AN INTERCONTINENTAL MACHINE LEARNING ANALYSIS OF FACTORS EXPLAINING CONSUMER AWARENESS ABOUT FOOD RISK

Alberto Tonda[1], Christian Reynolds[2], Nisrine Mouhrim[1], Rallou Thomopoulos[3]

[1] MIA-Paris, INRAE, AgroParisTech, University of Paris-Saclay, France
[2] Centre for Food Policy, City, University of London, United Kingdom
[3] IATE, University of Montpellier, INRAE, Institut Agro, Montpellier, France
email: alberto.tonda@inrae.fr, nisrine.mouhrim@inrae.fr
christian.reynolds@city.ac.uk, rallou.thomopoulos@inrae.fr

KEYWORDS
Food Habits, Food Knowledge, Risk Perception, Survey, Random Forest

ABSTRACT

This paper investigates to what extent food safety is perceived as a concern at the household level in different countries. It aims to identify the factors that best explain food safety concern, among the various food-related questions asked through a survey. To do so, a machine learning approach is used. The results show that the most significant explanatory variables of safety concern are the estimates of carbon footprints and calories associated with food products and primarily with beef and chicken meat. These results tend to indicate that people who are most concerned about food safety are also those who are best aware of environmental and nutritional impacts of food.

INTRODUCTION

To approach households' habits and beliefs at the domestic level, questionnaires and surveys have been an effective means of study. Among the fields of study, practices and perceptions about food, and more specifically food safety, have been well represented for years now and widely concern the population (EFSA 2019, Vedovato et al. 2014, Medeiros et al. 2004).
The collection of big quantities of data and their analysis for scientific, social research or business purposes naturally moved to the digital world (Kurtz and Thomopoulos 2021a, Salliou et al. 2019). A large-scale example at the level of the country was that of NutriNet-Santé in France, aiming to understand the relations between nutrition and health (Nutrinet-Santé 2022). On the one hand, the ease of participation offered by internet and the IoT (Internet of Things) boosted this type of studies. On the other hand, the handling of the data collected benefited from the progress of storage technologies, while the boom of data science offers emerging efficient analysis methods (Thomopoulos et al. 2021, Kurtz and Thomopoulos 2021b).
Literature has addressed perceptions of food risk by consumers in different geographic areas (Haas et al. 2021, Tucker et al. 2006, Van Kleef et al. 2007). Most recent studies have focused on emerging food safety related topics, such as the perceived safety of novel sources of proteins (Jarchlo and King 2022) or the impact of the COVID-19 pandemic on food safety perceptions (Sollid et al. 2022). This paper aims to explore the question: "What variables best separate individuals who express some worries about food safety risk, from those who do not?". As a first step, this is a thus classification problem. As a second step, regressions were performed to get further insight into the classification results.

MATERIALS AND METHODS

The data collected

A survey, focused on how people cook and what they know about the most common food items, was carried out at the international scale (Reynolds et al. 2020, Armstrong et al. 2021). The results obtained from identical questions asked 7 countries, namely Argentina, Brazil, Colombia, Ghana, India, Peru and United Kingdom, are used as input data in this paper.
The Qualtrics platform was used to ask everyday people to provide their opinions about images of food. For each food, the questions asked cover cooking and preparation of the food, food safety, food waste, how much energy is in the food, and the environmental impacts of these foods. The sample size is 3,247 and goes from 204 (Ghana) to 539 (India) per country.

Data pre-processing

Since food habits differ from one country to another, only common food products are considered. These common food products are: Beef, Chicken, Chard, Beans, Rice, Green beans, Carrot, Tomato, Bread (roll). Most commonly used green leaves vary from one country to another (chard, collard greens, etc.) but were consid-

ered as equivalent. For India, the "Chapatti/Roti" item was considered as equivalent to "Bread" for the other countries.
Among questions concerning the socio-professional status of respondants, participants were asked to input their individual and household weekly income. As these data were provided in the local currency, to avoid issues related to monetary conversion, all information related to income was normalized with respect to all participants from the same country.

