Meet the Meat:
Studying the Carbon Footprint of Recipes

Alex Karpilow, alexandra.karpilow@epfl.ch
Chloe Dickson, chloe.dickson@epfl.ch
Nadine Gächter, nadine.gachter@epfl.ch

Abstract

With increasingly dire climate change forecasts, concerned individuals are asking how they can minimize their carbon footprint. Recent research suggests that reducing one's consumption of meat, in particular beef, is one of the highest impact actions an individual can take (Harwatt *et al., 2017*). To examine this topic, we explore the popularity and prevalence of meat in recipes. Specifically, we extract the ingredients from a recipe database and calculate the carbon footprint of each recipe. In that way we establish a rating that directly reflects the environmental impact of each recipe. Then we study how a higher carbon footprint is justified with a higher appreciation of a recipe reflected in its rating. Finally, we find that with clustering methods it should be possible to provide alternatives for high carbon footprint recipes that have lower carbon footprint and equivalent or better rating.

1 Introduction

Who of us thinks about the climate when planning a meal? Not everybody. And even if we do so, what does a climate-friendly meal look like? Our goal is to make the information on the environmental impact of each recipe available in an intuitive way. We want to classify and tag each recipe for its environmental impact analogously to the tagging for energy efficiency in electronic devices. We base our estimations on the paper of REFERENCE! where they calculated the carbon footprint of the highest impact ingredients. In addition, we propose an alternative recipe with lower carbon footprint and better rating for people looking at high carbon footprint recipes. In that way we hope to create an effortless change of people's lifestyles towards a greener future. To refine our work, we could include seasonal considerations for vegetables in order to encourage to buy more locally.

2 Data

In this work we are using a data collection of over 90'000 html files from wide used recipe webpages in the US. They include pages such as food network, epicurious and allrecipes. For this work we are only interested in the files that are recipes posted on the different pages. They feature information on the ingredients used, ratings, title of the respective recipe as well as the page it was published on. The most time consuming step of this work was the extraction through scraping of the quite differently formatted html files, the cleaning of the data and its organization in analyzable Panda data frames. In the following section we shortly describe our approach.

2.1 Data extraction

In order obtain a workable dataset, we used BeautifulSoup to scrape each html file for its title, rating, ingredients, and serving size. We developed code to extract information from the following websites: Allrecipes, Epicurious, Food Network, Food.com, BettyCrocker, myRecipes and Taste. We ignored the websites that included an aggregation of recipes or had a bizarre format (i.e. the website was not present in the title).

In the data extraction loop, we detect the website from the title then extract the desired features from the soup. While not all features could be found for all websites (i.e. myRecipes does not display the rating in an accessible format), we saved the recipe as long as the ingredients could be retrieved. The list of ingredients are then compared to a list of ingredient keywords that we are interested in: 'steak','lamb', 'beef', 'cheese', 'pork', 'turkey', 'chicken', 'tuna', 'egg'. If any of

these items are detect in the ingredient string and the string also contains a recognizable unit (i.e. pounds, grams, oz), all numerical values in the ingredient string are extracted using regular expressions and summed. The method used for detecting quantity is not foolproof. While inputs in the form '4 1/2 pounds beef' would correctly extract the quantity, ingredients with the format '2 kg (6 wings) chicken' which would yield a quantity of 8 wings. While we could later remove these outliers, it is perhaps advisable in future studies to take only the first number found, since doubling the quantity is more invalid than losing a fraction of the mass. All detected quantities were converted to kilograms and saved so that we know the quantity of each ingredient of interest.

We split the data into 10 subsets and ran each through the extraction loop, saving the resulting dataframes to csv for post-processing. Of the 90,701 samples we were able to extract the ingredients from 33,736. Further studies could be done to expand our scraping net.

2.2 Data cleaning

The carbon footprint and ingredient quantities were normalized by the serving size. Note that the serving size may be underestimated since we extract the first numerical value in the serving string scraped from the website. For example, if the servings are reported as '5-10 servings', the lower value is taken.

