Nutriperso project: Spring and fall 2018

Andrei Constantinescu
Under the supervision of profs. M. Sebag and P. Caillou

October 25, 2018

Abstract

This work aims to take profit of the Kantar WorldPannel dataset of consumers to derive useful public health insights. The dataset mixes socio-economic information as well as food purchases over the year 2014 of a sample of 24,000 households. Here we seek to investigate 3 questions : 1- what links are there between social variables and eating habits 2- are there certain types of food that have a significant influence on Body Mass Index and by extension general health 3- can we build a recommender system to substitute harmful products with healthier ones while respecting people's taste.

1 Introduction

1.1 Description of the data

The data this work is focused on comes from Kantar Worldpannel 2014. Kantar Worldpannel is a company that specialises on doing marketing surveys. On the website, we found the following description of the data we used. Although it is in French, we thought it provided very valuable insight into the design of this dataset, and which analyses it aims to make possible.

'[...] Worldpanel Usage vous permet de comprendre en profondeur vos consommateurs pour savoir, par exemple : Qui sont les gros, les petits consommateurs de votre marque? Quelle est la consommation des gros acheteurs de votre marché ? Quelle mixité de consommation entre les différentes marques du marché ? Comment votre marque est-elle partagée au sein du foyer ? Quelles sont les marques concurrentes en consommation ?

Nous avons à votre disposition toute une palette d'études spéciales pour vous permettre d'affiner votre connaissance des consommateurs de votre marché et avoir ainsi une vision complète de votre marque, via vos acheteurs et via vos consommateurs.

Les données Worldpanel Usage sont basées sur les achats réels des foyers et des questionnaires complétés par tous les individus du foyer, sur un échantillon très large (+ de 28 000 individus interrogés), ce qui rend nos données très robustes et vous permet d'avoir un accès à un niveau de détail fin.'

Kantar WorldPannel is a global brand, specializing in giving marketing experts valuable insight about their customers. Our data was collected with a specific business oriented mindset. Its aim is to help Fast-Moving Consumer Goods Companies answer the following questions: Who are my customers? What is their consumption patterns? How is the market divided among brands? How do I perform compared to my competitors?

The focus is therefore on brands and marketing. Brands are the end customers of Kantar, and so the data is tailored to their needs. As a result, products are described with a wealth of details: about flavor, packaging, etc., or all the aspects brands might deem valuable in their analyses.

We can see already a clash in intentions. Our goal, broadly speaking is related to public health, not market share. What we want is to better understand consumption patterns – which parts of the population are subject to bad eating habits risks -, and how these bad eating habits relate to health. By chance, description of the household include height and weight of the household members – therefore one can compute the Body Mass Index, an indicator of overweight, and by extension of general health. This disparity in intention will translate is several challenges and limitations for our research.

There are mainly 3 data-tables of interest that we will manipulate:

1. menages 2014: Socio-economic description of households
2. produits achats: Description of products – without nutritional information about the data
3. achats 2014: Purchase data where each line corresponds to one purchase of one product by one household

Furthermore, below are some interesting figures about our data:

Number of households	25,000
Number of products	170,000
Number of purchases	10,400,000

1.2 Goals of the study

This study is conducted from a public health perspective. Our goal is to use Machine Learning techniques to gain insights into national food consumption and how eating habits relate to health. Several questions come up to mind in this respect:

What links are there between socio-demography and eating habits? Which part of our lives influence the way we eat (age, education, etc.?)? Which are the segments of the population with the worst eating habits? Who should the government target in priority to improve eating habits?

Can we infer relationships between the food eaten and health – as measured by the BMI? Which foods are the most poisonous for our bodies, and which foods are the healthiest?

Is it possible to act on these eating habits? Can we build a sketch of a recommender system, where we would recommend food to people while taking into account both their taste and their health?

These questions will direct our study. Nevertheless we are aware that since the data was not gathered with such a perspective in mind, our answers will often be incomplete or partial. Nevertheless we still think that the answers we found constitute interesting leads to explore.

