An ecological study of the food industry

Veillard Jennifer
jennifer.veillard@epfl.ch

Lanzrein Johan
johan.lanzrein@epfl.ch

Freundler Nicolas
nicolas.freundler@epfl.ch

Abstract

As ecology grows as a more important concern, populations need not only to focus on driving less or consuming less, but also on how they consume food. It is a known fact that consuming products derived from palm oil, or meat in general is much more damaging for the environment than growing your own vegetables. During this project, we aim to expose how to be more regarding towards the environment in our daily food consumption. The main goal is to investigate the world of food and understand what are some hints to take when we want to consume more eco-friendly.

1 Introduction

We decided to focus on the main ecological issues related with our daily food consumption. We first start to look into our dataset and notice how it is distributed and from this biased distribution we have to adjust our analysis. Afterwards, we compute the distance related information. We explore different aspect that are related to ecology, such as palm oil, vegan products, and labeling. A solution often presented is to eat locally. Thus we were interested in knowing what are the distance typically travelled by a product from its origin or manufacturing place until its selling place. Finally, we explore how Switzerland relates to its neighboring countries when it comes to labels.

2 Dataset Description

This project used the Open Food Facts database. This database compiles information about food around the world in an collaborative way. This means any user can add information about a food product she/he has to improve the database. In our study, we focus on mainly the following informations :

- Labels (vegan, green dot,...) described as $labels\_en$
- Palm oil information under fields such as $ingredients\_from\_palm\_oil\_n$
- Distance travelled from $origins\_tags$ or $manufacturing\_places\_tags$ to $countries\_en$

For each of those, we will make a link with the category of the product (column $categories\_en$).

3 Pre-processing

During the preprocessing phase, we create different datasets that each contain the relevant information to answer a specific problem. At this point, we notice how the dataset is flawed. As it is a collaborative dataset anyone can enter information and often we get fields that are empty which will clearly be detrimental for the analysis. To counter this we take a few measures. First of all, we decide to focus the study on France, as the database contains much more product from France then anywhere else.

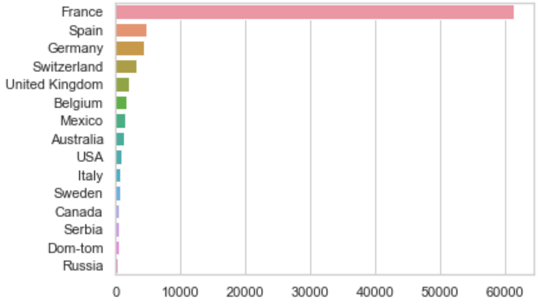

Figure 1: Number product sold in each country (top 15 countries)

In the end we have the following csv files that can be used throughout the project :

- *DistancePerProduct*: computed for each product with selling country and an origin or a manufacturing place specified. Basically

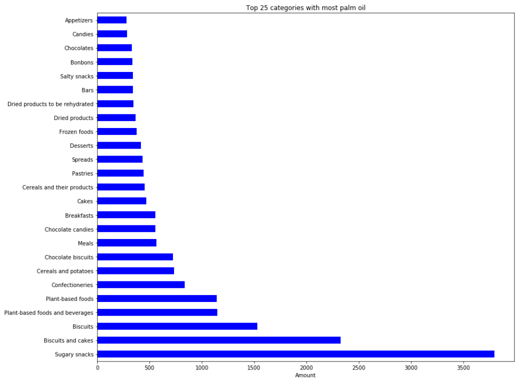

Figure 2: Categories consuming the most palm oil

the distance between both places is computed using the geopy library.

- *FrenchProduct*: all products sold in France.
- *labels* : all products with a labels
- *vegan* : all products which label is vegan
- *palm_oil* : all products with information about palm oil

In addition, the csv file *SeveralNamesOneCountry* contains a dataset used to build a dictionary to translate the countries names. Briefly, not all the nouns in *countries_en* are in English, this dictionary will give the English correspondence.

4 Data exploration

As a start, we compute how many articles we have total in the data set. We found a total of 665693 articles entered.

4.1 Palm oil

As palm oil is known for being a major factor for deforestation, we wanted to study what kind of products contain the most palm oil. We select items that contain palm oil and then look for the biggest category (see figure 2). As expected, categories such as cookies and snacks contain a lot of palm oil products. Moreover we also look at the evolution of number of palm oil articles over time. We notice that the number of articles containing palm oil follows a quadratic trend. So we could infer that there is more articles over time.

