Table 5, we all know that player 1 in coalition {1} is 'claiming' a median of player 1 for its survival rate, while player 5 and 6 in coalition {5} and {6} and are 'claiming' a median of 0.931 and 0.838, respectively. If we take into consideration order 1-5-6 of the players within the grand coalition, we are going to firstly allocate 1.000 for player 1; precisely the same amount as what it claims. to come to a decision what proportion should player 5 gets, we glance at Table 6 and see that 1.862 is that the value of coalition {1,5}. Since we already give player 1 a worth of 1.000, the remaining value of 0.862 are going to be the quantity which is given to player 5. confine mind that the sum of each player's value has to be capable the worth of the grand coalition. Since a complete of 1.862 has already been given to player 1 and 5, player 6 will get the rest of the grand coalition value; which is 0.764. Doing the identical procedures to each possible orders, we are going to get a Shapley value calculation table as shown below. Note that notation Ø denotes the Shapley value of every player involved within the grand coalition.

Table 7: Shapely value of 12-brood game, good years.

Possible Orders P	Player			
	1	5	6	Total
1-5-6	1.000	0.862	0.764	
1-6-5	1.000	0.857	0.769	
5-1-6	0.931	0.931	0.764	
5-6-1	0.950	0.931	0.745	
6-1-5	0.931	0.857	0.838	
6-5-1	0.950	0.838	0.838	
Ø	0.960	0.879	0.786	2.626

Source: Authors, (2019).

4. Now we compare the values we got from observing the common survival rates of every chick within the corresponding XY-brood data, which we denote as Observ. to the Shapley values we got from calculation. so as to induce these observation values, we want to appear at Table 5 (restricted approach) and find the common survival rate of every chick within the corresponding XY-brood. as an example, the observation values for player 1, player 5, and player 6 in 12-brood game in line with Table 5 are 0.950, 0.892, and 0.784, respectively. For easier comparison, we are going to add these observation values into the Shapley value calculation table we made within the previous step, resulting this table below:

Table 8: Possible orders for 12-brood game, good years.

Possible Orders P	Player			
	1	5	6	Total
1-5-6	1.000	0.862	0.764	
1-6-5	1.000	0.857	0.769	
5-1-6	0.931	0.931	0.764	
5-6-1	0.950	0.931	0.745	
6-1-5	0.931	0.857	0.838	
6-5-1	0.950	0.838	0.838	
∅	0.960	0.879	0.786	2.626
Observ.	0.950	0.892	0.784	2.626

Source: Authors, (2019).

5. Since there's almost no case within the chick broods where the marginal chicks are being fed before the core chick, now we fancy to jump over the unfeasible orders from the Shapley value calculation table and consider only the cases when the core chick are being fed before the marginals. We now have a brand new Shapley value calculation table with a collection of feasible orders F rather than possible ones. Below is that the new Shapley value calculation table for the 12-brood game:

Table 9: Feasible orders for 12-brood game, good years.

Feasible Orders P	Player			
	1	5	6	Total
1-5-6	1.000	0.862	0.764	
1-6-5	1.000	0.857	0.769	
∅	1.000	0.8595	0.7665	2.626
Observ.	0.950	0.892	0.784	2.626

Source: Authors, (2019).

Again, we compare the Shapley value ∅ with the observation value to see if parental favoritism exists within the case of specific XY-brood data. Notice that within the case of 12-brood, the Shapley values ∅ for the core chick that we dawned on both cases are larger than the observation values. Thus there is no tendency of parents favoriting the core chick in line with this Shapley value solution in 12-brood. On the alternative hand, the Shapley value for the marginal is almost always larger within the observations rather than within the calculation; aside from the second marginal within the case of taking all possible orders P into the calculation. Thus we may say that within the great year's 12-brood data, there is no indication of chick parents playing favorites between the core and thus the marginal chick. Note that the identical way of calculations can also be applied for every brood game, especially the smaller ones (with but three or four players in one brood).

Here we provide another example within the great year's data using the above Shapley value calculation procedure to see if parental favoritism could exist even within the great years. Consider the 21-brood game under the restricted approach where there exist two core chick as player 1 and a pair of and one marginal as player 5.

Following the same Shapley value calculation procedure, ordering the chick in 21-brood into order 1-5-2 means that we firstly give allocation for coalition {1,3} (by giving allocation for chick 1 first from v({1,5}) and the rest for chick 5), then lastly give the rest of the grand coalition value v({1,2,5} for chick 2 after being reduced by v({1,5}). Table 10 below will list all possible orders additionally because the Shapley value for the 21-brood game mentioned earlier (see Table 3 for all the possible coalition values of this 21-brood game). In the end, we also compare the worth we got with our observation value for the 21-brood data (see the corresponding average of the survival rate for every chick involves within the 21-brood game from the A data).

Note that the sum of the observed average of all chick' survival rate within the corresponding game (see **d8 val** data in Table 2) is capable the worth of the grand coalition.

Table 10: Possible orders for 21-brood game, good years.

Possible Orders P	Player			
	1	2	5	Total
1-2-5	1.000	0.867	0.598	
1-5-2	1.000	0.603	0.862	
2-1-5	0.867	1.000	0.598	
2-5-1	0.603	1.000	0.862	
5-1-2	0.931	0.603	0.931	
5-2-1	0.603	0.931	0.931	
∅	0.834	0.834	0.797	2.465
Observ.	0.859	0.859	0.747	2.465

Source: Authors, (2019).

From Table 10, we see that the marginal chick gets a bit but its Shapley value solution, while the cores get a bit more within the observation. It implies that in keeping with the Shapley value solution concept and by considering every possible orders of feeding the chick within the good years 21-brood, we may say that the fogeys are quite 'favoriting' the core ones. Now we are going to jump over the unfeasible orders and consider only the feasible ones. Coalitions during which any marginal is fed before any core don't seem to be feasible. Thus, erasing orders 1-5-2, 2-1-5, 5-1-2, and 5-2-1 from our calculations will give us the table below:

Table 11: Feasible orders for 21-brood game, good years.

Feasible Orders P	Player			
	1	2	5	Total
1-2-5	1.000	0.867	0.598	
2-1-5	0.867	1.000	0.598	
∅	0.9335	0.9335	0.598	2.465
Observ.	0.859	0.859	0.747	2.465

Source: Authors, (2019).

We can see in Table 11 that if we remove the unfeasible coalition orders, the result's the opposite way around. Here the marginal gets way more in point of fact instead of what it speculated to get supported the Shapley value solution that we calculate.

Therefore, we'd like to test these two conditions on our Shapley value solutions: Whether the one claiming more will always get quite the one claiming less, and whether the one claiming more will always lose quite the one claiming less.

IV. RESULTS AND DISCUSSIONS

For the case of fine years 12-brood game, from Table 2 we will see that solely, players 1, 5, and 6 are claiming 1.000, 0.931, and 0.838, respectively. this suggests player 1 claims the foremost, while player 6 claims the smallest amount. When considering all possible orders P, the Shapley value solutions are 0.960, 0.879, and 0.786, respectively. Since in keeping with these solutions player 1 gets the foremost while player 6 gets the smallest amount, the primary property of the CG-solution is satisfied. However, once we consider the lose (i.e. the difference between the claim and also the reward) that each player has, player 1 loses 0.040, while players 5 and 6 equally lose 0.052. The loss of player 1 who claims the foremost, is of course smaller

than the loss of two other players who claim less. Thus, the second property is unfortunately not satisfied. Therefore, when considering all the possible orders into the calculation, this Shapley value solutions of the great years 12-brood game isn't a CG-solution. we will also easily check for the case of removing the unfeasible orders and should attain the identical conclusion. within the case of the great years 21-brood game, we also get the identical conclusions when considering only the feasible orders into the Shapley value calculation.However, we get a special result once we consider all the possible orders. Claiming 1.000, 1.000, and 0.931 respectively, player 1 and player 2 equally get 0.834, while player 5 gets 0.797 in their Shapley value solutions. Claiming the foremost, players 1 and a pair of lose 0.166, while player 5 loses 0.134. we are able to easily see that this point, the 2 properties are satisfied. Thus, we may say that the Shapley value solutions of the great years 21-brood game are a CG-solution once we consider all the possible feeding orders into the calculation.

V. CONCLUSIONS

This paper gives a mild and gentle intro to cooperative game theory in chick's brood food allocation since the modeling and calculation choices we made using the Shapley value solution concept have proved that parental favoritism does exist in most cases of the brood data. The restricted structure of the brood datasets also enables the Shapley value solution concept to give a reasonable fit within a reasonable time. The brood data we are taking, the results of the experiments have tendency that small increase on the food allocation for the marginals could increase the chick's probability of survive a lot more, while giving more food to the core chick who already has a high survival rate does not give a different output as the core already has a great chance of surviving. To summarize the results of the experiments, the better the game fits the solutions, the more we can trust the resulting system. Shapley value solution concept to tackle the chick brood food allocation problem. We also successfully translate the biological problem of chick's food allocation into a cooperative game approach using various techniques known in literature.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Md Obaidul Haque.
Methodology: Md Obaidul Haque, Anwar Hossen.
Investigation: Md Obaidul Haque, Sharmin Akter.
Discussion of results: Md Obaidul Haque, Anwar Hossen.
Writing – Original Draft: Md Obaidul Haque.
Writing – Review and Editing: Md Obaidul Haque.
Resources: Md Obaidul Haque, Sharmin Akter.
Supervision: Anwar Hossen.
Approval of the final text: Md Obaidul Haque, Anwar Hossen, Sharmin Akter.

VII. ACKNOWLEDGMENTS

Acknowledge the continuous support and comments received from Dr. Mohaamad Anwar Hossen, Associate Professor, Department of mathematics, Comilla University and Sharmin Akter, Department of mathematics, Comilla University. The study is supported by Department of mathematics, Comilla University.

VIII. REFERENCES

[1] Md Obaidul Haque,S. Akter and A. Hossen, "A cooperative game theory application in chicks brood food allocation by using shapely value method in bad years data", American Journal of Innovative Research and Applied Sciences, vol. 10, pp. 222-227, 2020.

[2] Shapley L.S. A Value for n-Person Games, "In Contributions to the Theory of Games", Annals of Mathematics Studies. vol. II, no. 28, pp. 307-317, H.W. Kuhn and A.W. Tucker, eds., Princeton University Press, 1953.

[3] Haake C-J., Kashiwada A., and Su F.E. "The Shapley value of phylogenetic trees", Working papers, Bielefeld University, Institute of Mathematical Economics, 2005.

[4] L.S. Shapley, A Value for n-Person Games, "In Contributions to the Theory of Games", Annals of Mathematics Studies, vol. II, no. 28, pp. 307-317, H.W. Kuhn and A.W. Tucker, eds., Princeton University Press, 1953.

[5] D. Monderer, D. Samet, and L.S. Shapley, "Weighted Values and the Core", International Journal of Game Theory, vol. 21, pp. 27-39, 1992.

[6] A. Kar, M. Mitra, and S. Mutuswami, "On the coincidence of the prenucleolus and the Shapley value", Mathematical Social Sciences, Elsevier, vol. 57, no. 1, pp. 16-25, January 2009.

[7] S. Forbes, "Sibling Symbiosis in Nestling Birds", Auk 124, pp. 1-10, 2007.

[8] H. Bengtsson and O. Rydén, "Parental feeding rate in relation to begging behavior in asynchronously hatched broods of Great Tit Parus major", Behavioral Ecology and Sociobiology, vol. 12, pp. 243–251, 1983.

[9] B. Glassey and S. Forbes, "Muting individual nestlings reduces parental foraging for the brood", Animal Behaviour 63, pp. 779–786, 2002.

[10] L. H. Brown, V. Gargett, and P. Steyn, "Breeding success in some African eagles related to theories about sibling aggression and its effects". Vol: 48, pp. 65–71, 1977.

[11] Klaus Nehring and Clemens Puppe. A theory of diversity. Econometrica, vol. 70(3), pp. 1155–1198, 2002.

[12] E. H. Dunn, "The relationship between brood size and age of effective homeothermy in nestling House Wrens", Wilson Bulletin, vol. 88, pp. 478–482, 1976.

[13] E. H. Dunn, "On the variability in energy allocation of nestling birds", Auk 97, pp. 19–27, 1980.

[14] R. M. Evans, M. O. Wiebe, S. C. Lee, and S. C. Bugden, "Embryonic and parental preferences for incubation temperature in Herring Gulls: Implications for parent- offspring conflict", Behavioral Ecology and Sociobiology, vol. 36, pp. 17–23, 1995.

[15] S. Forbes and B. Glassey, "Asymmetric sibling rivalry and nestling growth in Red- winged Blackbirds (Agelaius phoeniceus)", Behavioral Ecology and Sociobiology, vol. 48, pp. 413–417, 2000.

[16] S. Forbes, S. Thornton, B. Glassey, M. Forbes, and N. J. Buckley, "Why parent birds play favourites", Nature 390, pp. 351–352, 1997.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.6 n.25, p. 27-31, Sep/Oct, 2020

DOI: https://doi.org/10.5935/jetia.v6i25.696

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

TESTING AND VALIDATING A TOOL TO MEASURE PRODUCT INVOLVEMENT FROM ITS ANTECEDENTS

Ashutosh Anil Sandhe*[1]

[1] Department of Commerce & Management, TeamLease Skills University, Tarsali (Vadodara)-Gujarat, India.

[1] http://orcid.org/0000-0003-3702-8325

Email: *ashutoshsandhe@yahoo.com

ARTICLE INFO

Article History
Received: September 11th, 2020
Accepted: October 19th, 2020
Published: October 30th, 2020

Keywords:
Involvement,
Cognition,
Intention,
Awareness,
Credence.

