Methodology for the Analysis of Agricultural Data in the Mexican Context: Study Case of Marigold

Cristal Galindo Durán[1][a] and Mihaela Juganaru[2][b]

[1] *Escuela Militar de Ingeniería, Universidad del Ejército y Fuerza Aérea, Lomas de San Isidro, Naucalpan, Mexico*

[2] *Department ISI, Institut Henri Fayol, IMT - Mines de Saint Etienne, Saint Etienne, France*

Keywords: Data Collection, Research Methodology, Agricultural Production, Data Analysis, Knowledge Extraction, Data Visualization.

Abstract: Agricultural production data for multiple crops is available as open data; However, to discover information in the data it is necessary to consider methodologies, methods and tools that allow guiding the research work to specifically explore agricultural data. This article aims to propose an adaptation of the CRISP-DM and OSEMN methodologies to the agricultural context, which helps to study any crop. In addition, to apply the proposed methodology to the agricultural production of an endemic Mexican product that is the marigold flower, *Tagetes erecta*.

1 INTRODUCTION

Currently, various national and international government agencies generate and make available a large amount of data a lot of domains such as: health, transport, tourism, economy, environment, agriculture, etc., which are accessible (open access) by multiple organizations and researchers to manage them. The treatment of said data allows to discover information and the relationships that underlie them, allowing to answer research questions; as well as, check, verify and contrast facts on a specific issue.

It's easy to find raw data on the annual production of a product, agricultural or industrial, for a specific period. However, this type of raw data can be provided with a variety of variations: by area, company, mode of production, period of production, etc. Our reading and learning mode is changing and we often are trying to verify or to check facts.

Particularly in the agricultural domain, raw data can be found on the annual production of a specific crop; however, raw data can provide a significant amount of variation, depending on the organization that publishes it.

Such is the case of the Mexican government, specifically the Secretaría de Agricultura y Desarrollo Rural (SADER, 2023), which through the Servicio de Información Agroalimentaria y Pesquera (SIAP, 2023) makes available data on more than 300 different crops that they have been produced in the Mexican countryside since 1980.

Specifically, in this article, the data on the cultivation of Mexican marigold scientifically known as *Tagetes erecta* or also called *cempasúchil* or Day of the Dead flower for its use in this celebration are taken up (Vázquez, 2016). The interest in analyzing the production of this crop is because various sources (Páramo, 2017), (Zamarrón, 2021), (Luna, 2021) report a decrease in production displacing Mexico, despite being an endemic flower of this country and positioning countries such as China (75%), India (20%) and Peru (5%) of global production. The international inclination for this crop is that it is used not only in cultural matters, but also in the cosmetic, pharmaceutical and food industry as a coloring, flavoring (Méndez, 2021) among others.

This work aims to propose a methodology based on the CRISP-DM (Hotz, 2023b) and OSEMN(Hotz, 2023a) data science methodologies, that allows analyzing agricultural data in the Mexican context[1] and suggesting relationships between them to discover relevant information. This methodology can be applied to any other crop and considered for another

[a] https://orcid.org/0000-0002-2119-8947

[b] https://orcid.org/0000-0002-4329-3101

[1] Mexican context means having different production cycles and having more of one crop by land, the same plant or another, during a year

Durán, C. and Juganaru, M.
Methodology for the Analysis of Agricultural Data in the Mexican Context: Study Case of Marigold.
DOI: 10.5220/0012257800003598
In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - Volume 1: KDIR*, pages 453-459
ISBN: 978-989-758-671-2; ISSN: 2184-3228

data source that shares similar metadata. For the application of the methodology, the historical records of more than 40 years of the cultivation of the Mexican marigold flower are considered as a case study to verify and contrast their production data, as well as their relationships.

The article is organized as follows: Section 2 presents the proposed methodology and its explanation. Section 3 shows the application of the methodology to the case study of the cultivation of the Mexican marigold flower. Section 4 presents the results obtained from the case study. Finally, section 5 presents the conclusions and some future works.