1.3 Caveats and preprocessing of the data

The data does not come into a amenable format for our study. This impacts both our representation of products, and our representation of households.

Ages mbr foyers	Famille	Geo/ habitation	équipement ménager	animaux compagnie	cat socio p	equipement info	résidence secondaire	phys
Nbr enfants – 3ans (en3)	famille recomposée (fare)	aire urbaine (aiur)	Lave-vaisselle (lvai)	Nbr de chats (cha)	Classe socio-économique ocde (scla)	nombre d'ordinateur fixes et portables (mor)	arbres fruitiers résidence principale (fru1)	poids de l'individu I du foyer (ipdsi)
Nbr enfants -6 ans (en6)	nombre de personnes au foyer (nf)	Département (dpts)	Lave-linge indépendant (malt)	Nbr de chiens (chie)	catégorie socio-profesionnelle individu i (cspc)	nombre d'inidividus possesseurs de téléphone portable (tlpo)	disposition d'une résidence secondaire (rs1)	Taille de l'individu I du foyer (ihaui)
Nbr enfants -15 ans (en15)		Type d'habitation (thab)			niveau d'étude individu i (etuc)	nombre de téléviseurs (tvc1)		
Nbr enfants -25 ans (en25)		statut d'occupation du logement principal (socc)			activité profesionnelle individu i (itra)	nombre de voitures (voit)		
age du chef de foyer (agec)					revenu mensuel brut du foyer (rve)			
age du panelliste (agep)								
position dans cycle de vie (cycle)								

Figure 1: Description of socio-demographic variables

1.3.1 Preprocessing of households

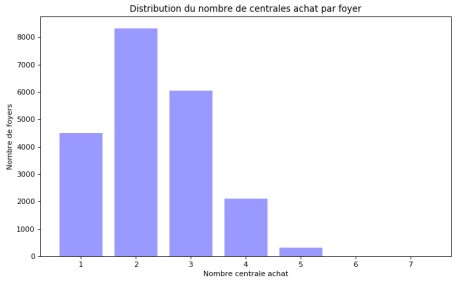
Socio-demographic data comes in form of a table where each line is a household, and each columns is a socio-demographic variable of the household (age of member 1, education of member 1, etc.). The household representation is the easier problem to handle. Households are represented with a wealth of detail about several aspects of the daily life. We kept most of the features handed to us unchanged, processed some, and discarded the features we found were irrelevant for our study - information about house appliances, pet sizes, etc. Figure 1 provides a description of the features we kept, subdivided into broad categories.

Sadly, there are also many features we would have loved to have, mainly about key life style aspects that heavily inform overall health – smoking, physical activity, etc., For these features, we could not do much, except acknowledging the imperfection of our data,.

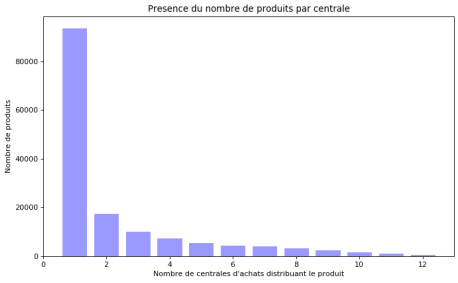
1.3.2 Preprocessing of products

Data about purchases and products comes into another form. The purchase data comes as a table where each row is a purchase of a product, and each column is information about the purchase and the product. We created a separate table for information about products. The issue at hand is that the representation of our purchase data is too sparse to obtain meaningful results. Indeed, we have 25,000 households that have purchased some of 170,000 products. This gives place to an extremely sparse purchase matrix, where only 0.2% of entries are not null. Moreover we are here in a pathological case for most machine learning algorithms, since we have largely more features than samples. Therefore, we will strive to reduce this dimensionality, by using prior knowledge.