Input and output variables

Classification step
The question about risk perception is formulated as follows: "According to your best guess, please rate how safe to eat the foods listed below are? i.e. how likely is it that eating them will damage your health due to risks such as contamination, food poisoning, improper handling, food fraud, mislabeling etc.". This question is asked on a scale from 0 (low risk) to 10 (high risk), for 5 foods: Beef, Chicken, Chard, Rice, Beans.
The output variable in our study is the maximum rate expressed over the 5 foods. We define two classes of respondents for this output variable:

- respondents for whom the output variable has a value below 5. This class (class 0) corresponds to individuals who express a low level of concern about food safety risk;
- respondents for whom the output variable has a value of 5 and more. This class (class 1) corresponds to individuals who express worries about food safety risk.

The rest of the variables of the survey are used as input variables.

Regression step
To get better insight into the interpretation of the classification results, regressions are then performed. Their objective is to further analyze the top-ranked variables of the classification results. The target variables of the regressions performed, are transformations of the top 4 explanatory variables obtained in the classification step. The transformations computed are detailed in the "Result confirmation with regression analysis" section below.

The analysis methods

Random Forest (RF) was used as the reference classifier for the experiments, taking into account both average accuracy in a stratified 10-fold cross-validation and interpretability of its results. RF (Breiman 2001) creates an ensemble of decision trees, training each one on a subset of the available data, thus reducing bias and delivering more robust predictions. RF determines relative variable (feature) importance, by evaluating the frequency of appearance of a variable in the splits of all the decision trees: The more a variable appears, the more important that variable is for the final classification of the ensemble. For all experiments reported in this work, the RF classifier has default parameters (Pedregosa et al. 2011, Scikit-learn 2021), using a total of 100 decision trees.
Most classifiers, alongside their predictions, are also able to return a ranking of the relative importance of the variables in the problem, with the ones that best explain the variance in the results among the top. In order to obtain a more reliable ranking, all the classification and regression experiments described in this work were performed with a 10-fold cross-validation, where available samples were randomly split into 10 groups (folds), and the system was iteratively trained on nine folds and tested on the one left out. The results from the ten folds were then averaged, and a standard deviation of the performance metric was computed, in order to obtain a more reliable estimate of the algorithm's efficiency.
During the cross-validation, data was normalized by removing the mean and scaling to unit variance, with this normalization learned on the nine training folds, to avoid information leakage towards the test fold.

RESULTS AND DISCUSSION

Major explanatory variables of food risk perception

The barplots displayed on Figure 1 show, on the X-axis, the list of variables selected in the 10 folds of RF stratified cross-validation to best discriminate class 0 (in green) from class 1 (in orange). The variables are ranked by decreasing importance, based on classification accuracy. The Y-axis provides the normalized mean value of each variable and shows the standard deviation. Although this standard deviation is relatively high, the classifier performance is still good, with an accuracy of 0.7 obtained.
The survey questions corresponding to these top-ranked variables are detailed in Table 1, in the same order as in Figure 1.

Result interpretation

The results reveal that the variables which best explain people's concern about food risk are the perception of carbon footprint and the perception of calorie content, for food in general (beef or lamb, chicken, rice, green leaves, beans) and most importantly for meat products (beef or lamb, chicken) which represent the top 4 explanatory variables.
Considering that meat is known to play a key part in the ecological impact of food (Godfray et al. 2018, Poore