There are also a number of recipes where the ingredient quantity could not be determined since the unit was irregular (i.e. '1 cup chicken broth', '5 chicken wings'). We replaced the missing quantities with the normalized quantity of each ingredient of interest.

2.3 Carbon Footprint

In order to evaluate the meat ingredient's carbon footprint, we used a dataset from the Environmental Working group. (ewg, 2011). In their report, they have performed Life Cycle Analysis (LCA) for different animal protein, evaluating the entire life-cycle of the product from raising and feeding the animal to processing, including transportation, refrigeration and packaging, and many more factors of influence. Following LCA guidelines, the assumptions are based on products sold in the USA with average transportation and production methods taken into account.
Interestingly, one can note that eating local (meat from 161km versus 2,253km) has very little influence on the carbon footprint (less than 1%) of the meat consumed.

Meat ingredient	Carbon Footprint [kg CO2]
lamb	39.2
beef	27.0
cheese	13.5
pork	12.1
turkey	10.9
chicken	6.9
tuna	6.1
egg	4.8

Table 1: Carbon footprints per meat ingredient or animal protein source we used in our project, as synthesized by (GreenEatz.com, 2018).

2.4 Design of the Carbon footprint Tags

In order to establish the categories for our 'Greenness' label, we group the recipes containing ingredients of interest in five quantiles with respect to their carbon footprint. We observe, that the difference between two quantiles and hence between two tags corresponds to an order of magnitude for the value we calculated for the carbon footprint. The definition of the tags is summarized in table 2.

Green-Label	Carbon Footprint [kg CO2]
AA	0
A	0 - 0.0144
B	0.0144 - 0.0576
C	0.0576 - 0.5741
D	0.5741 - 2.608
E	2.608 - 329.7

Table 2: Summary of Carbon footprint value for the respective tags normalized for the number of Servings. The carbon footprint has been calculated following the calculation described in section 2.4

3 Analysis

How does the carbon footprint influence the appreciation of a recipe? In this section we carefully analyze the data we collected to find out, in what way cooking greener puts a strain on our tastebudds.

3.1 Data exploration - What tastes better, meat or no meat?

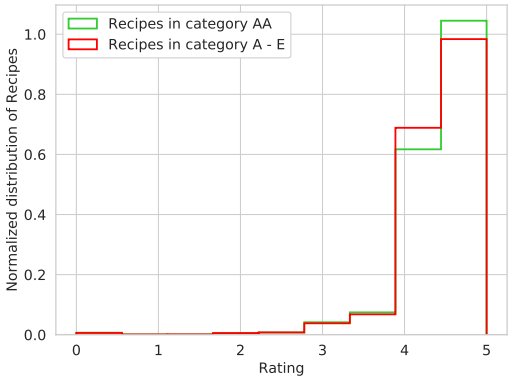

Figure 1: Distribution of recipes with respect to their rating.

Firstly, we look at the correlation between rating and carbon footprint and compare the mean rating in the two categories with and without ingredients that figure in the list for high carbon footprint identified in (ewg, 2011). We find a mean rating in group AA of 4.43 ± 0.55 and for groups A-E 4.40 ± 0.53, where the error is the standard deviation from the mean. The difference is not high enough to conclude that one is better than the other. Nevertheless, neither group disqualifies the other with respect to the ratings. In figure 1 you find the distribution of the recipes with respect to their rating for the categories with and without meat. As already indicated by the descriptive statistics they behave almost the same with a slightly higher percentage of excellent recipes in the category without high carbon footprint ingredients. Secondly, we display a density plot of the recipes with respect to their rating and carbon footprint (figure 2). We exclude all recipes with zero carbon footprint in order to exploit the properties of a log plot. We can see that some pattern emerges where the number of recipes is concentrated around two nodes in 'rating - carbon footprint' space. More importantly one can see that, for many recipes on the high carbon footprint side, there exist several alternative recipes on the lower carbon footprint side with equal or higher rating. Since the scales are logarithmic, these recipes can have an environmental impact that is over 100 times lower.