We noticed especially two aspects in our data that guided and motivated our clustering of products - the technique of dimensionality reduction we used. These aspects are highly linked to the nature of our data. The first goal of our clustering was to overcome the divide of products by



(a) Distribution of households by how many retailers they go to



(b) Distribution of products by how many distributors they are sold in

Figure 2

distributors - by distributors we mean retailers, such as Carrefour, Auchan, etc. We noticed two things in our data. Figure 2 – left graph - shows that households tend to go always to the same distributor for their shopping. Very few are the households go to more than 3 different distributors on a regular basis. Be it taste, geographical convenience, or something else, they tend to be loyal to a few distributors only. These few distributors represent more than 90% of their purchases – out of more than 10 distributors in the market. The second thing is that distributors tend to offer in some large extent exclusive products, products that can only be found in their outlets- right graph. Distributor A will be the only one selling brand A ham for instance – often it is a brand baring the name of the distributor brand, for instance Jambon Carrefour- and distributor B the only one selling brand B ham. Figure 2 right graph shows the distribution of products by how many distributors they are sold in. On sees that most products at hand only appear in one distributor. This is harmful because it artificially makes our households live in different dietary universes. Since such a big chunk of products are particular to a distributor, each distributor delineates in effect its own dietary space, with only partial connection points with other such spaces.

The second challenge is the problem of marketing. For a same category of product, brands will strive to differentiate themselves in order to seize market share. One can think of the example of sodas. There exists a wealth of brands, flavors, packaging of sodas. Nevertheless one might think that they have very similar effects on health. This is a point that is particularly damaging for us. The Kantar Worldpannel dataset is especially precise in this respect, giving for each product a wealth of details not for nutritional purposes, but for marketing ones, to help brands identify precisely which kinds of products are popular with which kinds of customers. The data distinguishes between Coca-Cola and Pepsi, between Pepsi in small size or in big size, between Pepsi with one taste and another, etc., and this yields the huge number of individual products - over 170,000. We argue here that although this aspect is important from a marketing perspective – to understand which specific product is successful on which demographic, from a nutritional perspective one can assume that all these products have the same impact. Therefore, in order to extract most information out of our data, in order to make households who go to different distributors and who are differently influenced by marketing comparable, we decided to operate a quite restrictive clustering. In effect, we want to blend back all these types of food that have artificially been separated. We found that it is not only useful from a Machine Learning perspective to cluster the data, but also from a practical point of view as well.

Aide Culinaire	Alimentation BB	Apiculture	Assaisonnement	Biscuits	Boisson	Charcuterie	Chocolat	Confiserie	Confit	Dessert	Fromage
Aide A La Patisserie, Aide culinaire pour Sauce, Autre aide culinaire, Beurre, Bouillon Et Court-Bouillon, Chapelure, Crouton, CrÃ¨me en Bombe, CrÃ¨me fraiche, Farine, Fecule, Gelee A Preparer, Graisse Conditionnee, Huile, Jus De Citron, Lait De Coco, Margarine, Marinade, Produit sucrant, PÃ¢te Ã Garnir, Pate Ã Tarte, Vinaigre	Alimentation BB, Biscuit BB, Boisson BB, Dessert BB, Farine BB, Petit Dejeuner BB	Gelee Royale, Miel, Pollen	Ail Echalote Oignon, Ail frais, Aneth fraiche, Autre assaisonnement, Autre herbes fraiches, Basilic frais, Cerfeuil frais, Ciboulette fraiche, Coriandre frais, Echalotte fraiche, Epices, Estragon frais, Herbes, Menthe fraiche, Melange d'herbes fraiches, Melange pour Assaisonnement, Oignon frais, Persil frais, Poivre, Sel, Thym frais	Barre Cerealiere, Biscuit Aperitif, Biscuit Sucre, Bouchee Cerealiere	Aperitif, Autre Alcool, Biere, Brsa, Cafe, Champagne, Chicoree, Cidre, Cocktail, Eau, Infusion, Lait, Mousseux /Petillants, Poire, Punch, Rhum, Sirop, The, Vins, Whisky-Bourbon	Andouille, Andouillette, Assortiment de charcuterie, Autre Porc Cuit, Autre charcuterie, Bacon, Bouchee A Croquer, Boudin, Cervelas, Chair A Saucisse, Chorizo, Foie Gras, Fromage De TÃªte, Lardon Poitrine, Lardons, Lot Mixte Pate, Mousse, Pave, Poitrine, Pate, Rillettes, Rosette, Roulade, Roti, Salami, Saucisse, Saucisson, Specialite Italienne de charcuterie, Tripes, Viande Des Grisons, Viande Froide	Avec Un Objet, Barre, Bouchee, Calendrier De l'Avent, Moulage, Tablette, Oeuf Surprise	Bonbon, Bubble Gum, Chewing Gum	Confit Frais, Confit en conserve	Dessert Frais, Dessert Surgele, Dessert en conserve	Fromage, Fromage camenbert, Fromage cheve, Fromage coulommier, Fromage fondu

