

COMPLIANCE STUDY OF HAZARD ANALYSIS AND CRITICAL CONTROL POINT SYSTEM

Tatiana Sousa Pinto¹, Brigida Monica Faria^{1,2}, Luis Paulo Reis^{2,3}, Henrique Lopes Cardoso^{2,3}
and Tiago Santos^{2,3}

¹*Escola Superior de Saúde – Instituto Politécnico do Porto (ESS-P.Porto), Porto, Portugal*

²*Laboratory of Artificial Intelligence and Computer Science (LIACC), University of Porto, Porto, Portugal*

³*Faculty of Engineering, University of Porto (FEUP), Porto, Portugal*

ABSTRACT

Hazard Analysis and Critical Control Point (HACCP) system is based on a preventive methodology to avoid potential hazards that can cause harm and to ensure that unsafe food is not made available to consumers. This system is recognized by the Economic and Food Safety Authority, a criminal police responsible for food safety and economic inspection in Portugal. Every day, Economic and Food Safety Authority generates a large and complex volume of data from inspections and complaints, also in its classification, registration and in monitoring until the end of the process analysis. This study focuses on the reported entities that are related to non-compliance with HACCP, and tries to understand the most common infractions. Results show values between 30% and 37% related to non-compliance to HACCP. As main conclusions, from 2014 to 2018, the number of these infractions maintained the same level and it will be important to understand if the relationship between these problems are related to legislation understanding or application.

KEYWORDS

Public Health, Food Safety, Data Analysis Applications, HACCP

1. INTRODUCTION

Nowadays, modern society is more concerned and alert about the dangers, lack of security, hygiene, control of systems and services that can affect the consumer. Hazard Analysis and Critical Control Point (HACCP) system is based on a preventive methodology, in order to avoid potential hazards that can cause harm to consumers by eliminating or reducing problems, and to ensure that unsafe food is not made available to the consumer (Wallace and Mortimore, 2016). HACCP system is based on the application of technical and scientific principles in the production and handling of food aspects, from the beginning of their production to the final disposal to the consumer (Lawley, R. et al., 2013).

In order to prevent, eliminate or only reduce the hazards that may contaminate the food during its production process and subsequent distribution, requirements which, once contemplated, have led to the effective application of the HACCP system should be taken into account ('The food safety hazard guidebook', 2013).

The prerequisites control the hazards associated with the surrounding environment in the food production process, while the HACCP system controls the hazards associated with the production process(Stankovic, 2015): Structures and Equipment; Sanitation Plan; Pest Control; Water supply; Collection of waste; Materials in contact with food; Personal Hygiene; Training.

Economic and Food Safety Authority is a criminal police body specialized in the areas of food security and economic inspection in Portugal (ASAE, 2018b) (Mil-Homens, 2007). It is responsible for the evaluation and communication of risks in the food chain, as well as for legal compliance with economic activities in the food and non-food sectors (ASAE, 2018c). The Economic and Food Safety Authority 's mission is to supervise and prevent compliance with the regulatory legislation for the exercise of economic activities in the food and non-food sectors, as well as the evaluation and communication of risks in the food chain (ASAE, 2018c). Economic and Food Safety Authority has followed, with special attention, the growing of technological innovation and today is created the need to have capacity to control and store the data produced in the daily tasks of all its stakeholders, with the aim of improving public health and automation of inspector's tasks.

Economic and Food Safety Authority generates a large and complex volume of data in the various inspections it performs, in the reception of complaints, in its classification, in its registration and in the monitoring until the end of the process. This factor enhances the use of specialized storage tools capable of transforming data into information that positively influences decision making on inspection days (Lele, 2019). According to this expectation and with the predominant characteristic of the large volume of data, the techniques and tools for the analysis of these arise as a response, more concretely, through Data Analysis (Provost and Fawcett, 2013). To achieve formal declarations, the organization intends to use Data Analysis techniques, since it recognizes its capacity to respond to large amount of data (IRMA, 2016). Moreover, Economic and Food Safety Authority is the only organization in the world that brings economic security together with food security. Another aspect about Economic and Food Safety Authority it is the national connection body with its counterparts at European and international level (ASAE, 2018a). In fact, Economic and Food Safety Authority cooperates with other similar organizations around the world (ASAE, 2018a). As a criminal police body receives tens of thousands of complaints per year that convert into a lot of data. It is important to analyze the data available in the datasets and evaluate how can be an improvement in public health. Economic and Food Safety Authority generates a large and complex volume of data in the various inspections, in the reception of information requests and complaints, in its classification, in its registration and in the monitoring until the end of the process. This factor enhances the use of specialized storage tools capable of transforming data into information that positively influences decision making on inspection days (Regulamento (CE) n.º 852, 2004). According to this expectation and with the predominant characteristic of the large volume of data, the techniques and tools for the data analysis are related with Knowledge Discovery in Database and Data Analysis (Han, Kamber and Pei, 2012). To understand from the many data provided by Economic and Food Safety Authority those that are related to non-compliance with the HACCP system and to analyze its recurrence is the motivational factor of this study.