Finally, we separately analyze the properties of excellent recipes with rating between 4.8-5.0. The

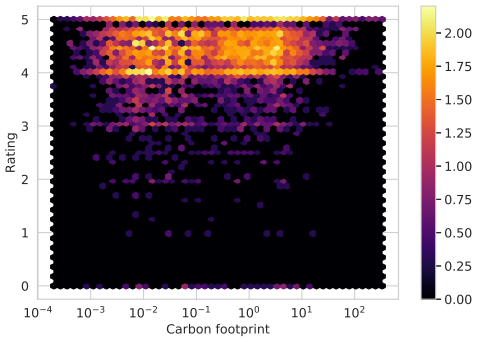

Figure 2: Distribution of recipes with non zero carbon footprint with respect to their rating and carbon footprint. (Categories A - E)

goal is to cook meals that have only little environmental impact, however the meals should be tasty! The histogram of excellent recipes with respect to their carbon footprint is shown in figure 3. We observe that there is a great number of excellent recipes with low carbon footprint. Furthermore in figure 4 we display the percentage of excellent recipes for each greenness label. It becomes clear that both groups AA and E show the highest percentage of excellent recipes. The amount of terrible recipes with rating lower than 3.0, on the other hand, is with differences of less than 0.01% equally present in all groups regardless of its environmental impact. To conclude, we have shown that there is a big pool of excellent recipes with very low carbon footprint that can be proposed as an alternative to the high carbon footprint recipes.

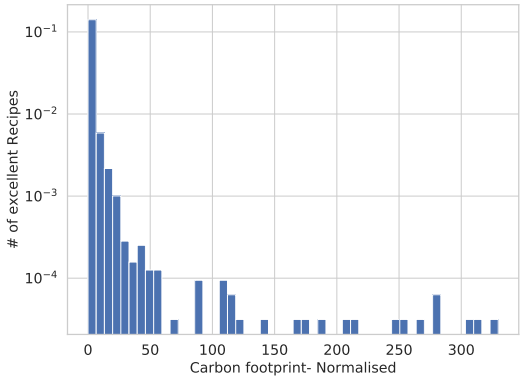

Figure 3: Histogram of excellent recipes (Rating 4.8-5) with respect to their carbon footprint.

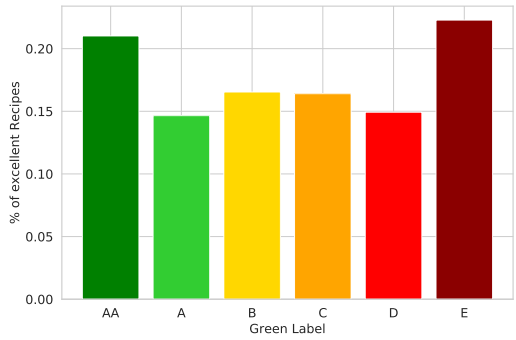

Figure 4: Percentage of excellent recipes with rating between 4.8 - 5.0 within each group.

3.2 Clustering Recipe Titles

By clustering the recipe titles we can potentially identify similar recipes with lower carbon footprints. The titles are often in the format 'Thai-Style Vegetable Curry Recipe — MyRecipes.com' so we needed to perform some cleaning. Stop words and punctuation were removed as well as the website title, the word 'recipe' and the names of several prominent chefs. We then created tokens of the titles, a vocabulary of keywords and an array with the simplified titles.

Using wordCloud we can observe the most frequent words in the recipe titles, as shown in figure 5. Thematic words like 'chicken', 'salad', 'soup' and 'cake' dominate the text while descriptive adjectives like 'spicy' and 'creamy' also show up frequently.