Figure 3: Groups (1st row) and subgroups(2d row)

In order to do this blending, we have been guided by a table providing information about the products. There is structure in this table, as our products are divided in 30-odd groups, that further subdivide into 200-odd subgroups. Figure 3 gives a sample table of these groups (1st row) and subgroups (2d row, divided by commas).

Further, we have other attributes for each product. These attributes are consistent only within subgroups, and change from one subgroup to another. For instance, we will have different variables for wine (couleur, cepage, region, degre d'alcool), than for beer(couleur, alcolise). Of course, in addition we have information about the brand, the packaging, and the maker.

Our clustering protocol was straightforward. We looked at each subgroup, selected manually 2 or 3 attributes of most interest - for beer it was color, and contained-alcohol, and considered products that had identical entries in the subset of attributes we selected as identical, in effect clustering them together. So for beer, we considered all blond alcoholic beers as the same thing. We discarded of course all information about brand and packaging (conditionnement), for reasons explained above, and with this approach could therefore reduce the space of products to 4000 – from an initial 170,000. We found that this scale was a good compromise between information lost, and inference power gained. This judgment was mainly based on iterated testing in downstream tasks such as BMI prediction, or unsupervised clustering of products. Nevertheless many other clustering of products are possible, and we can hardly argue that our clustering is the best.

2 Phase I : Relationship between eating habits and social condition

Figure 4 sketches out our status so far. What we have is effectively two different descriptions of our households. The first is a classical socio-demographic description, the other is a dietary description