ABSTRACT

Lot of research has been carried out lately with respect to the concept of involvement. This has become an area of interest from multiple strategic angles. Since, involvement is a concept which is related with the psychology and buying behavior of consumers, it is relatively cumbersome to measure. This paper has attempted to examine product involvement for real estate properties in the state of Gujarat, India. Data was collected in the form of structured questionnaires from respondents across the three most populated urban cities i.e. Ahmedabad, Surat and Vadodara. The tool had thirty questions asking respondents to provide their opinion on statements on a seven point likert scale. A total sample of 600 was considered appropriate with equal samples from the three cities. Data was collected on the basis of stratified random sampling where occupation was considered as the strata. The statements were formed considering three dimensions of involvement i.e. cognition, affection and behavioural. The data was found fit for factor analysis based on KMO and Bartlett's test. Factor analysis (Principal Axis Factoring) was carried out to test and validate factors that had an impact in determining product involvement. The results showed that from the three dimensions and thirty questions, five antecedents were found which were named as Affect (Af), Awareness (Aw), Intent (In), Credence (Cr) and Action (Ac). The tool developed was named as Involvement Antecedent Framework (IAF). Lastly, correlation analysis confirmed high levels of correlation between the antecedents of involvement.

I. INTRODUCTION

A lot of research has been carried out and will be carried out in future in the field of consumer behavior studies. The behavior of consumers is a complex phenomenon and keeps changing constantly with changes in the environment. For instance, what was assumed to be brand loyalty, was found to be habitual buying behavior (Kotler, 2015) [1]. With increasing choice and competition, the consumer has become more powerful and accordingly, marketers are trying to study them as closely as possible. One such area of study which is gaining prominence is involvement. Though originally the concept was found in psychology, soon researchers found the concept useful in the study of consumer psychology too.

II. PRODUCT INVOLVEMENT

The concept of involvement has been introduced in the field of marketing and especially consumer behaviour in the 1940s by Sheriff and others where they defined ego involvement [2]. Numerous definitions of involvement have been put forward by researchers over time. Freedman (1964) defined involvement as a general level of interest or concern about an issue [3]. The concept of involvement is to be applied to the person or consumer, in this case, and not the object. Therefore, it may be seen that two different persons may showcase different involvement levels for the same object [4]. Consumer involvement is important in the study of consumer behaviour since it can potentially become an important mediator (Andrew Mitchell, 1979) [5]. In the last few years, the concept of involvement has become considerable since it has

opened up a new dimension of consumer psychology. This concept has the potential to become a base of market segmentation [6]. It has been used to study the buying decision process for different offerings. It has been found that where involvement is low, products are purchased without detailed study of the product or offering. Evaluation, if any, is done on a superficial level (Lastovicka, 1979) [7]. Some of the most notable researches in this field can be dedicated to Zaichkowsky and Laurent and Kapferer who attempted to measure involvement with respect to advertisements and products. Laurent and Kapferer identified multiple kinds of involvement depending on product importance, risk, brand, pleasure value and so on [8]. The researchers developed a tool to measure involvement based on fourteen product categories. Another major contributor was Zaichkowsky who developed a tool called the personal involvement inventory (PII) [9] which was based on a bi-polar scale and had around thirty items. Later this tool was revised to twenty item scale which was named as the revised personal involvement inventory (RPII). The construct was applicable to product involvement, advertising involvement and also purchasing involvement [10]. Two scales of involvement have appeared recently in major marketing/consumer behavior journals. Of these, Laurent and Kapferer's (1985) scales assume multi-dimensionality in involvement; and Zaichkowsky's (1985) scale, while driven by a unidimensional view of involvement, is not unified [11]. Another dimension to measurement of involvement was put forward by Carmen Garcia et al (1996) when they designed a tool called the CIQ (Consequences of Involvement Questionnaire). Instead of asking the respondent his/her level of involvement with a product, they identified antecedents to involvement and through these antecedents, involvement levels were empirically calculated [12]. The concept of product involvement is ever evolving and still lot of insights are yet to be realized. Product involvement is closely related to perceived risk (Dholakia, 1997) [13]. Because of this, it has been observed that based on the levels of involvement, other factors like payment mechanisms and even shopping situations are likely to be different for low involvement products when compared to high involvement products (Ming-Chuan Pan, 2007) [14]. The concept of involvement can be used as a marketing tool to segment markets into low, moderate and high involvement customers and based on that each group can be targeted separately (Michaelidou et al, 2008) [15]. It is found that perceived risk fully mediates the effect of the importance dimension of product involvement on information search but not of the hedonic dimension. The effect of hedonic involvement on information search is direct [16]. The Elaboration Likelihood Model by Petty and Cacioppo suggested that in case of higher involvement with message, a central route of persuasion is adopted while where involvement is low, peripheral route of persuasion is adopted [17].

III. RESEARCH METHODOLOGY

The purpose of this study was to test and validate antecedents of product involvement for real estate properties in the state of Gujarat, India and present a standard tool to measure involvement in general. In this descriptive research, a sample of 600 respondents from across three most populated cities of Gujarat i.e. Ahmedabad, Surat and Vadodara was considered sufficient. Data was collected in the form of a structured questionnaire in which respondents were asked to provide their agreement to a set of thirty statements related to product involvement. Initially, three dimensions were identified in order to measure involvement. Cognitive dimension measured the information processing methods, knowledge and perception about the product. The affective or emotional dimension measured the feelings towards the product or brands and finally the behavioural dimension which measured the purchase intention and purchase purpose of the product. The statements based on these three dimensions were categorized into factors using principal axis factoring method of factor analysis. Before undertaking data analysis, reliability of the data was measured through Cronbach Alpha which was 0.837 suggesting high measure of reliability.

IV. RESULTS AND DISCUSSIONS

A structured questionnaire comprising thirty items was constructed. The purpose was to identify valid factors as antecedents of consumer involvement. The tool had statements which were to be rated on a seven point likert scale where 7 was "very strongly agree" and 1 as "very strongly disagree". By applying principal axis factoring method, valid factors were identified and analysed. One pre-determinant for conducting factor analysis is the correlation between the items. In absence of significant correlation, factor analysis cannot be conducted.

Table 1: KMO and Bartlett Test.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.934
Bartlett's Test of Sphericity	Approx. Chi-Square	7933.475
	df	435
	Sig.	0.000

Source: Author, (2020).

For the present study, Bartlett's test of Sphericity is found highly significant with 435 degrees of freedom and a Chi square value of 7933.475 (p=0.000). A significant value shows high correlations among variables tested. Along with Bartlett's test of Sphericity, another important test for factor analysis is the Kaiser-Meyer-Olkin (KMO) measure of sample adequacy which indicates whether sample is sufficient for the factor analysis to be conducted. In the present study, a KMO value of 0.934 is obtained which indicates that factor analysis can be conducted. Thus, the basic parameters for conducting a valid factor analysis are found to be favourable.

Table 2: Factor Analysis for Testing and Validating Antecedents of Involvement for Real Estate Properties.

Factor	Initial Eigenvalues		
	Total	**% of Variance**	**Cumulative %**
1	10.664	35.547	35.547
2	1.812	6.041	41.587
3	1.458	4.862	46.449
4	1.314	4.380	50.829
5	1.176	3.920	54.749
	Extraction Sums of Squared Loadings		
	Total	**% of Variance**	**Cumulative %**
1	10.159	33.864	33.864
2	1.346	4.486	38.351
3	0.972	3.239	41.590
4	0.781	2.603	44.193
5	0.668	2.226	46.419
	Rotation Sums of Squared Loadings		
	Total	**% of Variance**	**Cumulative %**
1	3.312	11.041	11.041
2	2.783	9.275	20.316
3	2.637	8.789	29.105
4	2.633	8.776	37.881
5	1.641	5.471	43.352
Extraction Method: Principal Axis Factoring.			

Source: Author, (2020).

From the table, a total of five factors were found to be having Eigen value more than 1. Out of the factors, first factor contributed the most with 35.547 percent of the total variance. The second factor contributed 6.041 percent to the total variance, while the third factor contributed 4.862 percent. The fifth factor had the least contribution of 3.920 percent to the total variance. The total variance of the five factors was 54.7490 percent which is an acceptable level for studies related to social science and consumer behaviour (Yong & Pearce, 2013) [18].

After validating five factors affecting consumer involvement, the rotated component factor matrix loading was examined in order to validate each item and allocate all the items to each of the five factors. While doing so, all items having factor loading less than 0.4 were ignored as having negligible or no influence. A total of 26 items were found valid from the initial 30 for which data was collected. Following Table 3 shows the factor loadings of each item of the tool.

Table 3: Rotated Component Matrix (Factors Identified).

Item	Factor				
	FI	FII	FIII	FIV	FV
This product is important for me				0.484	
I will prefer it if it fulfills my needs				0.534	
I love this product	0.479		0.406		
I very much have pleasure and enjoyment using this product	0.517				
I don't have any problem in spending money on this product			0.485		
This product makes me feel good	0.553				
This product has an important place in my life	0.427				
I have an emotional attachment with this product	0.465				
I have a strong interest in this product		0.643			
I take interest in collecting information about this product		0.642			
I compare various alternatives available in this product category		0.572			
I believe in going through a meticulous process of information collection for this product				0.645	
I like to spend time learning more about this product	0.508				
I love to get experts opinions and evaluations on this product	0.420				
I pay more attention to advertisements about this product			0.655		
I am able to evaluate the differences in various brands of this product			0.464		
I try to be up to date with latest information about this product			0.423		
I try to get maximum information about this product from various sources		0.658			
I am eager to buy this product in the right conditions					0.678
I feel that if I purchase this product, it will enhance my social standing	0.511				
Most of the people wanting to buy this product take a detailed process of buying		0.584			
I enjoy talking with knowledgeable people on this product				0.661	
I like to share ideas about this product with my friends	0.595				
Lack of this product makes me feel deprived	0.610				
I feel that this product is important for everyone			0.432		
I like/ would like to have this product					0.691
Eigen Value	**10.664**	**1.812**	**1.458**	**1.314**	**1.176**
% of Variance	**35.547**	**6.041**	**4.862**	**4.38**	**3.92**
Cumulative %	**35.547**	**41.587**	**46.449**	**50.829**	**54.749**
Extraction Method: Principal Axis Factoring. Rotation Method: Varimax with Kaiser Normalization.					

Source: Author, (2020).

Based on the factor analysis results, the three dimensions initially identified were scattered into five factors and a twenty six item tool to measure consumer involvement was finalized which was named as the Involvement Antecedent Framework (IAF). The five factors were named as Affect (Af), Awareness (Aw), Intent (In), Credence (Cr) and Action (Ac). Table 3b shows the complete valid tool having five distinct antecedents that affect consumer involvement towards real estate properties. The data clearly shows that the tool measures high on reliability for all the factors individually. The Cronbach alpha ranges between 0.860 for Affect (Af) and 0.705 for the factor Action (Ac). Since all the Cronbach values are above 0.60 which is considered as a standard for accepting reliability of the data, it can be said that the results obtained from the analysis were reliable for interpretation. The overall mean for the factors was found to be 5.733 on a seven point scale which indicated that product involvement for real estate properties in selected cities of Gujarat was high.

Table 3b: Rotated Component Matrix (Factors Named).

Factor	Item	Factor Loading	Cronbach	Mean	S.D.
Affect (Af)	I love this product	0.479	0.860	5.567	0.816
	I very much have pleasure and enjoyment using this product	0.517			
	This product makes me feel good	0.553			
	This product has an important place in my life	0.427			
	I have an emotional attachment with this product	0.465			
	I like to spend time learning more about this product	0.508			
	I love to get experts opinions and evaluations on this product	0.420			
	I feel that if I purchase this product, it will enhance my social standing	0.511			
	I like to share ideas about this product with my friends	0.595			
	Lack of this product makes me feel deprived	0.610			
Awareness (Aw)	I have a strong interest in this product	0.643	0.812	5.96	0.837
	I take interest in collecting information about this product	0.642			

	I compare various alternatives available in this product category	0.572			
	I try to get maxmum information about this product from various sources	0.658			
	Most of the people wanting to buy this product take a detailed process of buying	0.584			
Intent (In)	I love this product	0.406	0.796	5.468	0.927
	I don't have any problem in spending money on this product	0.485			
	I pay more attention to advertisements about this product	0.655			
	I am able to evaluate the differences in various brands of this product	0.464			
	I try to be up to date with latest information about this product	0.423			
	I feel that this product is important for everyone	0.432			
Credence (Cr)	This product is important for me	0.484	0.781	5.707	0.856
	I will prefer it if it fulfills my needs	0.534			
	I believe in going through a meticulous process of information collection for this product	0.645			
	I enjoy talking with knowledgeable people on this product	0.661			
Action (Ac)	I am eager to buy this product in the right conditions	0.678	0.705	5.963	0.846
	I like/ would like to have this product	0.691			

Source: Author, (2020).

The purpose of this research is not just to validate a tool to measure consumer involvement, but also to examine the overall consumer involvement for real estate properties in Gujarat. Through Bartlett's test, it is observed that there is high level of correlation between factors affecting consumer involvement. Correlation analysis confirmed this fact.

Table 4: Consumer Involvement for Real Estate in Gujarat.

		Affect (Af)	Awareness (Aw)	Intent (In)	Credence (Cr)	Action (Ac)
Affect (Af)	Pearson Correlation	1				
	Sig. (2-tailed)					
	N	600				
Awareness (Aw)	Pearson Correlation	.535**	1			
	Sig. (2-tailed)	0.000				
	N	600	600			
Intent (In)	Pearson Correlation	.725**	.550**	1		
	Sig. (2-tailed)	0.000	0.000			
	N	600	600	600		
Credence (Cr)	Pearson Correlation	.632**	.398**	.634**	1	
	Sig. (2-tailed)	0.000	0.000	0.000		
	N	600	600	600	600	
Action (Ac)	Pearson Correlation	.431**	.288**	.427**	.424**	1
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	
	N	600	600	600	600	600
Two tailed correlation is significant at 99%						

Source: Author, (2020).

As seen in Table. 4, there was significant positive correlation between all the antecedents of consumer involvement in the range of 0.725 and 0.288. The highest correlation was between Affect (Af) and Intent (In) while that between Awareness (Aw) and Action (Ac) was found to be the least. The mean for overall involvement for real estate properties in Gujarat was found to be 5.733 indicating high levels.