4.1.1 Labels and palm oil

When a consumer wants to avoid palm oil, he has no choice but to try and decipher the ingredients list. However there might be some indicators such as the label that could strongly suggest that the product contains palm oil. We explore this and found that there exists in fact an organization called RSPO (Roundtable on Sustainable Palm Oil), which delivers labels that can be found on products. The most prominent labels can be found on the figure 3.

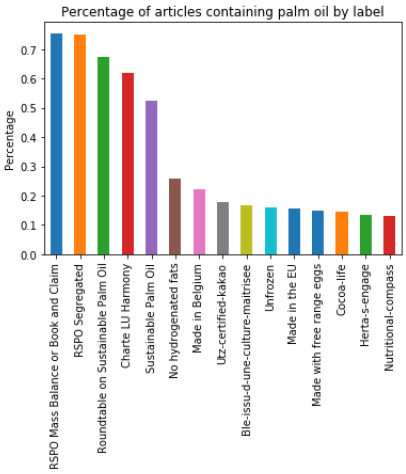

Figure 3: Percentage of articles containing palm oil by label

4.2 Labels

We found that labels concerning ecological and organic food products were the most popular however they represent only a fraction of the total dataset. There are a variety of labels. We decided to focus on two main labels that are related to sustainable food consummation. The first ones are labels that are related to ecological food preparation. Secondly, we investigate vegan product. Veganism has been argued to be an eco-friendly as it produces less waste and uses less ressources.

4.2.1 Ecological labels

Investigating labels related to ecological products, we found that :

- There are 37395 articles with label organic. This represents 0.056175 % of articles
- There are 13886 articles with label bio This represents 0.020859 % of articles

After our analysis, we saw that the most present labels were for instance Organic, Green Dot, and EU Organic. Hence even though there is a minority of products that are certified Organic, they are still the most present label in our dataset.

4.2.2 Vegan

Although the debate is still ongoing about sustainability of cutting dairies and meats from diet, people tend to adopt a vegan life-style, because of their anti- speciesism ideology or their will to reduce carbon footprint of their diet. For more information about this subject on link : The Independent.

In our data base, some products were labelled 'vegan' and were compared with the remaining data. One has to be careful that some product are labelled 'non-vegan' and too simple regex operation can take them as false-positives.

A first glance at origin distribution during preprocessing indicates that most of vegan-labelled products of our data base comes from France but Spain.

4.3 Distance relationships

Eating locally is an eco-friendly alternative. To understand clearly what kind of product travel a lot until reaching the supermarket we were interested in computing the distance. Answering this question was less straight-forward then the previous ones. A preprocessing set has to be done to select the data containing both a starting point (either origin or manufacturing place) and an arrival point (country, where the product is sold). In both cases, those data were transformed into coordinate thanks to geopy and then the distance computation was done using the same library.

In figure 4, we can see that products consummed by the french market are mainly from france, with Spain and Italy being behind.

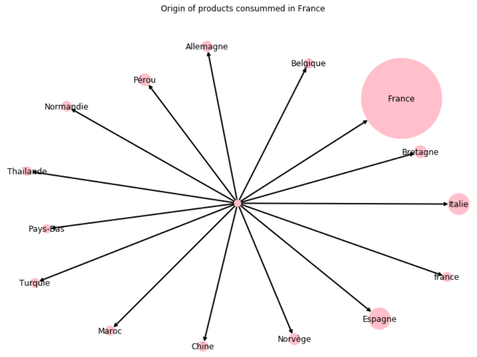

Figure 4: Network graph displaying origins of food products

4.3.1 Category and distance

In this section, we investigate which are the kind of product that travel more than others. We also focus only on french products. The hypothesis is that some category of products such as dairies and maybe meats will travel less than prepared meals for instance. To do this, we use the computations done beforehand and apply them to products before grouping them by category to see if we find any relations. We notice for instance that As can be seen in the figure 5, the distribution varies depending on the products. We noticed a major peak at the [0,2500] and the [7500,10000] bins. The distance from France to for example South America is in the bracket of 7500km and 10000km so we can suppose that exotic products such as cocoa beans for chocolate come from this area.

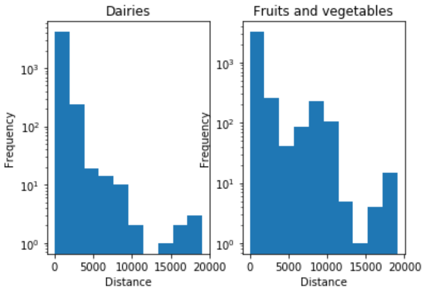

Figure 5: Comparison between dairies and fruits related products by distance

5 Statistical analysis

5.1 Distribution comparison

Distribution of fat and sugar was done between vegan-labelled and other products. The standard ones are paired-t tests, but they only apply to normally distributed population with same variances.