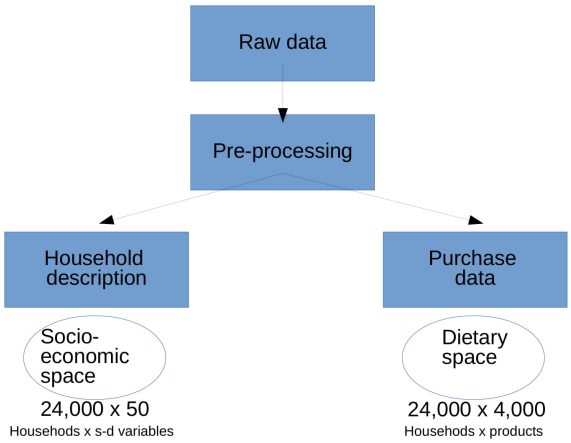


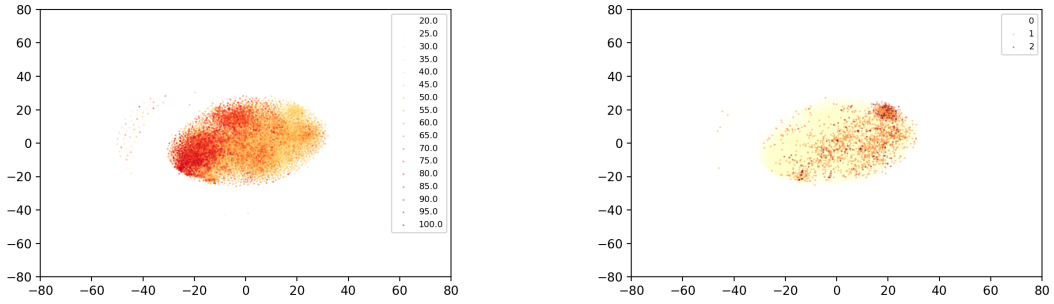
Figure 4: Sketch of our status

(which products have been purchased by the households during 2014). This is the same as saying that our households are two points in two different spaces, the socio-demographic space as well as the dietary space. What is of most interest to us now is to see the links between these two spaces? In effect, the question we will try to answer in this first phase is the following: is it possible to draw some links between social condition and eating habits? What impacts the way we eat? Is it age, occupation, geographical location, education? And of course, whom among these clusters of the French population has the worst/ best eating habits?

The first thing to note is that we have no indicator of quality of eating habits. Our products are not tagged with their nutritional attributes, or with a score that would indicate how beneficial they are for general health - such as the nutri-score. One thus cannot compute an overall score for the diet of a household. This limits a lot the scope of our study, but is not disqualifying either.

A first approach will be to seek which socio-demographic attribute affects people diets? Do people eat something because they are rich, old, have kids? Which dimensions of our lives impact the most the way we eat? This approach can lead to interesting sociological insights, and give levers of action in order to improve diets of people - by giving an insight into which mechanisms inform the choice of a diet.

A second approach that is less amenable to our data, will be to seek who among the French population has the riskiest eating habits. For this, since we do not have nutritional information about our products, we will base ourselves on the only piece of information we have : the BMI. The BMI is certainly dependent on what one eats, and thus translates - although imperfectly- the quality of the diet of the household. We will thus in a second time try to see which socio-demographic factors impact the most prediction of BMI. This will hopefully help us identify which subparts of the population one needs to target in order to improve public health.



(a) Households in dietary lsa space colored by age

(b) Households in dietary lsa space colored by number of babies

Figure 5

2.1 The LSA model

We want to know which aspects of our socio-demographics inform the way we eat. In order to answer this question, we will first project households in the dietary space and color them by relevant sociodemographic features. This entails thus 1- to find a representation of households in a low dimension in the dietary space 2- see if there is some features that is highly correlated with some cluster in this dietary space.

The idea of the LSA model is to approximate a high-rank matrix by a product of lower rank matrices, corresponding to the truncated SVD decomposition of the initial matrix. One sub matrix providing a latent representation of the households, the other a latent representation of the products.

$$S \approx U\Lambda V \tag{1}$$

with S $\in \mathbb{R}^{NxM}$, U $\in \mathbb{R}^{Nxd}$, $\Lambda \in \mathbb{R}^{dxd}$ and diagonal, V $\in \mathbb{R}^{dxM}$, and where typically d $\ll$ N,M.

One metric of how well the data lends itself to such schemes is the captured variance, i.e. how much of the variance is captured by the condensed representation of the matrix. It is usual to use this metric to decide of a cut-off dimension for d. Here 300 dimensions capture approximately 75% of total variance.

A useful tool to visualize this low-dimensional dietary representation of households in a classical 2-D space is T-SNE[1]. T-SNE is a dimensionality reduction technique that preserves local neighborhoods, enabling easy visualization of data in high dimension. Figure 5 is the T-SNE representation of households then products colored by relevant features - here age and number of babies in the family.