The remaining of the paper is structured as follows: section 2 presents the complaints flow, infractions classification and methodology used to understand and analyze data. Section 3 presents the results and discussion of the obtained result. The last part of this paper presents the major conclusions, and some directions for future work.

2. METHODOLOGY

This section presents the methods used to acquire more information about complaints that reach Economic and Food Safety Authority and which are related with non-compliance of HACCP system. First, a complaint is a formal complaint in order to expose an irregularity/non-compliance of economic group with the legislation (DL194/, 2012). All reports of the datasets were made online by filling in the form available on the Economic and Food Safety Authority website or sent by email. Then a decision circuit is made by inspectors to classify complaints in many aspects, such as competence (as can be seen in Figure 1) or type of complaints.

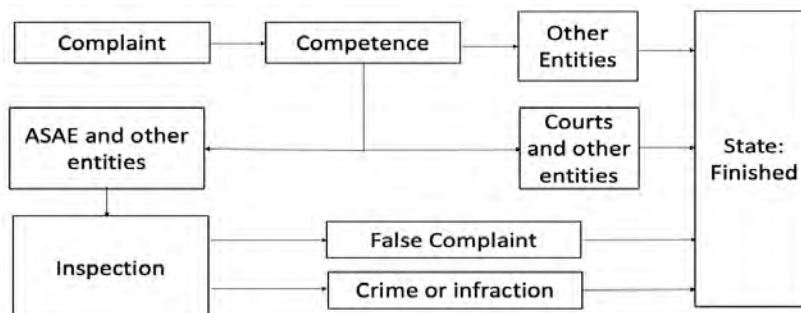
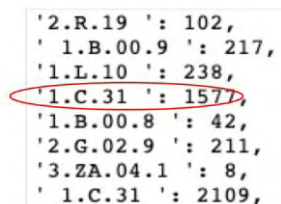


Figure 1. Complaint competence in decision circuit

In terms of competence, there are many complaints that should be treated by other entities such as courts or other institutional entities. In this work, only complaints that are from Economic and Food Safety Authority competence will be analyzed. The competence complaints have code assigned in order to classify the kind of infraction is presented.

Figure 2 presents an example of classification codes and frequencies in the year 2014. In this example the code 1.C.31 has digits and letters that represent different information: 1 represents “Segal”, an abbreviation for “Segurança Alimentar” that means food security; C represents hygiene and food safety; 31 represents “Failure to comply with general and specific hygiene requirements”.



'2.R.19'	': 102,
'1.B.00.9'	': 217,
'1.L.10'	': 238,
'1.C.31'	': 1577,
'1.B.00.8'	': 42,
'2.G.02.9'	': 211,
'3.ZA.04.1'	': 8,
'1.C.31'	': 2109,

Figure 2. Frequencies and infractions codes

In this example, the most frequent infraction in 2014 was a HACCP compliance. Also, in this example there are different formats related with the deep of detail in the code with two or three points dividing the fields of information.