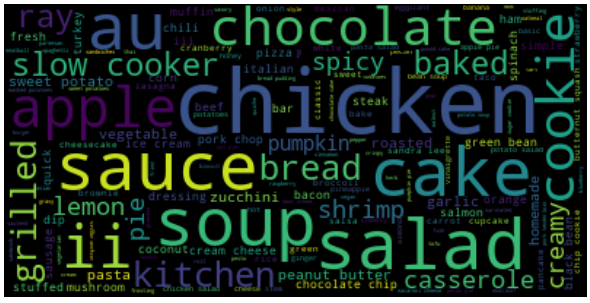

Figure 5: Most frequent words in recipe titles

The vocabulary was transformed into vector space with tf-idf and the titles were clustered using the k-means algorithm where k=100. Observing the top terms per clusters, we saw that the themes of the recipes are maintained. Recipes clusters include sugar cookies, bean salad, pork chops, corn chowder and many more. By limiting our data set to recipes with ratings over 4.8 and then selecting the recipe with the lowest carbon footprint from each cluster, we can select the best recipes with the lowest footprint for each cluster. However, it was difficult to visualize 100 clusters, so we performed the rest of our analysis with only 10. While this resulted in poorer resolution and prediction accuracy, the number of clusters proved to be more manageable.

With 10 clusters, we can see that the themes are much more general with clusters containing 'chicken, soup, bean' and 'casserole, baked, cheese'. Examining the median ratings as shown in figure 6, we see that the ratings are approximately the same across the clusters, indicating that no cluster is rated higher than the others. The carbon footprint on the other hand is heavily skewed towards the cluster 'chicken, soup, beans'.

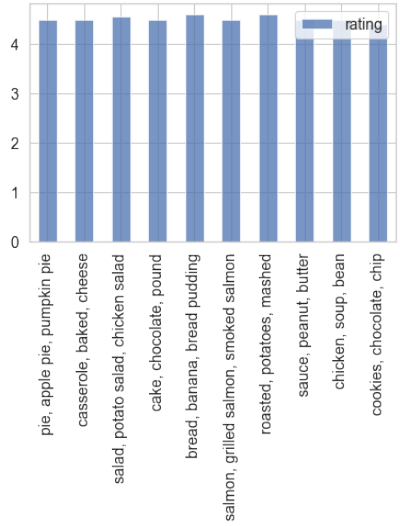

Figure 6: Median ratings of clusters

Finally, examining the number of recipes in each cluster, we found that the 'casserole, baked, cheese' cluster was orders of magnitude larger. The keywords indicate that the cluster is extremely broad, and could encompass a variety of recipes. However, as it would be preferable for the clusters to be more evenly sized, future steps might be taken to distribute the recipes more fairly. Once we remove this cluster, we see in Figure 8, that the distribution is much more even, with 'salad' being the next largest cluster and 'salmon' the smallest.

We also tested the predictive accuracy of our clustering model by inputting a new recipe title and extracting the suggested cluster. The accuracy of this method was only based on our own judgment which is not a reliable metric. How-