There exist a non-parametric paired-test for non-normal distributions based on ranks, called Wilcoxon test.

The histogram of fat and energy distributions on both classes strongly suggest non-normality of those features, as shown on 6. The Wilcoxon test was then performed. The hypotheses "Same distribution" and "Vegan more caloric" were rejected with p-values of respectively 9.60e-13 4.80e-13. Finally, the hypothesis "Vegan is less caloric" was not rejected with a p-value of 1.00.

The results for fat distribution are similar but a bit less stringer: for equivalent hypotheses for fat, the p-values are respectively 4.52e-5, 2.26e-5 and 1.00.

The equivalence between these two features is not surprising given that Spearman's correlation between energy and fat is 0.79 in vegan food and 0.74 in non-vegan data (p-values both 0.0). https://www.overleaf.com/1184642841swhbqhvmcvmx

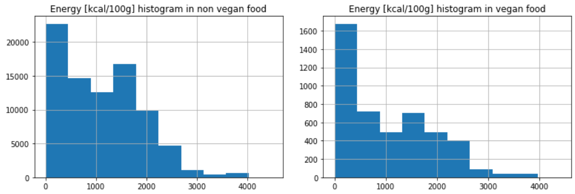

Figure 6: Energy distribution of non-vegan and vegan food

5.2 Association rule learning

Association rules are usually used in market basket analysis. The concept is to compute probabilities for a customer to purchase some items given that he/she has already purchase other ones. We adapted this method with two different features: the eco-friendly labels (Organic, EU Organic, Green Point) and the different categories (beverages, meals, ...). For instance, the association rule " EU Organic → Beverages " is equivalent to compute P(Category= Beverages | Label = EU Organic).

It is important to take in account only relevant local rules. To compute the relevance of rules "A → B" and "B → A", one computes the joined probability P(category = A, label = EU Organic). The twenty highest supports on the whole data set were kept for designing association rules.

Rules were computed for the total data set, for France and its neighbors. They were used a mean to compare eco-label consumption in those different countries.

As it can be tedious to compare association rules one by one, they were taken as a 1D vector. The similarity between two vectors can be computed with cosine similarity or linear correlation. As there is not any theoretical reason to prefer one method to an other one. We also adapted Tanimoto's distance that is normally used for bit strings, in order to test if it can be used for small real number vector comparison.

$$Tanimoto = \frac{\sum x_i y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i y_i} \tag{1}$$

The results show that the similarity assessment are not equivalent on our data: if country A is more similar to B in cosine similarity than to C, it is not necessarily true according Tanimoto's and linear correlation similarities. Belgium and France are more similar than Belgium and the rest of the world with linear correlation (0.76 vs 0.75) and cosine (0.55 vs 0.53) similarity, but it is the opposite with Tanimoto's (0.57 vs 0.58). In general France, Italy, Germany and the total data set have higher similarities measurements.

To visualize the results, the association vector were projected onto a 2D subspace. The immediate draw back of such a method is that dimensionality reduction might be too high and ignore some important information. Dimensionality was performed with three different techniques: principle componant analysis (PCA), independant componant analysis (ICA) and a autoencoder neural network (NN). For any of the techniques, data are normalized before execution.

Besides different scales of latent features, the three method show similar results, there are central clouds formed by Italy, Germany, France and World (the total data), whereas Spain, Belgium and Switzerland are outsiders, as shown in 7.

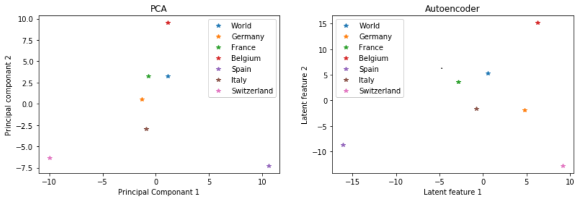

Figure 7: Dimensionality reduction of association rules by PCA and by autoencoder with 2x 128-neuron layers, 2x 64-neuron layers 1x 2-neuron layer and 300 epochs

6 Conclusion

Throughout this analysis of the Open Food Facts Database, we first noticed how incomplete real world datasets can be. This added the challenge of working with the incompleteness of the set. We learned how to complete information with external library such as geopy. Moreover we applied general data analysis technique seen during class to observe trends in the data. In the end, we used machine learning techniques to learn whether some European countries had similar behaviors.
The study of ecology in food is a subject that will be even more important in the future years. As society advances and the number of humans continues to augment, we need to look into sustainable and ecological way to produce food such that every human can have access to products.