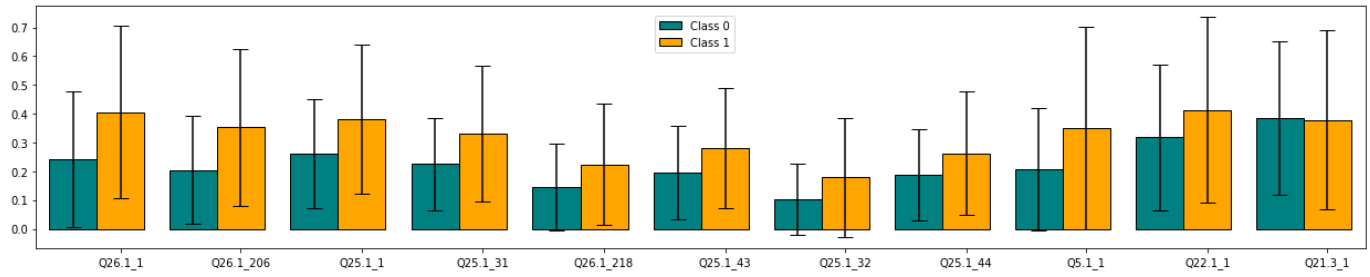

Figure 1: Top-ranked variables

Table 1: Most relevant explanatory questions identified

Id	Question
Q26.1_1	"According to your best guess, please estimate the carbon footprint (grams of CO2) embodied in the food portions that you typically eat - Beef or lamb"
Q26.1_206	"According to your best guess, please estimate the carbon footprint (grams of CO2) embodied in the food portions that you typically eat - Chicken"
Q25.1_1	"According to your best guess, please estimate the Calories (kcal) contained in the food portions that you typically eat - Beef or lamb"
Q25.1_31	"According to your best guess, please estimate the Calories (kcal) contained in the food portions that you typically eat - Chicken"
Q26.1_218	"According to your best guess, please estimate the carbon footprint (grams of CO2) embodied in the food portions that you typically eat - Rice"
Q25.1_43	"According to your best guess, please estimate the Calories (kcal) contained in the food portions that you typically eat - Rice"
Q25.1_32	"According to your best guess, please estimate the Calories (kcal) contained in the food portions that you typically eat - Green leaves"
Q25.1_44	"According to your best guess, please estimate the Calories (kcal) contained in the food portions that you typically eat - Beans"
Q5.1_1	"What is the most common way you usually purchase the food items listed below? - Beef"
Q22.1_1	"According to your best guess, please provide the typical method you used to cook the foods listed below when you eat them - Beef"
Q21.3_1	"According to your best guess, please estimate how long (in minutes) it takes you to actively prepare the foods listed below before you to cook and eat - Beef"

and Nemecek 2018, Vranken et al. 2014), the observation of these results raises the question of whether the "food risk concern" variable is a marker of the level of food education. Hence, we may hypothesize that the classification results obtained express a correlation between several variables representative of people's awareness of food-related issues.

Result confirmation with regression analysis

In order to further explore the above hypothesis, we considered 4 new variables, derived from the top 4 explanatory variables of the classification step. For each of these variables expressing respondents' estimates of greenhouse gas emissions or calories, for beef or chicken, we considered the difference, in absolute value, from the real greenhouse gas emission / calorie value of the given food. In other words, the 4 new variables measure how much the respondents are mistaken on the greenhouse gas emissions, and on the calories, of beef and chicken, respectively.

The regressions were performed in two configurations:

1. Including in the explanatory variables the questions on greenhouse gas emissions (group of questions Q26) and calories (group of questions Q25), for other foods than the targer one.

2. Excluding from the explanatory variables all the questions on greenhouse gas emissions and calories.

The R2 tests obtained are reported in Table 2.

As a result, we can state it is possible to well predict how much a respondent is mistaken about the greenhouse gas emissions and the calories of beef and chicken, using her/his answers about greenhouse gas emissions and calories for other foods. The prediction performance strongly declines if we remove these explanatory variables.

Table 2: Results of the regression experiments for the different target variables, using a 10-fold cross-validation. Mean values and standard deviation of test R2 are reported for each experiment. For reference, an R2 of 1.0 implies perfect predictions, while an R2 of 0.0 (or lower) corresponds to a poor predictive performance.