V. CONCLUSIONS

The data collected is found fit for further analysis in terms of preliminary tests. It is found that out of initially designed thirty items 26 items had factor loadings above 0.4 and based on that the tool is considered robust for further analysis. The three dimensions initially considered are split into five factors or antecedents which are named as Affect (Af), Awareness (Aw), Intent (In), Credence (Cr) and Action (Ac). The analysis of each antecedents provide valid and reliable results in terms of Cronbach alpha. The mean values of each antecedent is on the higher side of the seven point scale indicating that in this study the product involvement for real estate properties in Gujarat showes higher involvement levels. As shown by the Bartlett's test, there was high positive correlation between all the antecedents of product involvement as shown in Table 4. The highest correlation is found between Affect (Af) and Intent (In) (r=0.725), while the least correlation is found between Awareness (Aw) and Action Ac) (r=0.288). From the five factors maximum items (10) are related to the factor Affect (Af), the factor Awareness has 5 items whereas Intent (In) has 6 items. The least items are associated with the last factor Action (Ac) which is 2. The remaining factor Credence (Cr) has 4 items.

VI. LIMITATIONS

The research reveals some important insights into the antecedents of product involvement. This framework is applied to real estate properties. The same model can be tested with other products especially low involvement products or even services. A similar construct can be applied to a greater population to further validate the results. Based on this framework, a common construct to measure involvement levels and examine the factors associated with it can be propounded.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Ashutosh Anil Sandhe.
Methodology: Ashutosh Anil Sandhe.
Investigation: Ashutosh Anil Sandhe.
Discussion of results: Ashutosh Anil Sandhe.
Writing – Original Draft: Ashutosh Anil Sandhe.
Writing – Review and Editing: Ashutosh Anil Sandhe.
Resources: Ashutosh Anil Sandhe.
Supervision: Ashutosh Anil Sandhe.
Approval of the final text: Ashutosh Anil Sandhe.

VIII. REFERENCES

[1] Kotler, P., Keller, K. L., Koshy, A., Jha, M., & Pearson Education. (2013). Marketing Management: A South Asian perspective. New Delhi: Pearson Education.

[2] Sheriff M., Cantril H., (1947), The Psychology of Ego-Involvement, John Wiley and Sons, New York.

[3] Freedman, J. L. (1964), "Involvement, Discrepency and Change", Journal of Abnormal and Social Psyshology, 69, pp. 290-295.

[4] Hupfer N., D M Gardner, (1971), Differential Involvement with Products and Issues: An Exploratory Study, in "SV – Proceedings of the Second Annual Conference of the Association for Consumer Research, eds. David M. Gardner, College Park, MD: Association for Consumer Research, pp.262-270.

[5] Andrew A. Mitchell (1979), "Involvement: A Potentially Important Mediator of Consumer Behavior", in Advances in Consumer Research Vol.6, eds. William L. Wilkie, Ann Abor: Association for Consumer Research, pp- 191-196.

[6] Wim Verbeke, Isabelle Vackier, (2004), Profile and Effects of Consumer Involvement In Fresh Meat, Meat Science, Vol. 67(1), pp -159-168.

[7] John L. Lastovicka (1979), "questioning the concept of involvement defined product classes", in Advances in Consumer Research Vol.6, eds. William L. Wilkie, Ann Abor: Association for Consumer Research, pp- 174-179.

[8] Gilles Laurent and Jean-Noel Kapferer, (1985), Measuring Consumer Involvement Profiles, Journal of Marketing Research.

[9] Zaichkowsky, J. L. (1985). Measuring the involvement construct. Journal of Consumer Research, 12 (December), 341-352.

[10] Judith Lynne Zaichkowsky, (1994), The Personal Involvement Inventory: Reduction, Revision, and Application to Advertising, Journal of Advertising, Volume XXIII, Number 4, 59-70.

[11] Banwari Mittal (1989),"A Theoretical Analysis of Two Recent Measures of Involvement", in NA - Advances in Consumer Research Volume 16, eds. Thomas K. Srull, Provo, UT: Association for Consumer Research, Pages: 697-702.

[12] Carmen García, Julio Olea, Vicente Ponsoda y Derek Scott (1996), Measuring Involvement from Its Consequences, Psicothema, 1996. Vol. 8, No. 2, pp. 337-349.

[13] Utpal M. Dholakia (1997),"An Investigation of the Relationship Between Perceived Risk and Product Involvement", in NA - Advances in Consumer Research Vol.24, Eds. Merrie Brucks and Deborah J. Mac Innis, Provo, UT: Association for Consumer Research, pp. 159-167.

[14] Ming-Chuan Pan, The Effects of Payment Mechanism and Shopping Situation on Purchasing Intention - the Moderating Effect of Product Involvement, Proceedings of the 13th Asia Pacific Management Conference, Melbourne, Australia, 2007, pp.1-10.

[15] Michaelidou, Nina and Dibb, Sally, (2008), Consumer involvement: a new perspective. Marketing Review, 8(1), pp. 83–99.

[16] Chaudhuri, A., (2000), A Macro Analysis of the Relationship of Product Involvement and Information Search: The Role of Risk, Journal of Marketing Theory and Practice, Vol. 8, No. 1 (Winter, 2000), pp. 1-15.

[17] Petty R.E., Cacioppo J.T. (1986) The Elaboration Likelihood Model of Persuasion. In: Communication and Persuasion. Springer Series in Social Psychology. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-4964-1_1.

[18] An Gie Yong, Sean Pearce, (2013), A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis, Tutorials in Quantitative Methods for Psychology, Vol. 9(2), p. 79-94.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.6 n.25, p. 32-36, Sep/Oct, 2020

DOI: https://doi.org/10.5935/jetia.v6i25.700

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

INSULATING PROTECTIVE COVER FOR LIVE WORKING: MAIN CAUSES OF REJECTION UNDER LABORATORY CONDITIONS

Ramón González Guevara[1] and Boris Alba Valle*[2]

[1, 2] Electroenergetic Research and Testing Center (CIPEL), Faculty of Electrical Engineering, Technological University of Havana (CUJAE), Havana, Cuba.

[1] http://orcid.org/0000-0002-6439-2408, [2] http://orcid.org/0000-0002-5441-3877

Email: ramon@electrica.cujae.edu.cu, *borisav@electrica.cujae.edu.cu

ARTICLE INFO

Article History
Received: September 17th, 2020
Accepted: October 20th, 2020
Published: October 30th, 2020

Keywords:
Insulating cover,
Electrical protective equipment,
Live working,
Aerial electrical networks,
Electric individual tests.

ABSTRACT

Insulating covers are collective protective equipment widely used in the maintenance of energized aerial electrical networks. The insulating covers can be divided in rigid or flexible. Using them along with other insulating means they allow carrying out repairs without interrupting the electric service. The present work relates essential aspects established in the international normative to achieve the adequate performance of rigid and flexible insulating protectors. It describes the experiences obtained in the individual electrical laboratory tests carried out in the CIPEL High Tension Laboratory in a measurement period of five years. As a result, the main insulation rigid and flexible covers used in Cuba and causes of rejection are identified at the laboratory level and measures are proposed that, in the authors' opinion, can improve the good condition of the protective element and increase its useful life, which represents an increase in operator safety and an improvement in the quality of the executed works.

I. INTRODUCTION

The generalization of the work technique with energized power lines has been possible thanks primarily to the constant improvement of the operator's safety conditions when using materials and equipment with better characteristics and insulating qualities and also for the use of new technologies.

In recent times there has been an increase in the execution of energized power lines in Cuba. This is due to the benefits offered by this technique in terms of economic savings and positive social impact. These works are executed using individual and shared insulating protective equipment, including the rigid and flexible conductor covers which are the subject of this work [1].

As an importer of these protective elements the experimental work of high voltage laboratories is oriented to the realization of individual tests. These tests allow to comply with the annual quality certification of the equipment and "live" work tools guaranteeing the security of the energized work brigades of the UNE.

Insulating cover, including new ones, can't be used if it has been longer than twelve months after its last test. And when behaviours that may indicate loss of the insulating qualities of the cover are detected, it has to be sent immediately to a certified electrical laboratory [2-3]. This practice and work dynamics have closely linked the specialists of the CIPEL High Voltage Laboratory and the technicians of the different companies responsible for enforcing the annual checks established by the UNE.

The analysis and interpretation of the results obtained in the experimental work carried out under laboratory conditions for three years will allow identifying the main rejection criteria detected in each type of cover.

II. GENERAL CHARACTERISTICS AND TECHNICAL REQUIREMENTS OF RIGID AND FLEXIBLE CONDUCTOR COVERS

For a better understanding of the work done below, the definitions of rigid and flexible cover respectively [2-3] are presented:

Rigid covers: electrically insulated enclosures designed to be temporarily installed on various types of irregularly shaped equipment in order to protect personnel and equipment from accidental electrical contacts.

Flexible conductor covers: electro-insulating flexible tubes made of different materials such as elastomers, plastics or a combination of both. They present a longitudinal incision and are installed temporarily in electrical conductors to avoid accidental contacts when performing works on energized power lines.

The insulating covers can be classified according to their class depending on the electrical limits they support by design and according to their use depending on the shape of the element to be protected.

II.1 CLASS

Different manufacture insulating covers can be found internationally. In the CIPEL High Voltage Laboratory all certified insulating covers meet the criteria of the ASTM standards [2-3]. Rigid and flexible conductor covers are divided in classes according to their electrical limits according to the criteria shown in Table 1.

Table 1: Classes of rigid and flexible conductor covers according to their electrical limits.

Class	Rigid covers	Flexible conductor covers
	Phase to phase voltage (kV rms) [1]	Phase to phase voltage (kV rms) [2]
0	-	1
1	-	7,5
2	14,6	17
3	26,4	26,5
4	36,6	36
5	48,3	-
6	72,5	-

Source: [2-3].

In Table 1 it is observed that for both the rigid and the flexible conductor cover their operating voltage for the common classes does not considerably vary. However, there are differences in the classes that are manufactured.

Due to the predominance of electric networks with nominal voltage from line to line of 15 and 33 kV in Cuba, the most commonly used insulating covers are class 2, 3 and 4.

II.2 UTILIZATION

There is a wide diversity of insulating covers depending on the energized elements to be isolated from the operators. The most used rigid insulation covers in Cuba are shown in Figure 1.

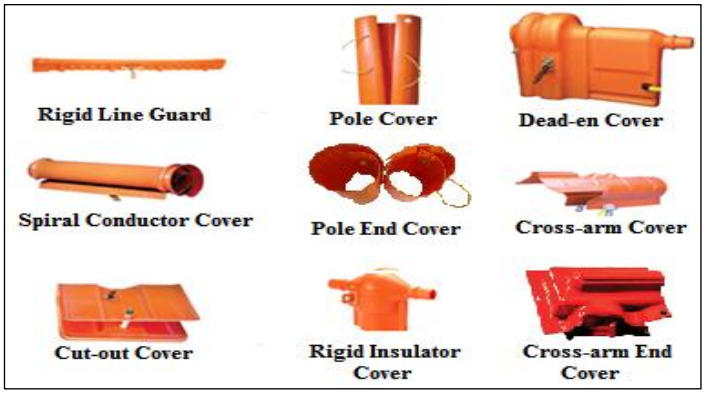

Figure 1: Rigid insulation covers most used in Cuba.
Source: [4-5].

The flexible conductor covers are used according to the styles shown in Figure 2.

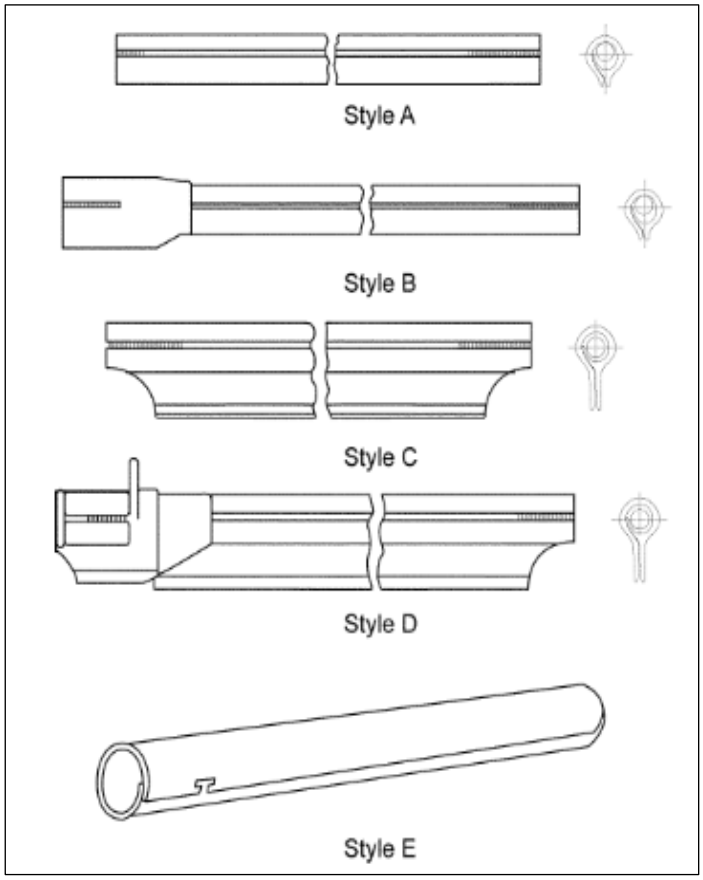

Figure 2: Different styles of flexible conductor covers.
Source: [3].

From Figure 2, the type A cover (short-lip) is the most used in our country for energized work [4-5]. Class 2 is the only one used for this type of cover. Other important aspects to keep in mind to maintain the quality of the sample in the processes of laboratory tests are: cleaning, drying, storage and transportation.

II.3 CLEANING AND DRYING

Each protective element must be inspected and cleaned in the field or laboratory conditions before being used. To fulfil this task, the following aspects must be taken into consideration [2-3]:

The cleaning of both types of insulating covers must be done with soap or light non-bleaching detergent or with a cleaner recommended by the manufacturer to prevent degrading of the insulating qualities of the cover [2-3].

Drying should be done by suspending each element in a way that facilitates drainage and air circulation. It is proposed that in cases where an automatic dryer is used, it should not have ultraviolet lamps, generate ozone or present interior surfaces that cut, wear or puncture the element [2-3].

II.4 STORAGE AND TRANSPORTATION CONDITIONS

Both types of covers must be stored and transported with extreme care avoiding mechanical stresses that could damage them. For storage, they should be placed in containers, boxes, or appropriate bags in cold, dry and dark places. It is important that the areas near the storage areas are free of ozone, chemical agents, oils, vapours, electrical discharges and heat sources [2-3].