2 PROPOSED METHODOLOGY

The proposed methodology for the analysis of agricultural production data is based on the CRISP-DM(Hotz, 2023b) and OSEMN(Hotz, 2023a) data science methodologies, which consists of 6 phases.

1. Definition of goal
2. Obtaining data
3. Cleansing Data,
4. Data Exploration
5. Data modeling
6. Interpretation of results.

The Figure 1 shows the diagram that represents the different phases and activities of the proposed methodology, where the phases are shown in solid line boxes, while the activities are in dashed boxes.

The development of the phases and their activities of the proposed methodology are detailed below.

2.1 Definition of Goal

At this phase, it is necessary to determine the type of crop to study, set the objective and scope of the study in question; as well as, set the research questions to be answered.

2.2 Obtaining Data

It consists of collecting data of interest from different sources, such as databases or Internet repositories of organizations (secretariats, ministries, agencies, etc.) dedicated to agriculture. In addition, by understanding the data dictionary provided by the different organizations dedicated to agriculture, said dictionary will provide information about the fields that make up the datasets; as well as their data types. It is also recommended to choose the data manipulation language.

2.3 Cleansing Data

This phase consists of identifying incomplete data (Vázquez and Juganaru-Mathieu, 2014), removing outliers and considering only data with the same units of measurement (tons, gross, plants, bundles, kilograms) . The above in order to guarantee their quality. In addition to identifying the characteristics of importance for the analysis, to later select or eliminate them.

2.4 Data Exploration

For this phase, different statistical estimators are applied to the quantitative variables of interest, such as: measures of central tendency (mean, mode, median), position (quartiles, deciles or percentiles), dispersion (standard deviation, variance, range of variation), variance and correlation matrix.

2.5 Data Modeling

This phase is based on using association, regression, classification or grouping techniques between variables in order to find relevant information between the relationships and propose models that allow predicting the behavior of any variable of interest, such as: production volume/cycle agricultural, crop/irrigation mode, planted/harvested area, etc.

2.6 Interpretation of Results

It consists of giving meaning and explanation to the results obtained in the previous phases of data exploration and modeling. Data visualization is extremely important, the final presentation can be as time series, overlay graphs, boxplot, etc. If the interpretation is not convincing, it is necessary to return to the previous stages for revision.

3 STUDY CASE

As already mentioned in the introduction for the application of the proposed methodology, the data sets of the crops produced by the Mexican countryside and disclosed by the Servicio de Información Agroalimentaria y Pesquera (SIAP, 2023) are taken as a case study. These data sets have more than 300 different crops between the years 1980 to 1941.

For the Definition Goal phase, the cultivation of the Mexican marigold flower was taken into account, a flower native to Mexico and important for its cultural weight in the traditions of the Day of