Regression target variable	**R2 of a 10-fold cross-validation (including Q25 and Q26 groups of questions)**	**R2 of a 10-fold cross-validation (without Q25 and Q26 groups of questions)**
Mistake on greenhouse gas emissions (beef), kg CO2	0.8007 +/- 0.0385	0.2139 +/- 0.0330
Mistake on greenhouse gas emissions (chicken), kg CO2	0.8281 +/- 0.0331	0.2241 +/- 0.0371
Mistake on caloric content (beef), kcal	0.7443 +/- 0.0395	0.2080 +/- 0.0359
Mistake on caloric content (chicken), kcal	0.7506 +/- 0.0266	0.2200 +/- 0.0449

This observation tends to confirm the existence of a correlation between a group of variables representative of the level of food education.

CONCLUSIONS

Based on a survey on food-related habits and opinions, carried out in 7 countries and 4 continents, this paper investigated the factors that explain people's concern about food safety. To do so, a machine learning approach was proposed in two stages.

In the first stage, classification was used to find out the variables that best separate people who worry most about food safety, from those who do not. Estimates of meat carbon footprint and of meat calories revealed to be the salient explanatory variables of food safety concern.

In the second stage, the hypothesis of a correlation between variables which are markers of people's awareness of food issues, was tested and confirmed using regressions. These regressions were performed on transformations of the top-ranked variables obtained in the first stage.

Correlation between several food-related concerns, observed in this paper, has also been pointed out in previous studies (Kurtz and Thomopoulos 2021b) and is thus confirmed in this study.

The results presented in the paper are quite homogeneous among the different participating countries, which represent different continents. However, some differences can be observed. In-depth presentation of these differences and exploration of their possible causes will be worth exposing in future developments.

REFERENCES

Armstrong B.; Reynolds C.; Martins C.; Frankowska A.; Levy R.; Rauber F.; Osei-Kwasi H.; Vega M.; Cediel G.; Schmidt X.; Kluczkovski A.; Akparibo R.; Auma C.; Defeyter M.; Tereza da Silva J.; and Bridge G., 2021. *Food insecurity, food waste, food behaviours and cooking confidence of UK citizens at the start of the COVID-19 lockdown. British Food Journal*, 123, no. 9, 2959–2978. doi:10.1108/BFJ-10-2020-0917.

Breiman L., 2001. *Random forests. Machine learning*, 45, no. 1, 5–32.

EFSA, 2019. *Food safety in the EU.* European Food Safety Authority, Publications Office. doi:10.2805/661752.

Godfray H.C.J.; Aveyard P.; Garnett T.; Hall J.W.; Key T.J.; Lorimer J.; Pierrehumbert R.T.; Scarborough P.; Springmann M.; and Jebb S.A., 2018. *Meat consumption, health, and the environment. Science*, 361, no. 6399, eaam5324.

Haas R.; Imami D.; Miftari I.; Ymeri P.; Grunert K.; and Meixner O., 2021. *Consumer Perception of Food Quality and Safety in Western Balkan Countries: Evidence from Albania and Kosovo. Foods*, 10, no. 1. doi:10.3390/foods10010160. URL `https://www.mdpi.com/2304-8158/10/1/160`.

Jarchlo A.I. and King L., 2022. *Alternative Proteins: Consumer Survey. Food Standards Agency.* doi:10.46756/sci.fsa.ncn554.

Kurtz A. and Thomopoulos R., 2021a. *Consumer perceptions of infant food safety in France.* doi:10.15454/ZPPOJH. URL `https://doi.org/10.15454/ZPPOJH`.

Kurtz A. and Thomopoulos R., 2021b. *Safety vs. Sustainability Concerns of Infant Food Users: French Results and European Perspectives. Sustainability*, 13, no. 18. ISSN 2071-1050. doi:10.3390/su131810074. URL `https://www.mdpi.com/2071-1050/13/18/10074`.