III. INDIVIDUAL LABORATORY TESTS

In the 2014-2019 periods up to a total of 1026 rigid insulating cover and 452 flexible conductor cover belonging to different companies in the country were tested in the CIPEL High Voltage Laboratory [6-7]. The insulating covers subjected to tests are manufactured by CHANCE, SALISBURY, KEARNEY and RITZ (ASTM standards) [8-10]. All of them have different types according to their use and belong to classes 2, 3 and 4. A summary is shown in Table 2 of the technical data of the insulating covers subjected to tests.

Table 2: Technical data of the insulating covers subjected to individual electrical tests.

Technical dates	Rigid covers		Flexible conductor covers
Manufactures	CHANCE, SALISBURY, KEARNEY AND RITZ		SALISBURY
Utilization	Rigid cover to: conductor, insulator, dead-en, cut-out, cross-arm, cross-arm top, pole and pole top		Short-lip flexible cover
Class	3	4	2
Operating voltage between phases (kV rms)	26,4	36,6	17

Source: Authors, (2020).

The individual electrical test of both rigid and flexible conductor insulators consists of two stages: visual inspection and dielectric test.

III.1 VISUAL INSPECTION

Visual inspection of the material should be performed to check for presence of: cracks, holes, foreign matter, cuts or abrasions. Only small depressions or bulges are accepted as long as they do not affect its operation. Cases that present some of these problems are immediately rejected.

Once the visual inspection has been successfully carried out, the insulating cover is subjected to a dielectric test to verify the state of the insulation of the material.

III.2 DIELECTRIC TEST

The test procedure consisted in the execution of the following steps:

1. An inner (potential) electrode is selected depending on the type or style of the cover as established in the corresponding regulations [2-3].
2. As an external electrode (grounding), a wet tissue blanket was used.
3. The distance between electrodes is established according to the corresponding regulations [2-3].
4. The test voltage specified in Tables 3 and 4 is applied according to the class of cover. The electric voltage is increased at a rate of 1kV/s from a value no greater than its 50%. Once the specified value is reached, the time established in Tables 3 and 4 is maintained.
5. If disconnection of the source does not occur, the voltage is reduced to zero maintaining the same rise rate.

The environmental variables during the trials were maintained within the following ranges:

Temperature: between 18 ºC and 28 ºC.

Relative humidity: between 45% and 75%.

Table 3: Voltages and test times specified for rigid covers.

Class	Test voltage (kV effective)	Duration (min)
3	24	1
4	32	1

Source: [2].

Table 4: Voltages and specified test times for flexible conductor covers.

Class	Test voltage (kV effective)	Duration (min)
2	20	1

Source: [3].

The rejection criterion established for the evaluation of the dielectric test performed to the material are:

Cover fault before the test voltage reaches its maximum value or during its application time.

- A significant increase in temperature.
- A perforation or discharge occurs.

If any of these are observed: cracks, abrasions, burns, perforations or symptoms that indicate degradation of the material once the test is finished.

If any of these problems occur, the covers are immediately rejected.

The generally used test schemes are shown in Figures 3 and 4.

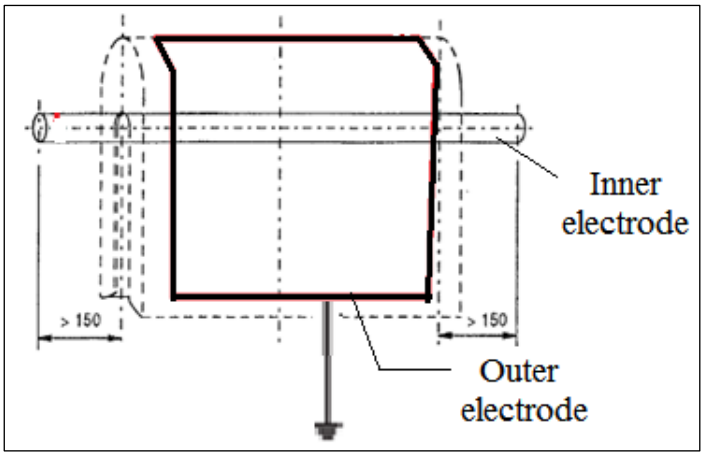

Figure 3: General test scheme for rigid cover certification.
Source: [2].

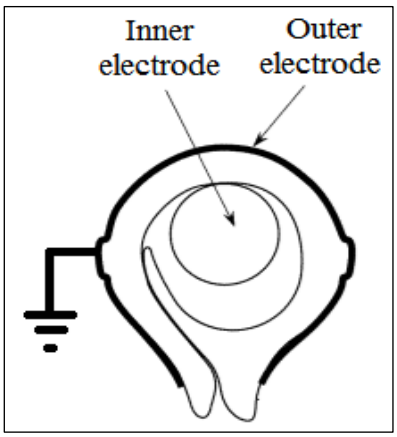

Figure 4: General test scheme for flexible conductor cover certification.
Source: [3].

IV. RESULTS AND DISCUSSIONS

Out of the 1026 rigid covers subjected to individual tests during the analysed period, 89 were rejected, representing a 8.7 % of the total. In the case of flexible covers, out of a total of 452 only 15 were rejected, representing a 3.3 % of the total. These percentages could be interpreted as general tendency of the covers to be resistant to mechanical and electrical stresses provoked by handling in field conditions.

Table 5 shows a synthesis of the results obtained at laboratory level in order to analyse which were the predominant causes that led to rejection in both cover types [7].

Table 5: Causes of rejection of insulating covers at the laboratory level.

Causes of rejection	Quantity of rigid insulating covers rejections	Quantity of flexible insulating conductor covers rejections
Cracks	71	2
Electrical perforation	18	13
Total	89	15

Source: Authors, (2020).

Next, an analysis is made of the percentage of the causes of rejection (Figure 5).

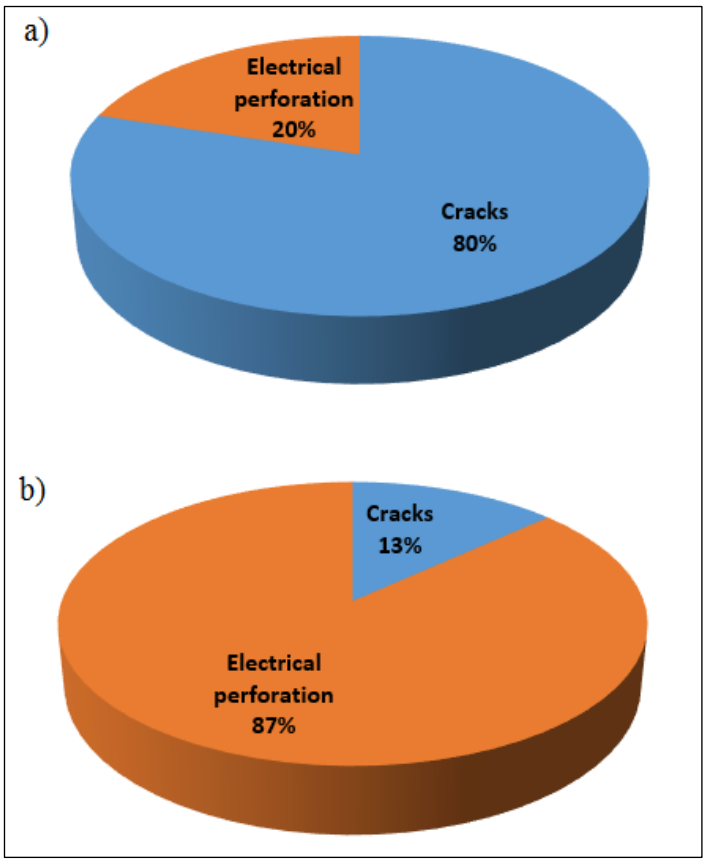

Figure 5: Cause of rejection percentages: a) rigid covers, b) flexible conductor covers.
Source: Authors, (2020).

Interpreting the behaviour of the results obtained in Figure 5 approximately 80% of the rigid cover's rejections are due to the presence of cracks and only 20% are due to the electrical perforation of the insulating material. However, in the case of flexible conductor covers, it is evident that 87% of the rejections are due to electrical perforation and 13% to cracking.

As a summary, it can be observed that the main cause of rejection of rigid insulating covers are the cracks present in the material while the main cause of rejection of flexible conductor cover is the electrical perforation of the material during the dielectric test.

The authors, based on the experience gathered working with specialists from different entities that use this type of protection, consider that these problems can be attributed mainly to deficient storage conditions, transportation and utilization conditions of the covers.

The problems related to storage, use and transportation of the covers are mainly due to lack of care owing to objective factors such as lack of material resources. Also, subjective factors such as the lack of technical culture prevent internalizing the extraordinary importance of the care that should be given to the protection means.

V. CONCLUSIONS

The main cause of rejection in the individual tests carried out in the CIPEL High Voltage Laboratory is the presence of cracks in the insulating material of the rigid covers and the electrical perforation in the flexible conductor covers. This reveals evidence of the existence of inadequate storage, transportation and handling conditions, therefor individual electrical tests have to be executed respecting the established periodicity. Also the training of the technical personnel ought to be continued and intensified through preparation courses to increase their technical culture. Furthermore, the implementation of measures by the UNE should be oriented to minimize the problems detected and increase the operator's safety and work quality.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Dr. Ramón González Guevara and MSc. Boris Alba Valle.
Methodology: Dr. Ramón González Guevara and MSc. Boris Alba Valle.
Investigation: Dr. Ramón González Guevara and MSc. Boris Alba Valle.
Discussion of results: Dr. Ramón González Guevara and MSc. Boris Alba Valle.
Writing – Original Draft: Dr. Ramón González Guevara.
Writing – Review and Editing: MSc. Boris Alba Valle.
Resources: MSc. Boris Alba Valle.
Supervision: Dr. Ramón González Guevara and MSc. Boris Alba Valle.
Approval of the final text: Dr. Ramón González Guevara and MSc. Boris Alba Valle.

VII. REFERENCES

[1] IEC 60743: 2013. "Live working IEC 60743: 2013. "Live working – Terminology for tools, devices and equipment". Edition 3.0. IEC Central Office. Geneva. Switzerland. Fecha de consulta: 07/05/2020, Available in: https://webstore.iec.ch/publication/3169

[2] ASTM F712: 2018. "Standard test methods and specifications for electrically insulating plastic guard equipment for protection of workers". ASTM International. West Conshohocken. United States. Fecha de consulta: 07/05/2020, Available in: https://www.astm.org/Standards/F712.htm

[3] ASTM F478-14a: 2019. "Standard specification for in service care of insulating line hose and cover". ASTM International. West Conshohocken. United States. Fecha de consulta: 07/05/2020, Available in: https://www.astm.org/Standards/F478.htm

[4] R. G. Guevara, and B. A. Valle, "Protectores para trabajos en líneas eléctricas energizadas", Monografía, Centro de Investigaciones y Pruebas Electroenergéticas (CIPEL), Facultad de Ingeniería Eléctrica, Universidad Tecnológica de La Habana "José Antonio Echeverría", Cujae, La Habana, Cuba, ISBN: 978-959-261-573-1, Dic. 2017.

[5] R. G. Guevara, B. A. Valle, M. A. R. Moreno, and G. H. Puente, "Equipos de protección aislantes para trabajo con tensión. Comportamiento en condiciones de laboratorio", Congreso Internacional en Alta Tensión y Aislamiento Eléctrico (ALTAE), La Habana, Cuba, ISBN: 978-959-261-572-4, Nov. 2017.

[6] IEC 60060-1: 2010. "High-voltage test techniques – Part 1: General definitions and test requirements". Edition 3.0. IEC Central Office. Geneva. Switzerland. 2010. Fecha de consulta: 07/05/2020, Available in: https://webstore.iec.ch/publication/300

[7] B. A. Valle, "Base de datos de resultados de ensayos del Laboratorio de Alta Tensión del CIPEL". Base de datos no publicada. Laboratorio de Alta Tensión, Centro de Investigaciones y Pruebas Electroenergéticas (CIPEL), Facultad de Ingeniería Eléctrica, Universidad Tecnológica de La Habana "José Antonio Echeverría", Cujae, La Habana, Cuba, 2019.

[8] CHANCE Catalog. "Tools and Grounding Catalog". Hubbell Power Systems, Inc. Missouri. United States. March 2020. Fecha de consulta: 07/05/2020, Available in: https://hubbellcdn.com/catalogfull/Tools_Grounding-CatFull.pdf

[9] SALISBURY Catalog. "Your #1 resource for personal electrical-safety protection". W.H. Salisbury Co. North Long Avenue. USA. 2000. Fecha de consulta: 07/05/2020, Available in: https://manualzz.com/doc/11501856/salisbury-electrical-safety-insulating-gloves

[10] RITZ Catalog. "High and medium voltage line tools and equipmet". RITZ do Brasil S.A. Teresópolis. Brasil. 2008. Fecha de consulta: 07/05/2020, Available in: https://studylib.net/download/18798434

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.6 n.25, p. 37-42, Sep/Oct, 2020

DOI: https://doi.org/10.5935/jetia.v6i25.701

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

DESIGN AND TUNING OF PID ALGORITHM FOR OPTIMUM PERFORMANCE OF PVTOL SYSTEM

Badri Narayan Mohapatra*[1], Shubham Gadekar[2], Rutuja Zate[3] and Dhavalsinh Bhosale[4]

[1, 2, 3, 4] University of Savitribai Phule Pune University, Pune, India.

[1] http://orcid.org/0000-0003-1906-9932, [2] http://orcid.org/0000-0002-2526-4497, [3] http://orcid.org/0000-0003-4453-7072
[4] http://orcid.org/0000-0002-5824-9664

Email: *badri1.mohapatra@gmail.com, shubhamgadekar27@gmail.com, rutuja.zate@gmail.com, bhosale.dhavalsinh831@gmail.com

ARTICLE INFO

Article History
Received: August 12th, 2020
Accepted: October 20th, 2020
Published: October 30th, 2020

Keywords:
Quadcopter,
VTOL,
Ziegler Nicholas,
Genetic Algorithm,
Simulink.