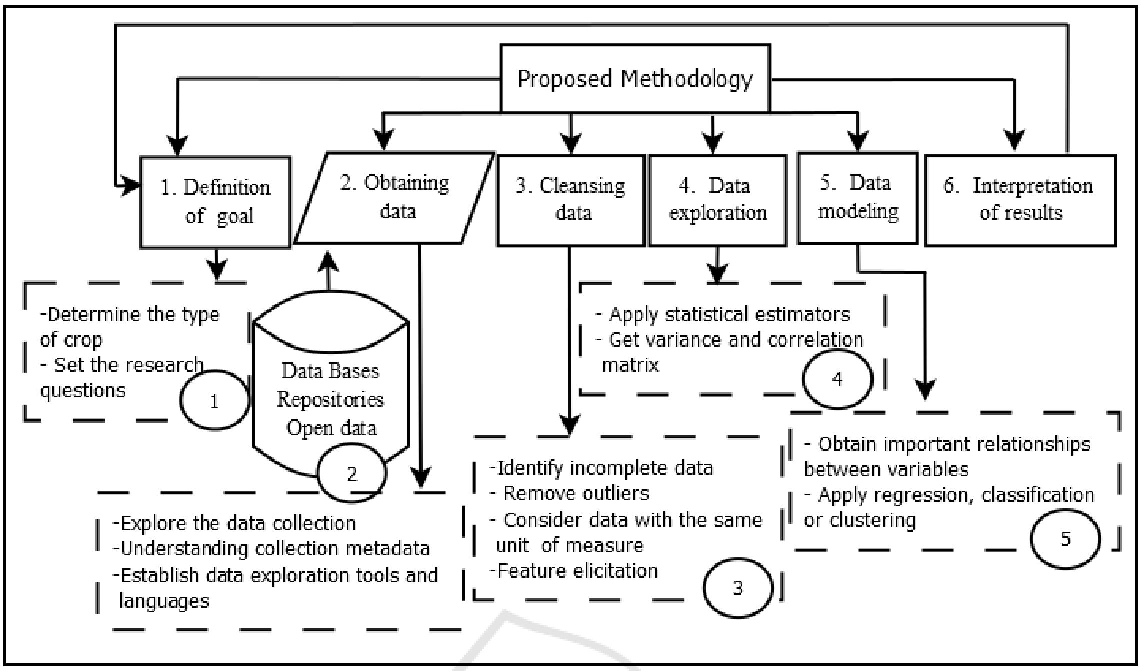
Proposed Methodology
1. Definition of goal
2. Obtaining data
3. Cleansing data
4. Data exploration
5. Data modeling
6. Interpretation of results
-Determine the type of crop
- Set the research questions
1
Data Bases Repositories Open data
2
-Explore the data collection
-Understanding collection metadata
-Establish data exploration tools and languages
- Apply statistical estimators
- Get variance and correlation matrix
4
-Identify incomplete data
- Remove outliers
- Consider data with the same unit of measure
-Feature elicitation
3
- Obtain important relationships between variables
- Apply regression, classification or clustering
5

Figure 1: Block diagram of the phases and activities of the proposed methodology.

the Dead (Mandujano, 2020). Likewise, various research questions could be formulated, such as: What is the historical production of said flower?, What is the minimum and maximum number of hectares planted/harvested?, What was the year in which more Mexican marigold?, What is your production forecast?

In the Data Collection phase, the 41 datasets offered by the SIAP were considered, available in `.csv` format, which were concentrated in a single file that had 1781 records and 18 fields (metadata), as: Year, IdState, Name state, IdCycle, Name cycle, IdModality, IdUnitMeasure, Name unit of measurement, IdCrop, Name crop, Planted, Harvested, Damaged, Volume of production, Yield, Price and Production value.

The most useful schema for data is a star schema, shows Figure 2. It is possible to see the foreign key integrated by four dimensions : $Time$, $State$, $Modality$ and $Cycle$. On the other hand, the main facts are: $Planted$, $Harvest$, $Damaged$ and $VolumeProduction$. The attribute $Yield$ can also be seen as a fact, but it's a calculate value by $VolumeProduction/Harvest$.

Keep data into a data mart allowed us to do grouping by dimensions the data and quick computing aggregate functions, mainly *sum*. The analyses about $Planted$ versus $Harvest$ and $Damaged$ and about $VolumeProduction$ are realized easily. However, applying the aggregate function *mean* about $Yield$ is non-sense, because it's a computed value. The mean of the yield production grouping by dimension ($Time$ or $Time$ and $Modality$ or $Cycle$) have to by computed as shown in the equation 1.

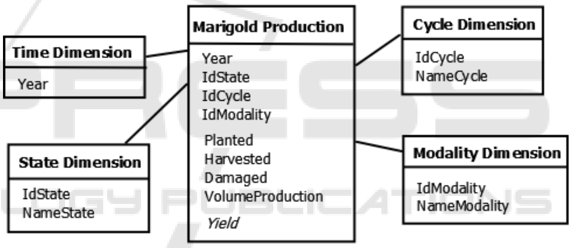

Figure 2: Star schema for Mexican marigold flower agricultural data.

$$\frac{\sum_{dimensions} VolumeProduction}{\sum_{dimensions} Harvest} \tag{1}$$

For the Cleansing Data phase, the units of measurement of 'gruesas', plants and bunch were discarded, taking into account only tons. Missing values were removed and outliers were ignored given the nature of the analysis.