Medeiros L.C.; Hillers V.N.; Chen G.; Bergmann V.; Kendall P.; and Schroeder M., 2004. *Design and*

development of food safety knowledge and attitude scales for consumer food safety education. Journal of the American Dietetic Association, 104, no. 11, 1671–1677. ISSN 0002-8223. doi:10.1016/j.jada.2004.08.030. URL https://www.sciencedirect.com/science/article/pii/S0002822304014038.

Nutrinet-Santé, 2022. *Official website of the Nutrinet-Santé study.* https://info.etude-nutrinet-sante.fr/siteinfo/, Accessed 25 January 2022.

Pedregosa F.; Varoquaux G.; Gramfort A.; Michel V.; Thirion B.; Grisel O.; Blondel M.; Prettenhofer P.; Weiss R.; Dubourg V.; et al., 2011. *Scikit-learn: Machine learning in Python. the Journal of machine Learning research*, 12, 2825–2830.

Poore J. and Nemecek T., 2018. *Reducing food's environmental impacts through producers and consumers. Science*, 360, no. 6392, 987–992.

Reynolds C.; Schmidt Rivera X.; Frankowska A.; Kluczkovski A.; Bridle S.L.; Martins C.; Akparibo R.; Auma C.; Bridge G.; Armstrong M.B.; Osei-Kwasi H.; Bockarie T.; and Mensah D., 2020. *Cooking as part of a global sustainable food system - a 6 country pilot survey.* Poster presented at the Nutrition and Cooking Education Symposium, 12 Jun 2020, Newcastle, Australia. URL https://openaccess.city.ac.uk/id/eprint/24351.

Salliou N.; Taillandier P.; and Thomopoulos R., 2019. *VITAMIN project (VegetarIan Transition Argument ModellINg).* doi:10.15454/HOBUZH. URL https://doi.org/10.15454/HOBUZH.

Scikit-learn, 2021. *RandomForestClassifier.* https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, Accessed 6 January 2022.

Sollid K.; Dostal A.; Paipongna M.; and Smith K., 2022. *Food Perceptions, Beliefs, and Behaviors Amid a Global Pandemic. Nutrition Today*, 57, no. 1, 26–33. doi:10.1097/NT.0000000000000524.

Thomopoulos R.; Salliou N.; Taillandier P.; and Tonda A., 2021. *Consumers' Motivations towards Environment-Friendly Dietary Changes: An Assessment of Trends Related to the Consumption of Animal Products.* In *Handbook of Climate Change Across the Food Supply Chain*, Springer.

Tucker M.; Whaley S.R.; and Sharp J.S., 2006. *Consumer perceptions of food-related risks. International Journal of Food Science & Technology*, 41, no. 2, 135–146. doi:10.1111/j.1365-2621.2005.01010.x. URL https://ifst.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2621.2005.01010.x.

Van Kleef E.; Houghton J.R.; Krystallis A.; Pfenning U.; Rowe G.; Van Dijk H.; Van der Lans I.A.; and Frewer L.J., 2007. *Consumer Evaluations of Food Risk Management Quality in Europe. Risk Analysis*, 27, no. 6, 1565–1580. doi:10.1111/j.1539-6924.2007.00989.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1539-6924.2007.00989.x.

Vedovato G.M.; Bastos D.H.M.; Mancuso A.M.C.; and Behrens J.H., 2014. *A scale to evaluate customer attitudes towards food risks in restaurants. Health Surveillance under Debate: Society, Science amp; Technology*, 2, no. 4, 53–61. doi:10.3395/vd.v2n4.337. URL https://visaemdebate.incqs.fiocruz.br/index.php/visaemdebate/article/view/337.

Vranken L.; Avermaete T.; Petalios D.; and Mathijs E., 2014. *Curbing global meat consumption: Emerging evidence of a second nutrition transition. Environmental Science & Policy*, 39, 95–106.