ABSTRACT

In this paper different design scheme for a PID controller have been introduced for a single axis of a quadcopter. This type of model is also known as PVTOL (planar vertical take-off and landing) system. The PVTOL system possess complicated roll control schemes, non-linearity, low stability and is a second order type process. This paper aims to present a comparison between different controllers used in a dynamic model of a PVTOL platform. Performance comparison of classical Zeigler Nicholas (ZN-PID) is done against Genetic Algorithm (GA) based controller optimization. The results are obtained using MATLAB and SIMULINK, the (ZN-PID) and (GA) based controller is designed for disturbance rejection, close loop response and set point tracking.

I. INTRODUCTION

During the years 1950 – 1970 experiments related to Vertical Take Off and landing (VTOL) vehicle came into existence. It is also named as Short run Take Off and Landing. VTOL consists of three types of configurations i.e. Wining type configuration which has fixed wings with vector thrust engine, Helicopter type configuration consist of moving wings with engine, Ducted type configuration consist of ducted rotor which helps to lift. All the configuration helps to take off and land vehicle vertically. VTOL has ability to fly slowly and land in small places as well. Due to increase in demand of VTOL vehicles, aircrafts and quadrotors, The Hover eye platform from Bertin Technologies was introduced. It was a first step towards Unmanned VTOL [1]. In [2] study of different types controllers for Unmanned Aerial Vehicle is given.

In [3] [4] stability control of VTOL system is discussed. Due to which usage of these types of system increased in various fields. As the demand of VTOL systems is increasing day by day it is used in field of aerospace and photography. According to recent research, VTOL uses batteries or motors instead of fuel, it reduces maintenance cost value. Usage of motors helps in reduction of noise pollution and gas emission as well. Nonlinear controller can be developed to stabilize the altitude by considering dynamic behavior of motors [5].

There are two forms of motors used for distinct purposes. The primary one is Brushed DC motor, typically known as Ordinary direct current motor. It acquires large space with giving less amount of torque with less efficiency and low power density. This type of motor is not suitable for developing generation equipment. New generation equipment must consist of great stability, low cost maintenance, reduced development time, adaptive to various conditions, as well as is must involves higher performance parameters such as high efficiency and reduced electromagnetic interferences [6].

After considering all the above required parameters, another form of motor adequately fulfils the conditions.

Another form of motor is Brushless DC Motor (BLDC). In [7] performances of BLDC motors are mentioned. BLDC motors are highly demanded in areas which requires interpretative

performance due to their small size, torque and compatible structure. In [8] parameters of motor and aspects related electronic devices is focused.

The BLDC motor works on same principles as of brushed DC motor which is internal shaft position feedback. In BLDC motor mostly we use hall effect sensor method. Here the current carrying conductor is stationary while the permanent magnet moves. When the stator coils are electrically switched by a supply source, they become electromagnet and starts to produce a uniform field in air gaps. Even the supply is DC, switching generates an AC waveform with trapezoidal shape. Most BLDC motors have three Hall sensors fixed firmly into stator on the non-driving end of the motor. When the magnetic pole of rotor passes near from hall sensors, they give a high or low signal. Due to switching of windings as high and low signal, corresponding winding energized as North and South poles. The motor produces torque because of development of attraction forces and repulsion forces. Due to this motor moves in a clockwise direction and compensation technique one can apply as per requirement [9].

Dynamic equation is established between rotor and stator to represent the synchronization error in [10].

The absence of brushes in a BLDC Motor is perhaps biggest advantage. It generates less noise and it is also less prone to sparking due to the lack of a commutator. The BLDC motor requires low maintenance than Brush DC motors. BLDC motors has higher speed ranges, high dynamic responses and it lasts longs in total operating hours. BLDC motors required less space and are lighter in weight. Working of BLDC motors under variable loads and reference speeds are given in [11].

This paper specifically aims to design a system of single axis quadcopter in which PID controller is used. The rest of the paper is arranged as follows. Section 2 shows the system and its description, in this section physical description of system is mentioned. Section 3 depict the dynamic modelling of system, in this section forces acting on the system is explained. Section 4 deals with designing of PID controller, in this different method for designing of controller is mentioned with its simulation model w.r.t MATLAB. While section 5 and section 6 deal with results and conclusion respectively.

II. SYSTEM AND DISCRIPTION

The PVTOL prototype is rather designed to move on an inclined plane. The general view of our experimental setup is depicted in Figure 1. The PVTOL moves on an inclined plane, which defines our two-dimensional (2-D) workspace. The size of the inclined plane is 60cm(L) * 5cm(W). A 10cm(L) threaded rod is used to fix the clamps at the center of the inclined plane. The threaded rod is fixed exactly at the midpoint of the inclined plane so that a pivot point is achieved. A pivot point is required so that the inclined wooden plane can roll in a particular axis. The inclination of the wooden plane is 30deg. The size of each propeller is 10-cm long. Two high speed and high torque BLDC motors are fixed at the end of the inclined wooden plane. One motor rotates in clockwise direction and other rotates in counter-clockwise. The two BLDC motors are driven by driven by two separate ESC (electronic speed controller). The total thrust is the sum of the thrusts of individual motors. The rolling moment is obtained by increasing the speed of one motor while decreasing the speed of the second motor and vice versa. Each motor is linked to a speed variator which is itself linked to an IMU (inertial moment unit).

Figure 1: System model (Hardware).
Source: Authors, (2020).

It is a device capable of measuring the force (acceleration) and speed. Generically it consists of an Accelerometer and a Gyroscope. The IMU is placed at the center of the inclined wooden plane, over the pivot point. The gyroscopes improve the maneuverability and the stability of the system. Dimensions of the components used in Figure 1 are listed in the Table 1.

Table 1: Parameters of the system.

Beam Length	60 cm * 5 cm
Beam weight	250gm
Propeller size	10 cm(L)
Motor operating voltage	12V
Motor weight	55gm
ESC output	30A
ESC weight	22gm

Source: Authors, (2020).

III. DYNAMIC MODELLING

The PVTOL is a model of a flying object that evolves in a vertical plane. More specifically, this aerial vehicle basically consists of two propellers at the end of the single axis. This model has minimum number of states and inputs but retains many of the features that need to be considered while designing control laws for a real aircraft. It has three degrees of freedom (x, y, φ) corresponding to its position and orientation in the plane. It has two thrusters that produce a force and a moment on the system as seen in Figure 1 [12]. Here F1 and F2 are the vertical forces and z is the rolling moment. The factor g denotes the acceleration due to gravity. Thus, PVTOL is an underactuated system with three degrees of freedom and two inputs.

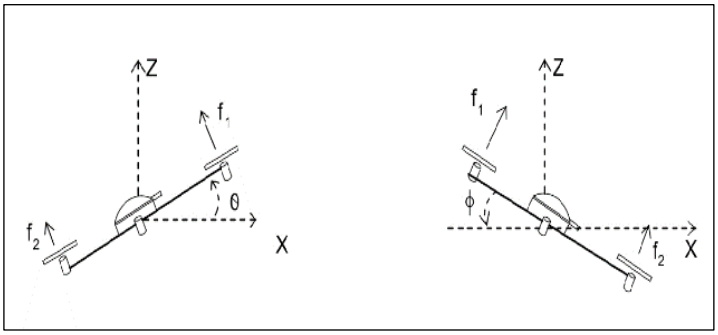

Figure 2: Schematic diagram of system.
Source: Authors, (2020).

We are going to propose the control of X axis rotation of the system, therefore we assume that the Y and Z axes as zero. Hence the equation of the rolling dynamics is:

$$\emptyset = F \cdot d - \mathcal{C}\,\ell_{\dot{\phi}} \cdot \dot{\phi} \quad (1)$$

Where, F=f1-f2 is the force produced by the thrust difference of the motors, d is the center of mass of each motor and Cl0 is the coefficient of damping on the x-axis. In this case Cl0 is 0.36.

The equations and parameters needed to generate the matrices are estimated based on empirical experimentation and measurments on control of unmanned vehicle. This work assumes the equations presented in [13] as described below.

The simplified transfer function that describes the rolling dynamics of the system is shown in Equation (2).

$$\frac{\phi(s)}{F(f)} = \frac{5}{S^2 + 25s} \quad (2)$$

IV. DESIGN OF PID CONTROLLER

In this section designing of PID controller is developed using different methods. The PID controller is most commonly used versatile technique. Given Figure 3 shows the simulation model of PID controller executed in Simulink. The form of PID controller is given in Equation (3) below.

$$\text{PID Output} = k_p + \frac{k_i}{s} + k_d s \quad (3)$$

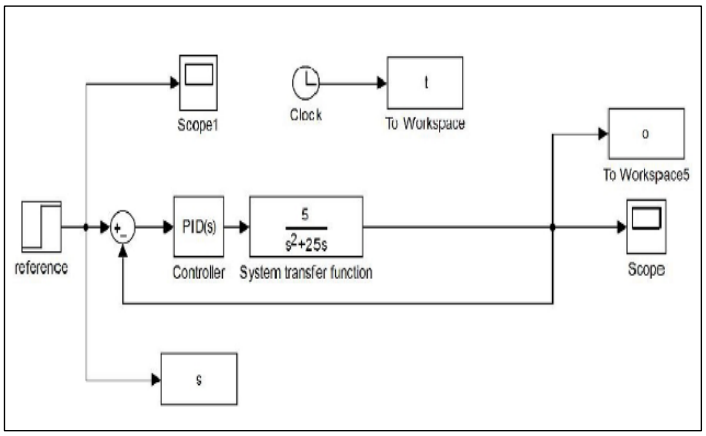

Figure 3: Simulation model.
Source: Authors, (2020).

IV.1 ZIEGLER NICHOLAS (ZN) - PID

The Ziegler–Nichols tuning method is an experimental method for tuning a PID controller by controlling. It is also known as ultimate cycle method. This tuning method is employed to avoid disturbance by PID loops. It is performed by adjusting the I (integral) and D (derivative) gains to zero. The proportional gain, is then increased from zero to the ultimate gain. The oscillation period are then used to set the P, I, and D gains depending on the type of controller used and its behavior [14]. The Ziegler-Nichols rule assumes that the system has a transfer function of the following form:

$$K \cdot \frac{e^{-sT}}{a + s} \quad (4)$$

This method can be easily put in to tune a PID controller by using the relations provided in Table 2.

Table 2: Relations between controllers.

Type of controller	k_p	T_i	T_d
PID	$0.6K_u$	$0.5T_u$	$0.125T_u$

Source: Authors, (2020).

Iterations are carried out with the help of MATLAB to find optimal values of K_u and T_u. Based on the results of

K_u and T_u, the corresponding values of controllers are:

$k_p = 150$
$T_i = 0.08$
$T_d = 0.02$

IV.2 GENETIC ALGORITHM (GA) - PID

A Genetic algorithm (GA) is a process of natural selection that belongs to the larger class of evolutionary algorithms (EA). Genetic algorithm has the same role as Artificial Intelligence. Genetic Algorithm is sometimes considered to be robust. Genetic Algorithms are known for its high performance in complex areas without experiencing the difficulties. In [15] the GA tuned (proportional, integral and differential) PID controller surpass the developed PID controller. GA adjusts on changing inputs and also be able to handle noisy or fuzzy input. GA is able to handle complex problems better. Genetic algorithms simulate the process of natural selection that means those parameters who can adapt to external changes and are able to work and give required output. The GA based optimization tool in MATLAB is used for finding optimum PID settings are given in this paper. The fitness function is designed according to optimize the performance criteria.

```
1/5/10 10:25 AM  C:\Users\vmsdemonuser\Doc...\pid_optim.m   1 of 1

function [J] = pid_optim(x)
s=tf('s');
num=[5];
den=[1 25 0]
Plant=tf(num,den);

Kp=x(1)
Ki=x(2)
Kd=x(3)

cont = Kp+Ki/s+Kd*s;

step(feedback(Plant*cont,1));

e = 1 - step(feedback(Plant*cont,1),t);

J=sum(t'.*abs(e)*dt);
```

Figure 4: MATLAB code.
Source: Authors, (2020).

Table 3: Performance indices comparison.

Performance Index	ZN – PID	GA – PID
Settling time(s)	0.525	3.81
Peak Response	1.37	1.08
Overshoot %	37.3	7.51
Rise time(s)	0.0513	0.517
Final Value	1	1

Source: Authors, (2020).

Based on the results of MATLAB corresponding values of controllers are:

$k_p = 4.51$
$T_i = 13.51$
$T_d = 4.51$

The effects of lead and lag compensator shows in oscilloscope Figure 2 and 4. Normally damping which will be more in case of led compensator. Due to this less rise time and less overshoot.

With the observation from oscilloscope one can easily find the lead and lag phase difference. Depending on the values of register and capacitor, the circuit behaves as lead and as well as lag. Any type of correction if a system will require then compensator circuit is very useful. Some cases lead lag compensator that is combination circuit is also useful. Satellite lunching, automobile, robotic control it's use is more.

V. SIMULATIONS AND RESULTS

In this section, we present some simulation results using MATLAB and SIMULINK in order to observe performance of proposed systems. We have considered two different types of controlling systems. The performance evaluation of the controllers is done by rating based on settling time, overshoot and peak response for open loop response and set-point changes and disturbance rejection. The Performance indices obtained for ZN-PID and GA are presented in Table 3 and Figure 5-6 show responses of the system for reference tracking set-point change and disturbance rejection.

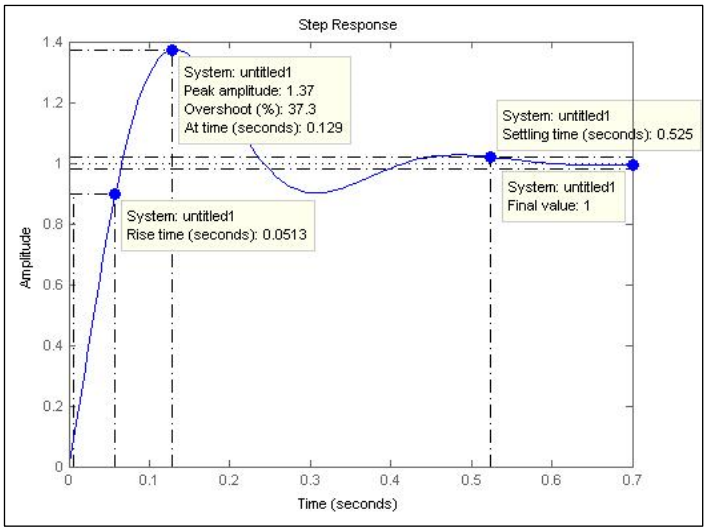

Figure 5: Ziegler Nicholas Simulation model.
Source: Authors, (2020).