In the Data Exploration phase, the metadata of interest were determined by selecting 10 of the 18 fields that make up the dataset. Table 1 shows the most important metadata chosen to answer the questions posed in phase 1. Once the metadata ware chosen, we proceeded to obtain the measures of central tendency, dispersion and minimum and maximum values of the variables of interest: hectares planted, harvested and damaged, tons of production (volume of production) and yield. In addition, the correlation matrix was obtained, showing the intensity and type of relationship of the variables.

By the other hand, for the Data Modeling phase, the following relationships were examined:

Volume of Production.

- Volume of production/ cycles. This relationship represents the total sum of tons of the crop produced in the cycles: spring-summer/autumn-winter per year.
- Volume of production/planted. It shows the relationship of the tons obtained compared to the hectares planted per year.
- Volume of production / states. It is the sum of the production volume of the crop by state.
- Volume of production/modality. Represents the sum of the tons obtained by type of water modality: irrigation/temporary.

Yield.

- Yield/cycle. It presents the total tons per hectare obtained in both cycles: Spring-Summer/Autumn-Winter per year.
- Yield/modality. It is the sum of the production volume divided by the sum of the harvested area, grouped by water type and year.
- Yield/state. It exposes the relationship between the sum of tons per hectares that each state.

Planted.

- Planted/damaged. It shows the relation of planted hectares compared to the damaged ones per year.
- Planted/harvested. It exposes the correspondence between the hectares planted and harvested per year.

In addition, the correlation matrix was obtained where significant or very strong relationships are identified between the harvested and planted variables; as well as between plantations and production volume; harvested with the volume of production in all three cases it is a direct relationship.

Subsequently, evaluate the correlation between the variables, a simple linear regression was carried out between the variables of planted-harvested and planted production volume, with the intention of specifying the development of the methodology.

4 RESULTS

Regarding the case study of the Mexican marigold flower, the results of the data exploration and data modeling phase are presented in the following subsections.

4.1 Data Exploration

The results obtained in this phase were based on obtaining the measures of central tendency, the measures of central tendency, standard deviation, and minimum and maximum values of the variables of interest were obtained: planted, harvested, damaged, and production volume.

The Table 2 presents a concentrate of some simple statistical measures.

From the Table 2 it can be noted that both the sum and the mean of the planted areas in hectares is greater compared to the harvested and damaged areas. This is also reflected in the frequency graph shown in the figure 3 that contrasts the historical behavior of the variables through 41 years. We can also see that the Planted area are varying from a state, a year, a modality and a cycle.

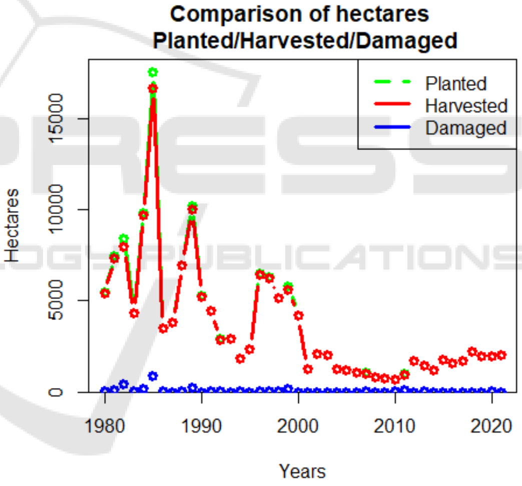

Figure 3: Comparative graph of the planted, harvested, damaged areas of the Mexican marigold crop.