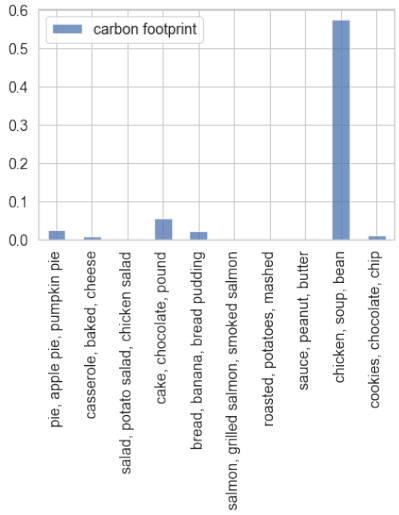

Figure 7: Median carbon footprint of clusters

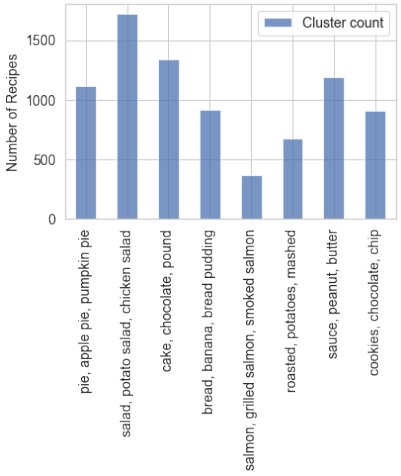

Figure 8: Number of recipes in each cluster with dominant cluster removed

ever, we were able to get a sense for how well the model performed: in some cases when keywords like 'cheese' or 'chicken' are present, the model could accurately detect its cluster. While some recipes appeared to be mis-classified, this is likely due to the overlap and limited number of clusters. After predicting the cluster, we can then return a selection of 3 other recipes from that cluster with high ratings and lower carbon footprints.

We can also view the clusters in 2D space by projecting the title vectors onto two features using Principal Component Analysis (PCA). The resulting distribution of samples and their corresponding clusters is shown in Figure **??**. We can see that sweets are generally grouped to the left, including banana bread, cake, and cookies, while savory recipes cluster to the right (with chicken, salads, and casseroles). However, the overlap suggests that our clusters are not very distinct and several recipes are either shared or on the border between clusters. The low resolution of our clusters affects our prediction accuracy and could be addressed by increasing the number of clusters created.

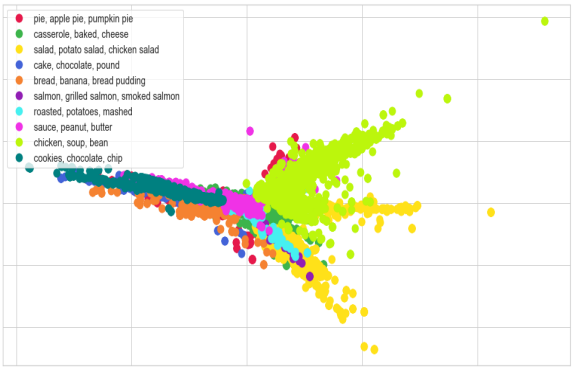

Figure 9: Word clusters projected into 2 dimensions

4 Conclusion

In this report we explored the relationship between recipe ratings and their carbon footprint in order to draw conclusions about diet and conscious living. With the goal of increasing environmental awareness we investigated whether low impact recipes are preferred over high impact ones. After categorizing the recipes based on carbon footprint, we found that the difference in ratings between the high and low impact recipes was not significant, indicating one is not preferred over the other. Furthermore, we found that there are just as many highly rated recipes with low carbon footprints as there are with high carbon footprints, indicating that consumers have a significant number of options when it comes to selecting a low impact meal.

Out of curiosity, we also examined the results from clustering the recipe titles using vectorizers and the k-means clustering algorithm. We confirmed that the ratings of the different clusters are nearly indistinguishable while the carbon footprint varies significantly. We were also able to roughly predict the cluster of a new recipe title. However, our results were hard to draw conclusions from since our clusters were too broad. Further studies should be done with more clusters to improve our resolution.

There are numerous ways to improve this study including expanding and modifying our web-

scraping code so that we can obtain more data and increasing the number of clusters so our prediction model performs better. It would also be useful to extract additional information from the websites such as number of times visited or number of ratings.

Overall, we can conclude that low impact recipes are both numerous and highly rated. In every category, from soups and sandwiches to salads and casseroles, there are various options for those looking to decrease their carbon footprint while still enjoying a delicious meal.

References

Helen Harwatt, Joan Sabaté, Gidon Eshel, Sam Soret, William Ripple 2017 *Substituting beans for beef as a contribution toward US climate change targets.*

Environmental Working Group (ewg) 2011. *Meat Eaters Guide: Methodology 2011 (page 19).* `http://static.ewg.org/reports/2011/meateaters/pdf/methodology_ewg_meat_eaters_guide_to_health_and_climate_2011.pdf` (last consulted: 16.12.18)

GreenEatz.com *Food's Carbon Footprint* `https://www.greeneatz.com/foods-carbon-footprint.html` (last consulted: 16.12.18)