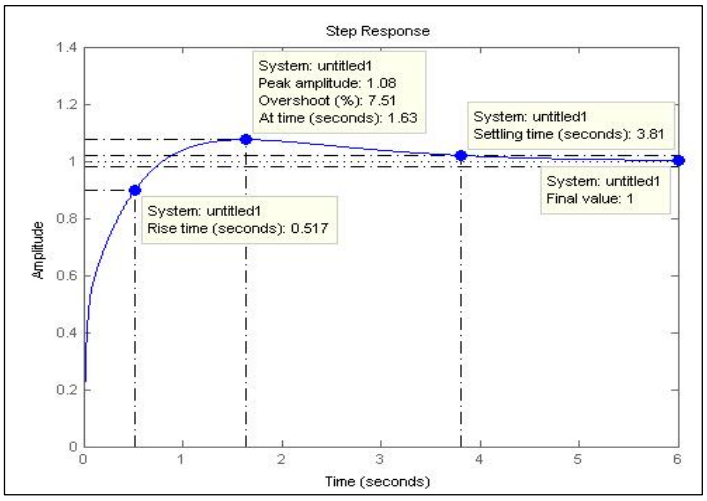

Figure 6: Genetic algorithm simulation model.
Source: Authors, (2020).

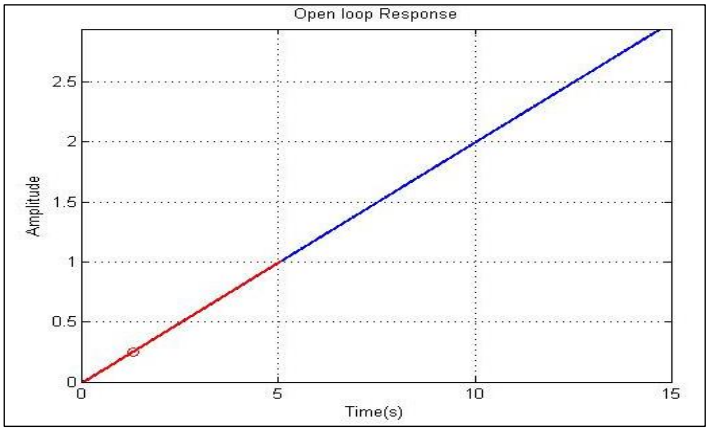

Figure 7: Open loop response model.
Source: Authors, (2020).

Figure 7 depicts the open loop response of the proposed system. It shows as the input values increases output increases accordingly, results given by the system in open loop graph depicts the graph in linear way which shows that the input given is directly proportional to the requires output as there is no feedback given. Present system is unstable as pole and zero lies on right side.

The output response of system shows the relation between input given to the system and Rolling effect.

In this paper single axis of a quadcopter is consider it deals with rolling effect (i.e. tilt left or tilt right) of an axis. From Figure 7 it shows that Rolling effect of the system increases with increase in input.

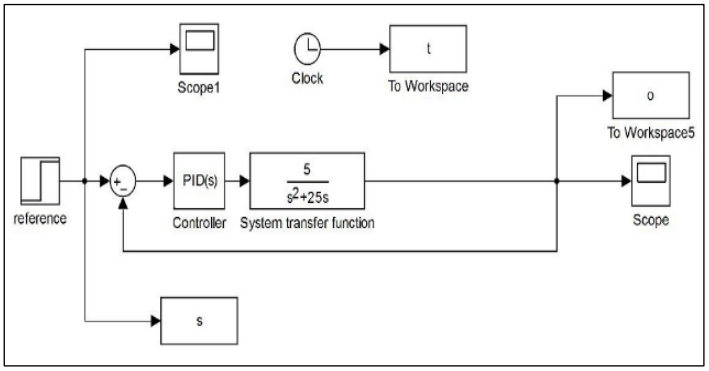

Figure 8: Close loop simulation model.
Source: Authors, (2020).

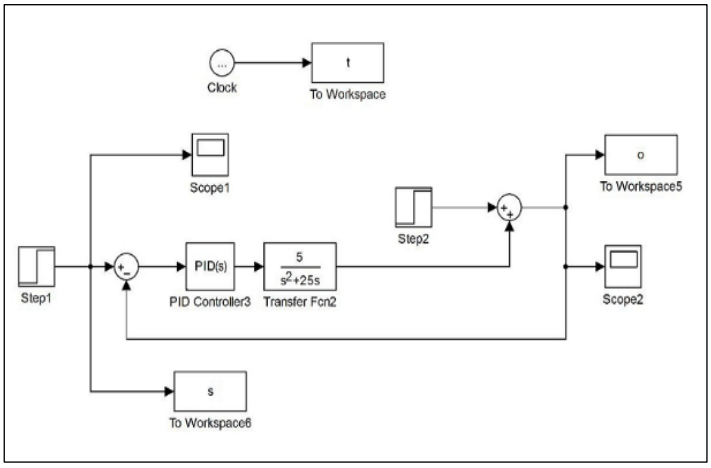

Figure 9: Disturbance rejection simulation model.
Source: Authors, (2020).

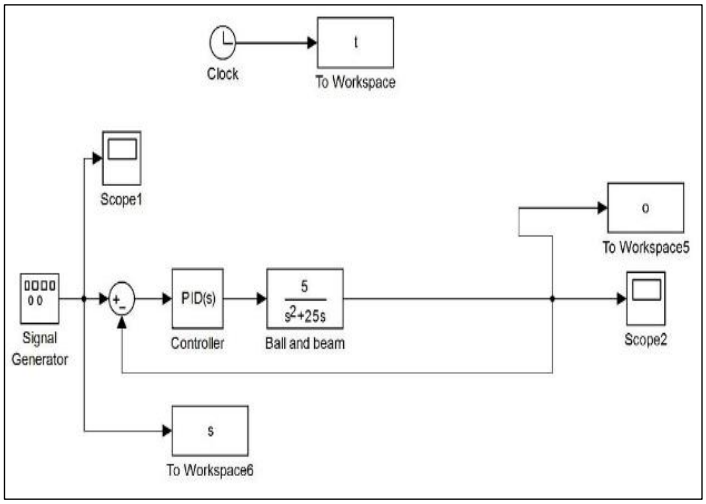

Figure 10: Set-point simulation model.
Source: Authors, (2020).

The Figures 8-10 shows Simulation model for a system proposed for finding closed loop response, set point tracking and disturbance rejection of the system. For closed loop response a step signal is used whereas for set point tracking a square wave with 0.5HZ frequency is used and for disturbance rejection a spike of 0.5 amplitude is introduced in the system at 3 seconds time interval.

The above Simulink models are used to generate system results which are given below:

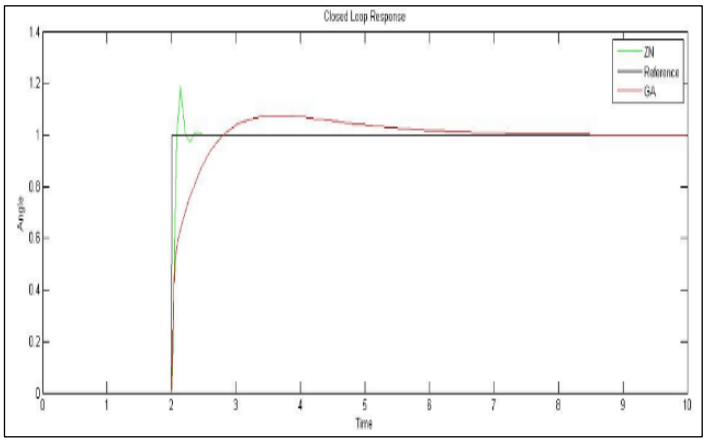

Figure 11: Close loop response.
Source: Authors, (2020).

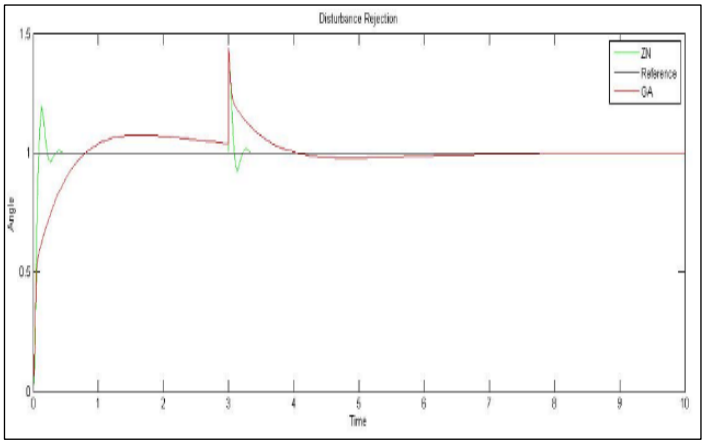

Figure 12: Disturbance rejection response.
Source: Authors, (2020).

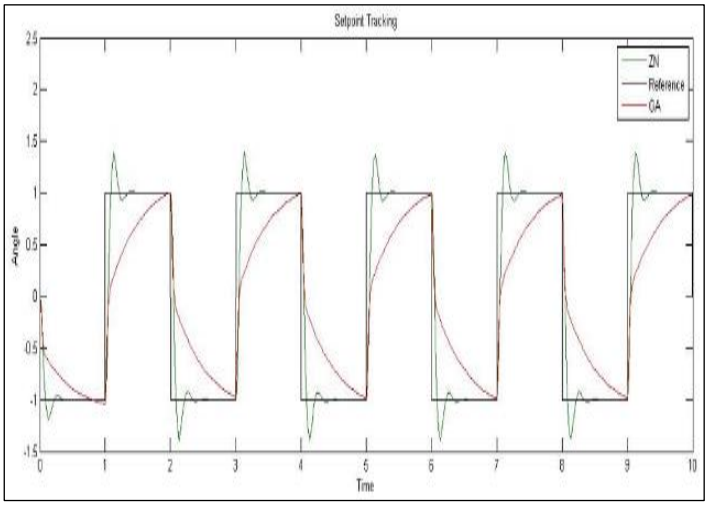

Figure 13: Set-point response.
Source: Authors, (2020).

VI. CONCLUSIONS

The PID controller was tuned using Zeigler Nicholas and Genetic Algorithm optimization method and the simulations are carried on MATLAB SIMULINK environment. Considering the analysis of the data found in this work the different control methods present satisfactory results. Performance evaluations of both the controllers are done on the basis of set point tracking and disturbance rejection. A square wave is used as a reference for the set point tracking and disturbance rejection is evaluated by introducing a spike in system at three seconds interval. Each of this controller provide singular characteristics that makes it difficult to say which one is the best. The settling time obtained from Zeigler Nicholas tuning method is less, while the overshoot is large which cause non-linearity in the system. The Genetic Algorithm optimization method has comparatively higher settling time, but the overshoot is less making the system more responsive in the speed control of linear brushless DC motors.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Dr. Badri Narayan Mohapatra, Shubham Gadekar and Rutuja Zate.
Methodology: Dr. Badri Narayan Mohapatra and Shubham Gadekar.
Investigation: Shubham Gadekar and Dhavalsinh Bhosale.
Discussion of results: Dr. Badri Narayan Mohapatra, Shubham Gadekar, Rutuja Zate and Dhavalsinh Bhosale.
Writing – Original Draft: Shubham Gadekar and Rutuja Zate.
Writing – Review and Editing: Dr. Badri Narayan Mohapatra and Shubham Gadekar.
Resources: Rutuja Zate and Dhavalsinh Bhosale.
Supervision: Dr. Badri Narayan Mohapatra and Shubham Gadekar.
Approval of the final text: Dr. Badri Narayan Mohapatra, Shubham Gadekar and Rutuja Zate.

VIII. ACKNOWLEDGMENTS

I would like to thank AISSMS IOIT undergraduate students that they have complete on timely collection of compilation of results in the research work of their final year BE Project. I have appreciated their motivation towards this Project research.

IX. REFERENCES

[1] Y. Wang, "The design, prototyping, modeling and flight tests of a vertical take-off and landing (VTOL) unmanned aerial vehicle (UAV)." DOI: 10.14711/thesis-991012535962203412.

[2] R. Czyba and G. Szafranski, "Control Structure Impact on the Flying Performance of the Multi-Rotor VTOL Platform - Design, Analysis and Experimental Validation." International Journal of Advanced Robotic Systems, vol. 10, no. 1, p. 62, 2013, DOI: 10.5772/53747.

[3] Feng Lin, W. Zhang, and R. Brandt, "Robust hovering control of a PVTOL aircraft." IEEE Transactions on Control Systems Technology, vol. 7, no. 3, pp. 343-351, 1999, DOI: 10.1109/87.761054.

[4] A. Palomino, P. Castillo, I. Fantoni, R. Lozano, and C. Pegard, "Control strategy using vision for the stabilization of an experimental PVTOL aircraft setup." IEEE Transactions on Control Systems Technology, vol. 13, no. 5, pp. 847-850, 2005, DOI: 10.1109/tcst.2005.847347.

[5] X. Huo, M. Huo, and H. R. Karimi, "Attitude Stabilization Control of a Quadrotor UAV by Using Backstepping Approach." Mathematical Problems in Engineering, vol. 2014, pp. 1-9, 2014, DOI: 10.1155/2014/749803.

[6] "Design of Speed-Controller for Brushless DC-Motor Based on Grey Predictor-PID Controller." Engineering and Technology Journal, vol. 36, no. 8, 2018, DOI: 10.30684/etj.36.8a.9.

[7] Mohapatra, Badri Narayan, and Rashmita Kumari Mohapatra. "Application of DC motor as speed and direction control." (2020).

[8] "WSEAS Transactions on Systems and Control." DOI: 10.37394/23203.

[9] B. N. Mohapatra and J. Joshi, "Design performance of lead and lag compensator using OPAMP and root locus approach through simulation tool." ITEGAM- Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), vol. 6, no. 24, 2020, DOI: 10.5935/jetia.v6i24.683.

[10] Z. Wu, H. Lyu, Y. Shi, and D. Shi, "On Stability of Open-Loop Operation without Rotor Information for Brushless DC Motors." Mathematical Problems in Engineering, vol. 2014, pp. 1-7, 2014, DOI: 10.1155/2014/740498.