2.1 Data and Ethical Issues

The available data range from 2014 until 2018 and is divided into two datasets: data on complaints and information requests and data related to type of complaints associated with infractions. Statistical analysis of the complaints file and information requests, a comparison was made between the number of complaints and information requests, including complaints and pending information requests per year. This statistical analysis was made using Python libraries, Pandas and NumPy (McKinney, 2018). The first descriptive analysis was applied to five excel files related to the years of 2014 to 2018 with approximately 30MB each with a total of 120661 entries of complaints and information requests. This analysis was simple and with no setbacks because there was not missing values or incorrect data in any of the files. In 2018 the available data was only until October. Concerning the file related to denounced entities a 670mb file containing 2882629 reported entities was analyzed. The file's variables include:

- RNPC - National Registry of Legal Entities;
- Relation - If belongs to a head office or to an establishment;
- NIPC / NIF - Identification Number of Collective Person/ Tax Identification Number;
- Type - Type of Entity;
- Name - Entity Name;
- Postal Code;
- Community/Region;
- Address;
- District / County / Town;
- Legal Nature – RNPC;
- Main CAE – Main Code of Economic Activity;
- Secondary CAE – Secondary Code of Economic Activity.

The descriptive analysis about these variables focus on the type of entity most reported/denounced and in which districts most of the complaints were made. The Economic and Food Safety Authority has an epigraph file containing the meaning of each letter or code number associated with each infraction to make possible the analyses of the dataset related to type of complaints associated with infractions. Analyzing this file, it was verified that infractions beginning with the number "1" are related to HACCP. Each year between 2014 and 2018 corresponds an excel file with each report made, including the infringement code assigned to each report. After importing each excel file, per year, into the Jupyter Notebook (McKinney, 2018) an initial observation was made to identify the columns of interest. Given the amendment to the General Data Protection Regulation,

the directive and legislation on data protection have been replaced and all those involved in these works have signed terms in accordance with the new legislation to ensure the right of secrecy of both the complainant and of the complaint.

2.2 Data Analysis Methods

Descriptive data analysis was made using Pandas, a software library written for the Python programming language for manipulation and data analysis (DataCamp, 2001). It was also used NumPy that provides the data structures, algorithms, and it is needed for most scientific applications involving numerical data in Python (McKinney, 2018).

2.2.1 KDD Methodology

The data volume, generated and stored in the course of any activity, exceeds the capacity of human analysis and makes it impossible to extract knowledge from the same data, without resorting to a system that automates this process (Fayyad, Piatetsky-Shapiro and Smyth, 1996). This context justifies the existence of the research area of Knowledge Discovery in Databases (KDD), generally defined as "the non-trivial process of identifying valid and potentially useful patterns, perceptible from the data" (Fayyad, Piatetsky-Shapiro and Smyth, 1996), in which the principles associated to it are influenced by the areas of Artificial Intelligence, Automatic Learning, Pattern Recognition, Statistics, Database, Information Systems, among others (Maimon and Rokach, 2010). The method used to search the patterns in the data is called Data Mining and is considered the core step of the process of Knowledge Discovery in Databases, which is developed in several phases, such as: selection, preprocessing, transformation, data mining, interpretation and evaluation (Witten, Frank and Hall, 2005). These patterns may or may not represent useful knowledge, being one of the phases that normally requires the participation of the user (Fayyad, Piatetsky-Shapiro and Smyth, 1996). Data Analysis phase represents 20% in the KDD process and is better supported by software than the other phases. All other phases, from the selection of data to the interpretation of the patterns found, are more a matter of "art" than a routine that can be automated (Jothi, Rashid and Husain, 2015).

2.2.2 Data Pre-Processing

The data has some problems that were tackled using pre-processing. These problems were related with missing/incomplete values as can be analyzed in Table 1.

Table 1. Variables vs Percentage Invalid/Null data information

Variables	Invalid/Null (%)	Variables	Invalid/Null (%)
RNPC	< 0,01	Address	0,53
Relation	< 0,01	District	2,40
NIPC/NIF	< 0,01	County	6,05
Type	0,01	Town	90,74
Name	< 0,01	Legal Nature RNPC	0,60
Postal Code	73,23	Main CAE	41,20
Community/Region	0,43	Secondary CAE	89,73

In terms of missing values were identified the following problems:

- 3-digit suffix of the postal code missing;
- District / County / Town" field the information for "County" and "Town" to be filled.
- In terms of incorrect and/or outdated values:
- Town as region;
- Incomplete / ambiguous values;
- Data corruption;
- CAEs missing values.
- There were also inappropriate separators and delimiters:
- Mixture of separation by spaces, by "," and by ";".