From the analysis of the Figure 3, it can be observed that the historical maximum of the planted and harvested surfaces was in the year 1985 with a total of 17,577 and 16,673 hectares respectively. Likewise, in 2010 it was its historical minimum with a total of 673.60 hectares planted and 657.60 harvested. Therefore, in general, a considerable decrease can be seen in the production of Mexican marigold, which has been exorbitant from 1980 to 2021 (despite the fact that a slight recovery is seen in the 2010 period). In this way, it can be ensured that the production of this crop in Mexico has been reduced in recent 30 years.

Table 1: Metadata considered for the analysis of Mexican marigold production.

Name of column	Type	Signification
Year	Numeric	Year of information between 1980 and 2021
IdState	Numeric	Code that identifies the federative entity or state of production
Name State	Text	Official name of the federal entity or state
IdCycle	Numeric	Code that identifies the agricultural cycle: Spring-Summer (1), Autumn-Winter (2)
IdModality	Numeric	Name Hydro Modality: irrigation (1), temporary (2)
Planted	Decimal	Surface in hectares planted with the crop
Harvested	Decimal	Area in hectares harvested from the crop
Damaged	Decimal	Area in hectares affected by the crop
Volume Production	Decimal	Production volume of the harvested area whose unit of measure is tons
Yield	Decimal	The unit of measure is tons per hectare. $Yield = VolumeProduction/Planted$

Table 2: Statistical measures for the variables: harvest, planted, damaged; the mean, standard deviation, min and max as computed by year, by state, by modality and by cycle.

Variable	Sum	Mean	SD	Min	Max
Harvest	153503.5	95.5	389.4	0	6571
Planted	156097.1	97.0	399.2	0.5	7162
Damage	2593.6	1.6	19.6	0	591
Volume of Production	1779721	1106.8	5594.1	0	121548

4.2 Explicit Results

Regarding the relationships between the different variables, the following relevant information could be found:

Production Volume/Cycles. The number of tons produced is now higher in the Spring-Summer agricultural production cycle than in the Autumn-Winter cycle, as well as the number of planted surfaces, see Figure 4. We also can see a significant change over the time : more less planted in cycle Autumn-Winter since 1990.

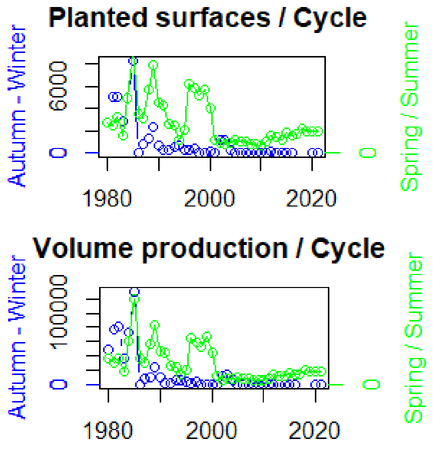

Figure 4: Analysis by Cycle during the years regarding the planted surfaces, volume of production and yield.

Production Volume/States. The 5 states that have had the highest production over 41 years are: Sinaloa, Guanajuato, Puebla, Michoacán and Sonora. The five states with the lowest production are: Zacatecas, Chiapas, Baja California, Baja California Sur and Nuevo León. That is, marigold flower production is more favorable in dry, semi-dry and subhumid climates; On the contrary, this crop is not suitable in hot, humid or very dry climates.

Production Volume/Modality. A greater quantity of Mexican marigold flower is produced by irrigation than by rain.

Yield/Cycle. There is no progress in yield through the years [2] and there are more variations in yield in the Autumn-Winter cycle, see figure 5.

Yield/Modality. A higher yield is sometimes obtained in the irrigation modality than in the temporary one, see Figure 6.

Yield/State. The 3 states with the highest yield have been Sinaloa with $16.55t/ha$, Guanajuato with $14.47t/ha$ and Puebla with $13.20t/ha$ during the period 1980-2021. The best yield production was obtained in 2021 in Morelos, Durando and Guerero. Since 2012 and 2013 some states had stopped their production, even they had an excellent yield.