[11] M. Ramesh, G. Rao, J. Amarnath, S. Kamakshaiah, and B. Jawaharlal, "Speed torque characteristics of brushless DC motor in either direction on load using ARM controller." ISGT2011-India, 2011, DOI: 10.1109/iset-india.2011.6145385.

[12] C. Venkatesh, R. Mehra, F. Kazi, and N. Singh, "Passivity based controller for underactuated PVTOL system." 2013 IEEE International Conference on Electronics, Computing and Communication Technologies, 2013, DOI: 10.1109/conecct.2013.6469304.

[13] Jiřinec, Tomáš. "Stabilization and control of unmanned quadcopter." (2011).

[14] J. G. Ziegler and N. B. Nichols, "Optimum Settings for Automatic Controllers." Journal of Dynamic Systems, Measurement, and Control, vol. 115, no. 2, pp. 220-222, 1993, DOI: 10.1115/1.2899060.

[15] D. Fogel, "An introduction to genetic algorithms Melanie Mitchell. MIT Press, Cambridge MA, 1996. $30.00 (cloth), 270 pp." Bulletin of Mathematical Biology, vol. 59, no. 1, pp. 199-204, 1997, DOI: 10.1016/s0092-8240(96)00095-x.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.6 n.25, p. 43-47, Sep/Oct, 2020

DOI: https://doi.org/10.5935/jetia.v6i25.702

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

BEHAVIOR OF BIOMATERIALS IN RELATION TO THE CONSERVATION TIME AND TEMPERATURE: STABILITY TEST OF A BIOPRODUCT

Joab de Souza Arouche*[1]

[1] Federal University of Amazonas - UFAM. Manaus-Amazonas, Brazil.

[1] http://orcid.org/0000-0002-4254-0529

Email: *arouchejoab16@gmail.com

ARTICLE INFO

Article History
Received: September 23th, 2020
Accepted: October 20th, 2020
Published: October 30th, 2020

Keywords:
Biomaterials,
Bioproducts,
Stability test,
Bioactive,
Evaluation parameters.

ABSTRACT

All products before reaching the consumer's table go through a battery of analyzes, in order to check if these materials resist the weather, temperature, climate changes, friction, including transportation. Whether they really support physical, chemical and biological factors and also with the idea of estimating the useful life of these materials. Biomaterials are no different, all bioproducts are also subjected to resistance tests to reach industry, or long-term production. Among the numerous possibilities that exist for the application of biomaterials, recently the bioactive ones encapsulated in polymeric nanoparticles of controlled action stand out with their several alternatives of application, among them as natural insecticides, aiming at the reduction of conventional pesticides, including fungicides, used in the activation and controlled release of the drug. Thus, this research aims to develop a formulation and verify its behavior as a function of time and temperature (25 °C), as well as the application of different preservatives. , [phenoxyethanol-2-methyl-2H-isothiazolin-3-one-NE), citric acid and thymol] in order to verify its influence on the stability of the formulations as a function of time. Among all tested formulations, the one containing the preservative NE was the only one approved in all parameters.

I. INTRODUCTION

All products before reaching the consumer's table go through a battery of analyzes, in order to check if these materials resist the weather, temperature, climate changes, friction, including transportation. Whether they really support physical, chemical and biological factors and also with the idea of estimating the lifetime of these materials [1]. Biomaterials are no different, all bioproducts are also subjected to resistance tests to reach industry, or long-term production.

Biomaterials have an expressive representativeness in the health area, not being restricted only to the development of mechanical devices (robots - prototypes), but also in the creation of bioproducts with controlled action [2].

Recently, it has become normal to hear the term biomaterials, which are those applied to biological and health issues. For the development of biomaterials (bioproducts), joint action is required between health professionals (doctors, dentists, physiotherapists and speech therapists) and professionals in the exact area (engineers, chemists, physicists and biologists).

Among the countless possibilities that exist for the application of biomaterials, recently the bioactive ones encapsulated in polymeric nanoparticles with controlled action stand out with their several alternatives of application, among which as natural insecticides, aiming at the reduction of conventional pesticides, fungicides, as well as medicinal application with controlled activation and release. Figure 1 shows an illustrative example of a natural asset covered (encapsulated) by a biodegradable polymeric layer.

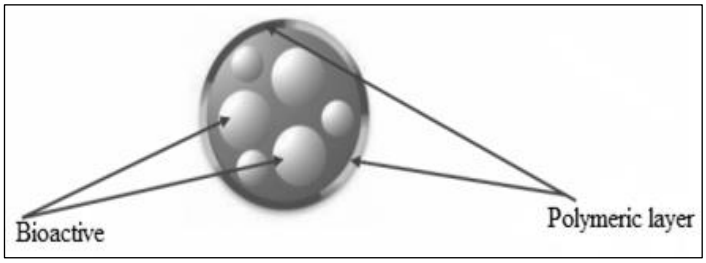

Figure 1: Illustrative image of a natural asset encapsulated by polymeric nanoparticles.
Source: Author, (2020).

Technological advances have allowed the development of biomaterials containing natural active substances [3]. Effective bioproducts, of low toxicity and low cost, as is the case of controlled release biocides based on the encapsulation of natural bioactives of various plant species rich in biologically active substances against fungi, bacteria, viruses, larvae, insects, agricultural pests and pathologies that affect human health [4-12].

One of the objectives of encapsulating biomaterials is to protect and prolong the durability of certain constituents of the formulation until the moment of its application [13], promoting a controlled action [14]. For this reason, this technique has been widely applied to encapsulate essential oils, as they are volatile substances susceptible to photodegradation.

Thus, this research aims to develop a formulation, based on the creation of a biomaterial from oil encapsulated in biodegradable polymeric nanoparticles and developed based on gelatin and poly-caprolactone (PCL). It was evaluated as a function of time and temperature (25 °C), as well as the application of different preservatives, [phenoxyethanol-2-methyl-2H-isothiazolin-3-one-NE), citric acid and thymol] with the intention of verifying its influence on the stability of formulations as a function of time. In addition, the stability assessment was also monitored by estimating the encapsulation efficiency (EE), electrical conductivity, turbidity and pH measurements, and evaluating the organoleptic properties (color and odor), periodically monitored to select the most stable formulation.

II. MATERIALS AND METHODS

Usually the preparation of biodegradable formulations is composed of two phases, aqueous and organic. For the aqueous phase, 1 g of gelatin was solubilized in 100 ml of distilled water followed by stirring and heating to 50 °C. In another beaker, 0.3 g of Tween 80 was dissolved in 50 ml of distilled water with magnetic stirring until total solubilization. Then, Tween 80 was poured into the solubilized gelatin. For the organic phase, the PCL was solubilized in 5 ml of dichloromethane under magnetic stirring. 0.02 g of SPAN, 0.1 g of TACC and 0.075 g of natural Amazonian active substance homogenized in 5 mL of dichloromethane were used. After stirring, SPAN, TACC and the natural asset were poured into the solubilized PCL giving rise to the organic phase. After the magnetic stirring was completed, the turrax-type ultra-disperser was used to pour the organic phase into the aqueous phase for 30 s. Then, 0.0935 g of transglutaminase was added under magnetic stirring, adjusting the pH to 8. After the preparation of the synthesis, it was divided into three vials and a preservative was added to each solution. The preservatives used were NE, citric acid and thymol, so the formulations were kept in an incubator with a controlled temperature of 25 °C. To choose the formulation with the most stable preservative and with the greatest temperature resistance.

The stability evaluation of the formulations was performed using a pH meter from BEL, Model PHS-BW; Turbidimeter Brand MS TECNOPON Model TB-1000; Conductivimeter Brand BEL, Model W12D; UV-vis Global Trade Technology Model UV 5100. The evaluation of organoleptic properties (color and odor) was carried out in order to verify physical-chemical changes associated with possible changes. For the EE estimate, the empty and loaded nanoparticles were separated by centrifugation (DAIKI, model DTR 20.000) at 15.000 rpm for 15 min. The supernatant was used to determine the amount of free natural assets.

III. RESULTS AND DISCUSSIONS

Were evaluated (I) the formulation containing the encapsulated natural active and the preservative NE, (II) the formulation containing the encapsulated natural active and the citric acid preservative, (III) the formulation containing the encapsulated natural active and the preservative thymol. The purpose of stability assessment was to verify possible changes in relation to destabilization of the formulations, in addition to checking which formulation is more stable.

The stability tests were started in a centrifuge by means of a behavioral stability study, being evaluated the state of the formulations after centrifugation. The behavioral stability study aims to verify if the formulations support pressure, mobilization, temporal changes without phase separation [15]. The formulations were submitted to 15.000 rpm and no phase disjunction was observed.

After the completion of the behavioral study of the formulations by the centrifugation test, the stability analysis was performed at 25 °C, and the parameters evaluated were organoleptic properties (color and odor), turbidity, efficiency of encapsulation, pH and electrical conductivity. All parameters were analyzed over time.

All products need to undergo several stability assessments, industrial or manipulated. The initial stability tests aim to verify the product quality before release to the consumer. These tests are performed to guarantee its stability until the end of its validity, that is, for ensure that the product effectively performs the action for which it was created [16].

The organoleptic properties were the first aspect to be evaluated (color and odor), were analyzed in comparison with the initial characteristics of the standard formulation to verify changes in color and odor, in addition to phase separation, excessive turbidity and ruptures.

Regarding organoleptic properties, the formulations containing the citric acid and thymol preservatives showed changes in color and odor after 25 days, as shown in Figure 2.

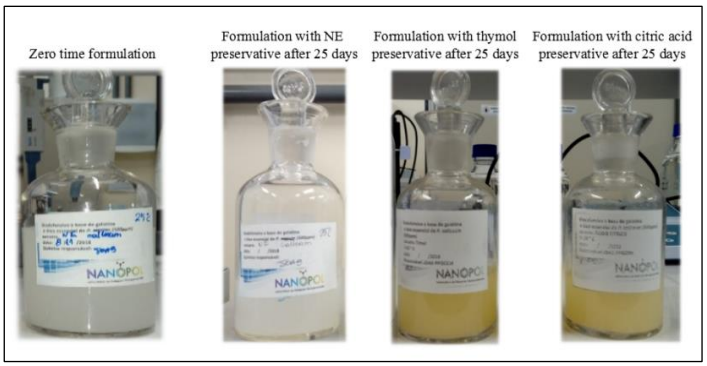

Figure 2: organoleptic properties, the formulations.
Source: Author, (2020).

These results suggest a possible proliferation of microorganisms [17]. As they also indicate that preservatives were not appropriate. However, the formulation containing the preservative NE did not change in the same evaluation period. Therefore, the formulation containing this preservative was the only one approved in the study of organoleptic properties. Corroborating the signs of destabilization verified by the evaluation of the organoleptic properties of formulations containing citric acid and thymol preservatives, the turbidity analyzes of these formulations also showed changes. It is known that the destabilization of biodegradable systems is related to physical, chemical and biological factors [16, 18-19]. However, in the formulation containing the preservative NE, no significant changes were observed in the turbidity values capable of compromising the stability of the bioproduct, corroborating the evaluation of organoleptic properties, as shown in Figure 3.

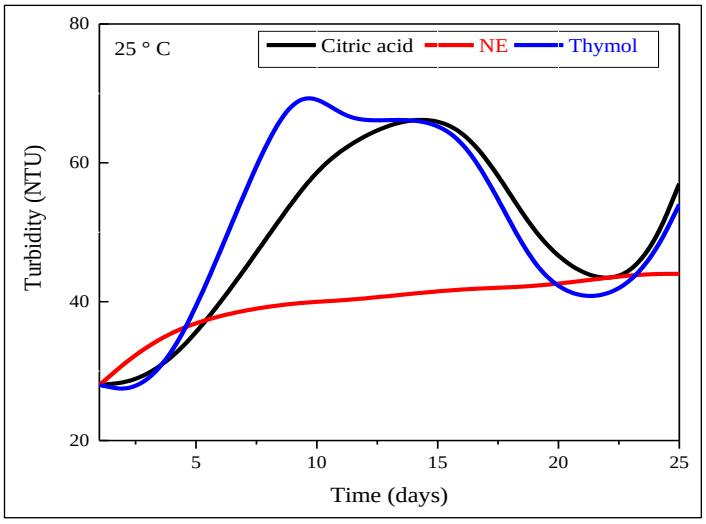

Figure 3: Turbidity (NTU).
Source: Author, (2020).

All formulations showed encapsulation efficiency (EE) greater than 90%, indicating that the combination of materials applied to encapsulate the natural asset was effective. The formulation containing the citric acid preservative showed an initial EE of (97 ± 2) % and a final value of (91 ± 2) %. The formulations containing the preservative NE and thymol showed an initial EE of (95 ± 2) % and, at the end of the analysis, (90 ± 2) %. The formulation containing the preservative thymol was also above average. Figure 04 shows the variation of the EE in an evaluation period of 25 days (25 ºC) for formulations containing preservatives NE, citric acid and thymol.

Polymeric biomaterials are considered efficient when their encapsulation efficiency is greater than 70% [20]. If we consider only the analysis of the EE, all formulations would indicate stability, since they presented high EE. For this reason, it is important to associate the stability assessment with other analyzes (such as organoleptic assessment and turbidity) and, as previously seen, the preservatives citric acid and thymol were not able to maintain the stability of the formulations, even with high EE.

Although the formulations containing the citric acid and thymol preservatives have shown destabilization proven by the organoleptic assessment and turbidity, this fact has not compromised the EE, probably because it is a bilayer carrier system, promoting greater protection of the natural asset. Some studies report that biodegradable polymeric systems that have two layers are more resistant to extrinsic factors than those that have only one layer [21-22].