The majority of data considered invalid or null is in the “CAEs”, the “Postal Code” and the “Town”. The missing data were filled matching it with online data bases with valid information. In case of unavailable information from online datasets, the lines with missing values were excluded from analysis.

From another dataset the column with the infractions codes is called "Infracções" as shown in Figure 3. In this column we have information about the infraction code, the type of infraction and a description that explains the meaning of the infraction code. All the following examples are from 2014 dataset.

Tipo	Ofícios para Denunciante	Produtos	Infracções	Fiscalizações Relacionadas	Processos Relacionados
E-MAIL	[S/290 /14/SC] Confirmação (competência da ASAE...	NaN	2.B.1 , CO, Exercício de actividades de anima...	NaN	NaN
E-MAIL	[S/303 /14/SC] Recepção - DIP	NaN	2.G.02.14 , CO, Violação das regras do contrat.	NaN	NUI/CO /001589 /15.5.EAPRT

Figure 3. Infractions column from 2014 dataset

2.2.3 Transformed Data

As the goal is to study the compliance analysis of Economic and Food Safety Authority’s Hazard Analysis and Critical Control Point system the focus of study will be in the column selected in Figure 3.

Each complaint is associated with an NID, which is a code assigned to identify each complaint. In each complaint, more than one infraction can be identified, as demonstrated in Figure 4, and it is necessary to separate the different infractions within each complaint.

NID	Infracções
0 E/8/14/SC	2.B.1 , CO, Exercício de actividades de animação turística sem registo\n 2.B.2 , CO, Exercício ...

Figure 4. NID and Infractions columns

As each infraction of the same complaint is separated by paragraph, the split is done by the paragraph as can be seen in Figure 5 the output.

NID	Infracções
0 E/8/14/SC	2.B.1 , CO, Exercício de actividades de animação turística sem registo
1 E/8/14/SC	2.B.2 , CO, Exercício de actividades não reconhecidas como turismo de natureza na Rede Nacional ...

Figure 5. Output from split by paragraph

Since the infractions are separated it is necessary to split each code so that we can proceed to the count.

Since the infraction code, the type of infraction, and the description of the infraction are separated by commas in the column “Infracções”, it is used the comma split. This split creates a new column (Figure 6) named ‘Codes’ that contains the infraction code that represents the infraction in each complaint. This column can be added to the original dataset for future studies in the database.

Using unique, one of the commands of the Pandas library, it turns out that split created a space at the end of each code. This space was removed so that it is possible to count the different infraction codes. In order to understand the meaning of each code it is necessary to analyze the file of epigraphs made available. By analyzing this file, it is verified that all the complaints started by the number ‘1’ are related to the HACCP system. In order to know how many complaints are related to the HACCP system a new split was made. The output of this new split gives a 26083 total number of rows in the dataset for the year 2014. In order to know the number of 9726 complaints with the infringement code related to the HACCP system was applied a count of these entries. The same procedure described for the year 2014 was applied for the years 2015, 2016 and 2018, trying to verify, through data analysis by descriptive statistics if there is any pattern over the years.

3. RESULTS

Data Analysis is a superset of Data Mining that involves extracting, cleaning, transforming, modeling and visualization of data with an intention to uncover meaningful and useful information that can help in deriving conclusion and take decisions (Abzalov, 2016). This Section presents the results obtained by analyzing the complaints and information requests and the complaints related to HACCP system from 2014 to 2018.

3.1 Descriptive Analysis about Infractions related to Complaints and Information Requests

From a total of 120661 entries between 2014 and 2018, 80358 of them are complaint and 40303 are information requests (see Figure 6).

“Complaints” represent the solved and archived complaints. “C_Hold” represent complaints on hold to be solved. “Information Requests” is related to information that the sender needs to be enlightened. “IR_Hold” represent information requests that have not yet been answered. The number of complaints has not decreased over the years. Complaints on hold (C_Hold) have not yet been solved. 2018 is the year with more “C_Hold” which is expected because there was no time for response yet. From a total of 120661 entries of denounced entities nearly 100% represent Economic Agents. To try to solve missing data related to entities addresses was used information in georeferenced websites. As a result of this cross-referencing of information 97,33% of data became complete.