Planted/Damaged. There is a greater number of hectares planted than damaged and some states such as: Baja California Sur and Nuevo León that in the year 2000 tried to plant this crop and it was damaged, so they did not try replanting again. As we can see in

[2] The variations of the yield of marigold over the time indicate that this type of crop doesn't benefit of the progress of any agricultural techniques, like better seeds, changes of technologies, etc.

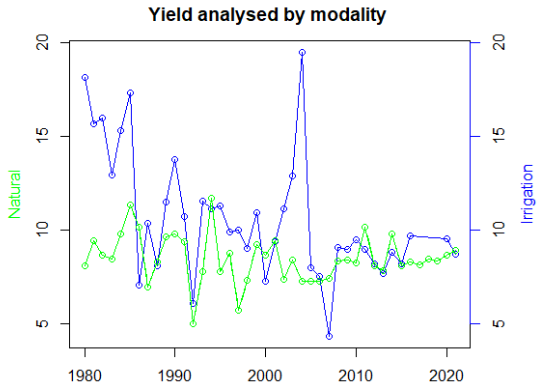

Figure 5: Analysis by Cycle during the years regarding the planted surfaces, volume of production and yield.

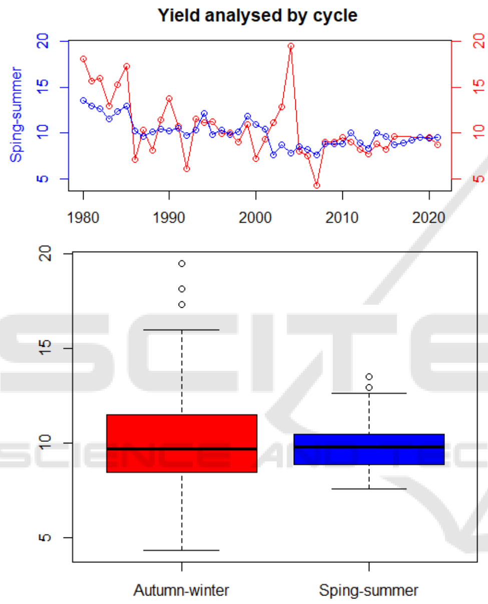

Figure 6: Analysis by Cycle during the years regarding the planted surfaces, volume of production and yield.

Figure 7 the risk of a damage is greater in the cycle Autumn-Winter; the risk is very low if grouping by modality and taking the modality irrigation.

Planted/Harvested. There is a direct relationship between the number of hectares planted and those that are harvested.

On the other hand, considering the production volume, planted, harvested and casualty variables are taken into account, making the correlation matrix and applying the Spearman coefficient to find out their relationship. From the data obtained, it can be seen that there is a significant or very strong relationship between the harvested and planted variables; as well as between planted and production volume and harvested with production volume, with correlation coefficients of 0.99, 0.96 and 0.97.

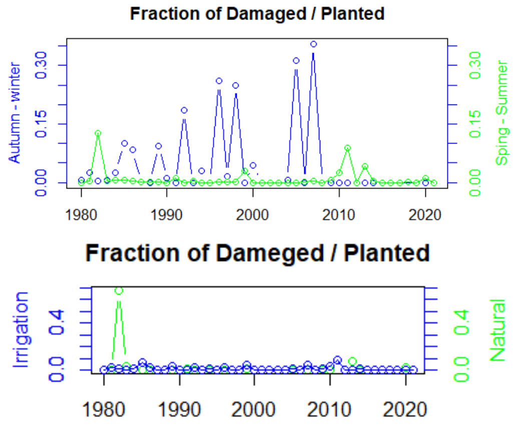

Figure 7: The proportion of damaged surfaces against the planted one grouping by cycle, respectively by modality.