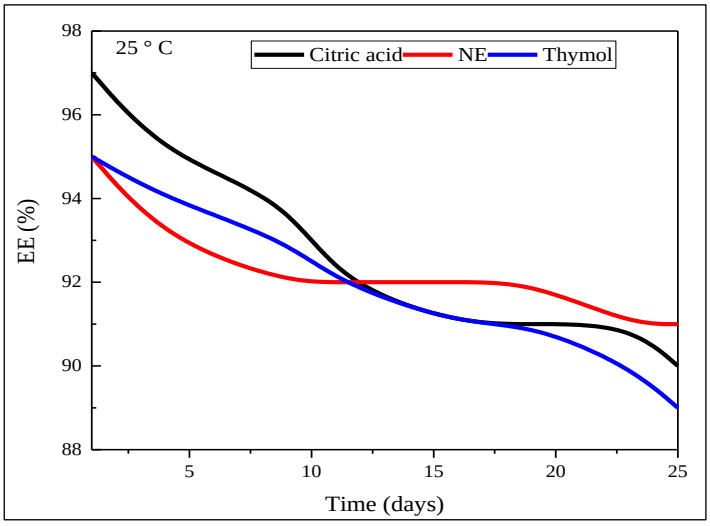

Figure 4: Encapsulation efficiency (EE).
Source: Author, (2020).

During the 25 days of analysis of the stability of the formulations, the pH values decreased, but remained above 6.3. Based on this result, it can be suggested that the natural asset remained protected by the bilayer during the evaluation period (as demonstrated by the high percentage of EE), since lower pH values indicate exposure of the asset to the medium (pH = 5, 5) and destabilization of formulations.

All formulations showed initial pH 8 and, in the final analysis, the results were: formulation containing preservative NE, pH = 6.7; formulation containing citric acid preservative, pH = 7.3 and formulation containing thymol preservative, pH = 6.5.

As shown in Figure 05, all formulations showed a decrease in pH over the 25 days of analysis. This fact is related to the exposure of the bioactive to the aqueous medium, resulting from the degradation of the capsules, causing a decrease in EE, as previously seen. Among all the formulations evaluated, the one that contained the preservative NE was the one that showed the lowest pH variation after the fifth day of analysis. However, as previously mentioned, a single isolated evaluation is not enough to analyze the stability of a formulation, several parameters must be applied for the results to be efficient.

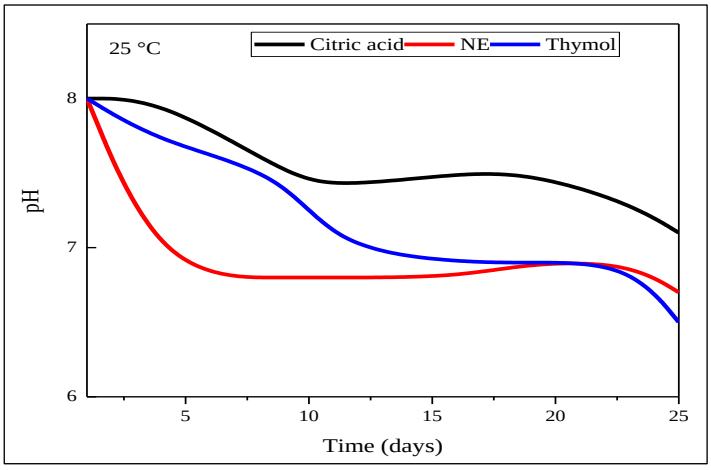

Figure 5: pH.
Source: Author, (2020).

However, the decrease in pH values is not only related to the exposure of the natural asset in the aqueous medium, since the pH of the asset is acidic and, naturally, from its exposure to the medium, the formulation will tend to its pH. The decrease in pH values can also be related to the proliferation of microorganisms in the formulation, such as bacteria and fungi, which naturally changes the pH [23]. In addition, extreme pH changes are suggestive of instabilities, as they indicate degradation of the polymeric material.

The formulation containing the citric acid preservative showed electrical conductivity in the range of 1.9 to 4.0 mS. The formulation containing the preservative thymol showed electrical conductivity from 0.8 to 1.3 mS. The formulation containing the preservative NE showed electrical conductivity between 2.1 and 2.7 mS, as shown in Figure 6. These values corroborate the evaluations of organoleptic properties, turbidity and EE, as unexpected changes in electrical conductivity are indicative of instabilities [15].

The increase in the values of electrical conductivity in systems containing natural assets encapsulated in biodegradable polymeric materials is related to the presence of a large amount of free charge (ions) in the solution. Thus, the increase in electrical conductivity is related to the exposure of the asset in the solution and, consequently, to the decrease in EE and destabilization of the formulation. In some cases, these electrical conductivity values can be changed later, depending on the temperature at which the formulation is packaged. The decrease in pH values can be correlated to the values of electrical conductivity, since this decrease is related to exposure to greater amounts of bioactive substances and, consequently, to increased charges [24].

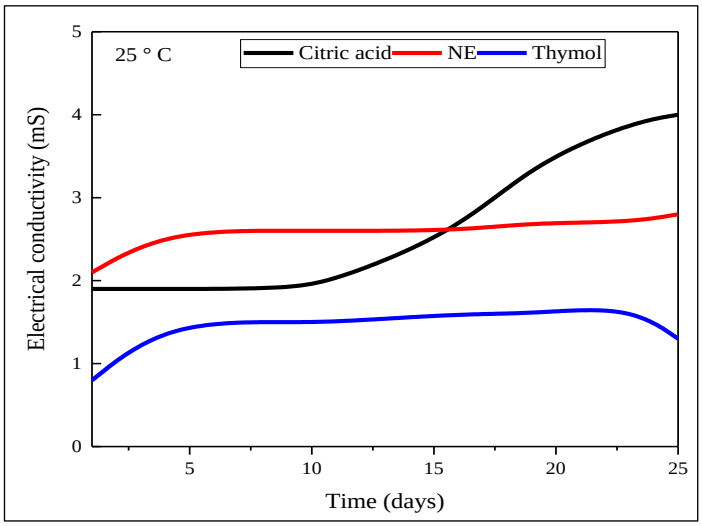

Figure 6: Electrical conductivity (mS).
Source: Author, (2020).

The next stage of the research aimed to choose the most stable formulation using all parameters previously evaluated for 25 days and at 25 ºC, namely: EE, pH, turbidity and electrical conductivity. The values of the parameters obtained in the evaluation are presented in Table 1. Of the formulations evaluated, the one that showed greater stability (that is, the one that was approved in all the evaluated parameters) was the one that contained the preservative NE. The systems containing the other preservatives did not pass all the evaluation parameters. Therefore, the bioproduct containing the preservative NE was selected for further study.

Table 1: Evaluation and preservative parameters.

Preservative	**Turbidity (NTU)**
NE	28 – 44
Citric acid	28 – 57
Thymol	28 – 54
Preservative	**EE (%)**
NE	95 – 90
Citric acid	97 – 91
Thymol	95 – 90
Preservative	**pH**
NE	6,7
Citric acid	7,3
Thymol	6,5
Preservative	**Electric conductivity (mS)**
NE	(2,1 - 2,7)
Citric acid	(1,9 - 4)
Thymol	(0,8 - 1,3)

Source: Author, (2020).

IV. CONCLUSIONS

The present work successfully developed a formulation based on biodegradable polymeric materials based on PCL/gelatin containing encapsulated natural assets and maintained at a specific temperature, based on the Brazilian climate. Initially, the formulations were developed and evaluated over time based on the best preservative, that is, the formulation with the preservative that contributed to keeping the biomaterial more stable within the evaluation parameters. This system was the one that contained the preservative NE. The stability data analyzed showed that the synthesis with the preservative NE is promising for future applications and that its initial durability remained stable in 25 days of handling. The system presented stability at 25 °C. It is hoped that this research can contribute to the proposal of an alternative formulation to be used as a low toxicity and effective bioproduct.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Joab de Souza Arouche.
Methodology: Joab de Souza Arouche.
Investigation: Joab de Souza Arouche.
Discussion of results: Joab de Souza Arouche.
Writing – Original Draft: Joab de Souza Arouche.
Writing – Review and Editing: Joab de Souza Arouche.
Resources: Joab de Souza Arouche.
Supervision: Joab de Souza Arouche.
Approval of the final text: Joab de Souza Arouche.

VI. ACKNOWLEDGMENTS

Thanks to the Federal University of Amazonas - UFAM, to Nanopol - UFAM, where this study was carried out, and to FAPEAM for the financial support for the development of this research.

VII. REFERENCES

[1] L. R. Oroqui, M. Mori and P. Wongtschowsk, "Guia para a detetrminação da estabilidade de produtos químicos", Química nova, vol. 36, no. 2, pp. 340-347, Feb. 2013. https://doi.org/10.1590/S0100-40422013000200023

[2] A. R. Pires, A. C. K. Bierhalz and A. M. Moraes, "Biomateriais: tipos, aplicações e mercado", Química nova, vol. 38, no. 7, pp. 957-971, May, 2015. http://dx.doi.org/10.5935/0100-4042.20150094

[3] GHASEMISHAHRESTANI, Z. et al. "Tunable Synthesis of Gelatin Nanopartcles Employing Sophorolipid and Plant Extract, a Promising Drug Carrier". World Journal of Pharmacy and Pharmaceutical Sciences, vol. 4, no. 7, pp. 1365-1381, 2015.

[4] CALO, J. R. et. al. "Essential oils as antimicrobials in food systems–A review. Food Control", vol., no. 54, pp. 111-119, 2015. https://doi.org/10.1016/j.foodcont.2014.12.040

[5] KHALIL, Noha et. al. "Chemical composition and antimicrobial activity of essential oils from selected apiary fruits". Future Journal of Pharmaceutical Sciences, vol. 4, no. 1, pp. 88-92, 2018.

[6] CORREA, M. S. et. al. "Antimicrobial and antibiofilm activity of the essential oil from dried leaves of Eucalyptus staigeriana". Arquivos do Instituto Biológico, vol. no. 86, pp. 1-8, 2019. https://doi.org/10.1590/1808-1657000202018

[7] ASIAEI, E. O.; MOGHIMIPOUR, E.; FAKOOR, M. H.; "Evaluation of Antimicrobial Activity of Eucalyptus camaldulensis Essential Oil Against the Growth of Drug Resistant Bacteria". Jundishapur Journal of Natural Pharmaceutical Products, vol. 13, no. 4, pp. 1-7, 2018. DOI: 10.5812/jjnpp.65050

[8] MOSTAFA, A. A. et. al; "Antimicrobial activity of some plant extracts against bacterial strains causing food poisoning diseases. Saudi journal of biological sciences", vol. 25, no. 2, pp. 361-366, 2018. https://doi.org/10.1016/j.sjbs.2017.02.004

[9] HAMED, S. M. et al. "Biosynthesis of silver and silver chloride nanoparticles by Parachlorella kessleri SAG 211-11 and evaluation of its nematicidal potential against the root-knot nematode; Meloidogyne incognita". Australian Journal of Basic and Applied Sciences, vol. 10, no. 18, pp. 354-364, 2016.

[10] FARIAS, P. K. S. et. al. "Antioxidant activity of essential oils from condiment plant sand their effect on lactic cultures and pathogenic bacteria". Rural Science, vol. 49, no. 2, pp. 1-12, 2019. https://doi.org/10.1590/0103-8478cr20180140

[11] WILLIAMS, M. A.; WYNER, S. N. "Antimicrobial Resistance: Facing the Rise of a Global Threat". AJPH BOOK & MEDIA., vol. 109, no. 4, pp. 246, 2019. Doi: 10.2105 / AJPH.2019.304981

[12] ESFANJANI, A. F.; JAFARI, S. M. "Nanoencapsulation of Phenolic Compounds and Antioxidants". Nanoencapsulation of Food Bioactive Ingredients, Afshin Faridi Esfanjani1 and Seid Mahdi Jafari. vol. 1, no. 1, p. 63-101, 2017. DOI:10.1016/B978-0-12-809740-3.00002-7

[13] DE JONG, W. H.; BORM, P. J. "Drug delivery and nanoparticles: applications and hazards". Int J Nanomedicine, vol. 3, no. 2, pp. 133-49, 2008. Doi: 10.2147 / ijn.s596

[14] KALEPU, S.; NEKKANTI, V. "Insoluble drug delivery strategies: Review of recent advances and business prospects". Acta Pharm. Sin B, vol. 5, no. 5, pp. 442-453, 2015. https://doi.org/10.1016/j.apsb.2015.07.003

[15] ANVISA. Guia de estabilidade de produtos cosméticos - séries temáticas. 1a ed. Brasília: Agência Nacional de Vigilância Sanitária, 2004.

[16] ANVISA. Guia de estabilidade de produtos cosméticos - séries temáticas. La ed. Brasília: Agência Nacional de Vigilância Sanitária, (2012).

[17] ISAAC, V.L.B. et al. "Protocolo para ensaios físico-químicos de estabilidade de fitocosméticos". Rev. Ciências Farmacêuticas Básica Aplicada, vol. 29, no. 1, pp. 81-96, 2008.

[18] THOMPSON, JUDITH E. A prática farmacêutica na manipulação de medicamentos. Porto Alegre: Artmed, 2006.

[19] STAUB et al. "Determinação da segurança biológica do xampu de cetoconazol: teste de irritação ocular e avaliação do potencial de citotoxicidade in vitro". Brazilian Journal of Pharmaceutical Sciences. vol. 43, no. 2, pp. 301-307, 2007. https://doi.org/10.1590/S1516-93322007000200017

[20] XING, F. et al. "Nanoencapsulation of capsaicin by complex coacervation of gelatin, acacia, and tannins". Journal of Applied Polymer Science, vol. 96, no. 6, pp. 2225–2229, 2005. ttps://doi.org/10.1002/app.21698

[21] Apolinário, A. C. et al. Polimerossomos Versus Lipossomos: A Evolução Da "bala Mágica". Química Nova, vol. 40, no. 7, pp. 810–817, 2017. https://doi.org/10.21577/0100-4042.20170054

[22] SOUZA, P. M. S. et al. "Desenvolvimento de nanocápsulas de poli-ε-caprolactona contendo o herbicida atrazina". Química Nova, vol. 35, no. 1, pp. 132–137, 2012. https://doi.org/10.1590/S0100-40422012000100024

[23] BRITO. G. F. et al. "Biopolímeros, Polímeros Biodegradáveis e Polímeros Verdes". Revista Eletrônica de Materiais e Processos, vol. 6, no.2, pp. 127-139, 2011.

[24] HIGA, L. O. S. et al. "Evaluation of Rhipicephalus (Boophilus) microplus (Acari: Ixodidae) resistance to different acaricide formulations using samples from Brazilian properties". Braz. J. Vet. Parasitol., Jaboticabal, vol. 25, no. 2, pp. 163-171, 2016. Doi: http://dx.doi.org/10.1590/S1984-29612016026