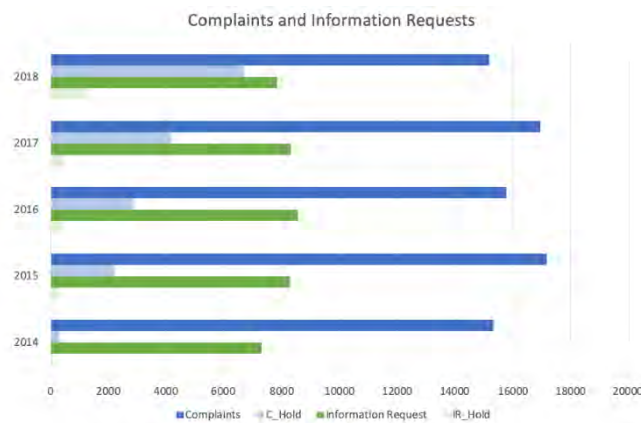


Figure 6. Complaints and Information Requests frequencies

3.2 Descriptive analysis about infractions related to HACCP system

The number of rows of all original datasets increased after the split (Fig. 6). Once each line represents a complaint and to each complaint is associated an infraction code, it means that there are complaints with more than one infraction code. In Figure 7 it is possible to compare the number of original rows with the number of rows created after the split from 2014 to 2018.

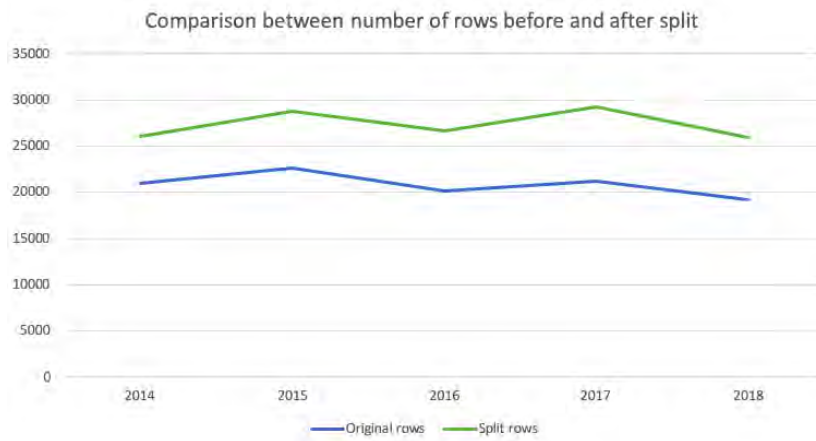


Figure 7. Comparison between the number of rows before and after split

As can be seen (Figure 7), in all datasets from 2014 to 2018 there are complaints with more than one infraction committed. As the objective is the analysis of compliance with the HACCP system, it is important to analyze the infractions committed related to this system in each year under study (Figure 8).

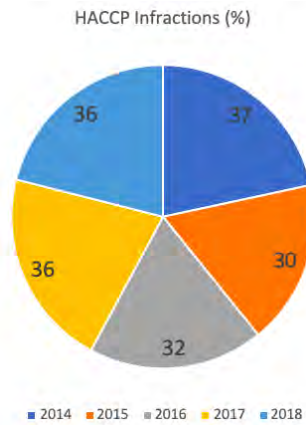


Figure 8. Percentage of infractions related to non-compliance with HACCP per year

The percentages of non-compliance of the HACCP system were rounded to the units. As can be seen, the variation between the interval of years was not accentuated. From 2014 to 2015 the compliance improved by 7%, however in the remaining years this percentage of compliance was disappearing.

4. CONCLUSIONS AND FUTURE WORK

Food and Economic Security Authority receives more complaints than information requests. It is important to try to understand where law enforcement is failing so that there is not a significant decrease in complaints over the years. Incomplete, incorrect and omitted data relating to the entities complained may be justified since the complainant does not always know the correct and complete data of the entity it intends to report. From the point of view of compliance with the HACCP system, values between 30% and 37% are relatively high since the Food and Economic Security Authority is responsible for food and economic security and not only for food security and hygiene. Although, as mentioned above, there was an improvement between the first two years under study, followed by worse results in the following years, the values related to the non-compliance rate did not have marked differences.