After the variables that have a significant relationship were identified, their linear regression between the planted and harvested variable was obtained; as well as planted and volume of production. In both models, the planted variable is the independent variable (X) and the dependent variables are the harvested area and the production volume (Y), being able to obtain the 2 equation and the 3 equation for the planted-harvested relationship and production volume respectively.

$$Y = 0.849774 + 0.974631X \tag{2}$$

$$Y = -216.849774 + 13.63151X \tag{3}$$

With the previous models, it is possible to predict the harvested area and production volume based on the planted area.

5 CONCLUSIONS

This article presents the proposal of a methodology based on Data Science, specifically on the CRISP-DM and OSEMN methodologies that consists of 6 phases: 1) Definition of objectives, 2) Data collection, 3) Data cleaning, 4) Data exploration , 5) Data modeling and 6) Interpretation of results, adapting it for the analysis of agricultural data in the Mexican context.

A series of relationships between several variables is also proposed that allow obtaining additional information about the cultivation of the marigold flower, providing information that can be used by the government to implement policies that strengthen the production of this crop, promoting planting in states that have warm, dry and semi-dry climates to promote projects that benefit their communities.

To validate the methodology, a case study focused on a particular crop is considered to validate the phases and activities of the proposed methodology.

As a crop for the case study, the Mexican marigold flower is taken into account with historical data of more than 40 years, where through the results obtained it is shown that there has been a considerable decrease in the volume of production of this flower due to what Mexico sees displaced by other countries. We can see also that the yield has not progressed since 1980, it is variable from year to year, and these variations are significant even for the irrigation modality.

On the other hand, the results obtained in the relationships between variables for the cultivation of Mexican marigold allow us to know the states where it is mostly produced, so it can be known which is the most favorable climate for its production; its most convenient water modality, which is irrigation, the most convenient cycle to obtain a greater production, which is spring-summer, among other factors. With the above information, it is possible to develop agricultural policies that encourage its cultivation and that the cultivation of this endemic flower is not migrated and banished from its country of origin.

In general, the application of the proposed methodology to analyze the agricultural data of a specific crop in the Mexican context was satisfactory, showing the feasibility of applying it to the analysis of other crops in order to understand their behavior. It could be also useful to build a tool that has a graphical user interface which allows anyone to perform the analysis of different crops in a more friendly way.

ACKNOWLEDGEMENTS

We want to thank the students that realized the integration of data: download all the files, one file by year, extract only Marigold's data and do the cleaning; the results consisted in the file used in this work.

REFERENCES

Hotz, N. (2023a). OSEMN Data Science Life Cycle. https://www.datascience-pm.com/osemn/.

Hotz, N. (2023b). What is CRISP DM? https://www.datascience-pm.com/crisp-dm-2.

Luna, D. (2021). Flor de cempasúchil ¿China? México no figura entre los principales productores. *Expansion Política*.

Mandujano, J. (2020). Cempasúchil, la reina de los altares que desaparece en su país. *Enrrezando*.

Méndez, F. (2021). Cempasúchil, materia prima acaparada por otros países. *Gaceta UNAM*.

Páramo, O. (2017). China, principal productor de cempasúchil del mundo. *UNAM Global*.

SADER (2023). Secretaría de Agricultura y Desarrollo Rural- Gobierno de México. https://www.gob.mx/agricultura.

SIAP (2023). Datos Abiertos del Servicio de Información Agroalimentaria y Pesquera - Gobierno de México. http://infosiap.siap.gob.mx/gobmx/datosAbiertos_a.php.

Vázquez, H. J. and Juganaru-Mathieu, M. (2014). Handling missing data in a tree species catalog proposed for reforesting mexico city. In *6th International Conference on Knowledge Discovery and Information Retrieval*, page 457–464. 6th International Conference on Knowledge Discovery and Information Retrieval.

Vázquez, M. (2016). Manejo de enfermedades foliares con Trichoderma ssp. y Bacillus subtilis en cempasuchil (Tagetes erecta) del valle de Toluca. *Universidad Autónoma del Estado de México, Facultad de ciencias Agrícolas*.

Zamarrón, I. (2021). Cempasúchil: una flor muy mexicana… hecha en China. *Forbes Mexico*.