As future work, it would be interesting to evaluate the density of distribution of the complaints by regions. Once this has been done, and in an attempt to improve public health, it is necessary to associate them to the HACCP system related complain. Then try to understand what the most common complaint is and see if the problem is related to understanding the legislation or if it is related to the difficulty of complying with the legislation because it is too restrictive.

ACKNOWLEDGEMENT

The authors would like to thank the IA.SAE – “Inteligência Artificial na Segurança Alimentar e Económica” project, funded by the FCT/MCTES through national funds (PIDDAC), as part of the “Programa Iniciativa Nacional Competências Digitais” e.2030 – INCoDe.2030, enrolled in the National Reform Plan.

REFERENCES

- Abzalov, M., 2016. *Exploratory data analysis*, in Modern Approaches in Solid Earth Sciences. doi: 10.1007/978-3-319-39264-6_15.
- ASAE, 2018a. ASAE-Cooperação Internacional Bilateral, <https://www.asae.gov.pt/cooperacao/internacional/bilateral/medio-oriente.aspx>.
- ASAE, 2018b. Como Actua a ASAE. Available at: <https://www.asae.gov.pt/inspecao-fiscalizacao/como-atua-a-asae.aspx> (Accessed: 30 May 2019).
- ASAE, 2018c. Missão, Visão e Valores. Available at: <https://www.asae.gov.pt/asae20/missao-visao-e-valores.aspx> (Accessed: 30 May 2019).
- DataCamp, 2001. *Python for Data Science*, Cheatsheet. doi: 10.1097/00007890-200105270-00005.
- DL194/ (2012) ‘Decreto-Lei no 194/2012 ASAE’, Diário da República 1a série, pp. 4641 - 4646
- Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996. *The KDD process for extracting useful knowledge from volumes of data*. Communications of the ACM, 39(11), pp.27-34.
- Han, J., Kamber, M. and Pei, J., 2012. *Data Mining: Concepts and Techniques*. doi: 10.1016/C2009-0-61819-5.
- IRMA, I. R. M. A., 2016. *Big Data: Concepts, Methodologies, Tools, and Applications*. Information Resources Management Association ISBN-13: 978-1466698406
- Jothi, N., Rashid, N. A. and Husain, W., 2015. *Data Mining in Healthcare - A Review*, in Procedia Computer Science. doi: 10.1016/j.procs.2015.12.145.
- Lawley, R.; Curtis, L.; Davis, J., 2013. *The food safety hazard guidebook*, Choice Reviews Online. doi: 10.5860/choice.46-3282.
- Lele, A., 2019. *Big data*, in Smart Innovation, Systems and Technologies. doi: 10.1007/978-981-13-3384-2_9.
- Maimon, O. and Rokach, L., 2010. *Introduction to Knowledge Discovery and Data Mining*, in Data Mining and Knowledge Discovery Handbook. doi: 10.1007/978-0-387-09823-4_1.
- McKinney, W., 2018. *Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython*, Climate Change 2013 - The Physical Science Basis. doi: 10.1017/CBO9781107415324.004.
- Mil-Homens, S., 2007. *ASAE - HACCP*. Available at: <http://www.asae.gov.pt/pagina.aspx?back=1&codigono=54105579AAAAAAAAAAAAAAAAAAAA> (Accessed: 19 January 2019).
- Provost, F. and Fawcett, T., 2013. *Data Science and its Relationship to Big Data and Data-Driven Decision Making*. Big Data, 1(1), pp.51-59.
- Regulamento (CE) n.o 852, 2004. Regulamento (UE) no 852/2004 do parlamento europeu e do conselho’, *Jornal Oficial da União Europeia*, 139, pp. 54.
- Stankovic, I., 2015. Codex Alimentarius, in *Encyclopedia of Food and Health*. doi: 10.1016/B978-0-12-384947-2.00180-X.
- Wallace, C. A. and Mortimore, S. E. (2016) ‘HACCP’, in Handbook of Hygiene Control in the Food Industry: Second Edition. doi: 10.1016/B978-0-08-100155-4.00003-0.
- Witten, I. H., Frank, E. and Hall, M. A., 2005. *Data Mining Practical Machine Learning Tools and Techniques*, Data Mining. doi: 0120884070, 9780120884070.