The right part of figure 5 acts more as a sanity test. Indeed, with the diet of a baby being very specific. Thus, any family with a new born baby should cluster together by virtue of this specific food they will have to buy for their baby. And indeed, we can see a strong dark cluster. The left part of figure 5 is also very interesting. Age is the only socio-economic variable that colored in a powerfully intuitive way our scatter plot. We see here that old people (darker) then to cluster together

Bébé	Dessert frais	Alcool
Alimentation BB : 0.31 | 0.99 | 100.35	CAZAUBON : 0.08 | 0.93 | 48.88	Vins : 0.28 | 0.59 | 24.98
Dessert BB : 0.21 | 0.98 | 98.47	DANONE.ACTIVIA : 0.07 | 0.4 | 21.03	Bière : 0.11 | 0.41 | 17.53
Petit Déjeuner BB : 0.04 | 0.96 | 96.75	LES JACQUINS : 0.07 | 0.47 | 24.78	Apéritif : 0.07 | 0.51 | 21.84
Biscuit BB : 0.01 | 0.95 | 95.4	DANONE.DANETTE LE LIEGEOIS : 0.04 | 0.32 | 16.91	Whisky-Bourbon : 0.05 | 0.67 | 28.66
Farine BB : 0.03 | 0.93 | 93.99	DANONE.DANIO : 0.04 | 0.72 | 37.74	Brsa : 0.03 | 0.02 | 0.73
Boisson BB : 0.08 | 0.93 | 93.54	PETIT BASQUE LE : 0.03 | 0.21 | 10.99	Mousseux/Pétillants : 0.02 | 0.44 | 18.89
Dessert en conserve : 0.03 | 0.04 | 4.31	DELISSE : 0.03 | 0.08 | 4.06	Biscuit Apéritif : 0.02 | 0.04 | 1.61

Figure 6: 3 topics output by the LDA with their most representative products

in their eating habits, and distinguishing themselves from younger people/ households. Here, since we can have several people in the household, we took the mean of the adult members of the household.

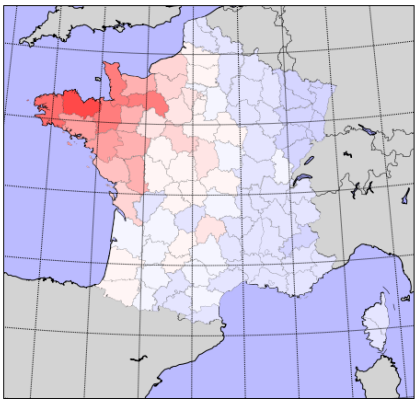
Nevertheless, this analysis is not satisfying. The LSA model is a crude approximation of a household diet's, that but poorly models the households' taste. The problem comes also from our data. The LSA model considers each household as a unity. But, it would be better to see them as composed of people, each with their specific taste, that each influence proportionnaly how the global household basket will end up. Mixture models are good at especially that, finding again diversity in unity, and among them the Latent Dirichlet Allocation model - LDA is a robust and well-used choice.

2.2 The LDA model

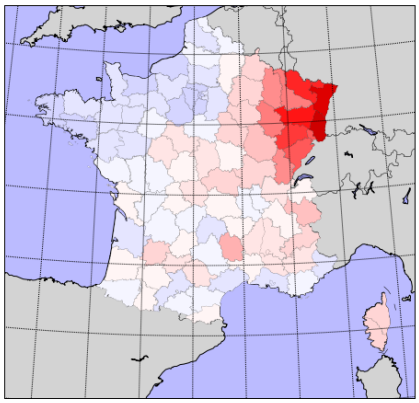
LDA is a technique introduced by Blei and al. [2] for unsupervised learning. Its aim is to model data generation as a mixture model. Each document – here household, is obtained by sampling randomly from a small number of topics, individual taste and preference. Topics are therefore distribution among products - or preference among products, which product am I more likely to buy, do I prefer? The idea is that purchases from a household will spring from various tastes or needs, and from then will blend together into the global basket of the household.

The result of the LDA analysis is thus two latent representations, one of products, and one of households. Households are represented as a probability mass on k topics, with only a small number of topics appearing. Products on the other hand are ventilated through the k topics, with each product appearing with a significant probability only in a small number of topics. To choose the optimal number of topics k - or clusters of meta preferences, we used a train – validation paradigm and optimized the perplexity parameter. The optimal number of topics found – i.e. a number that does not under-fit nor over-fit is 30.

About the products, below can be found 3 reasonably intuitive topics. We can see that the LDA has divided its topics in consistent product categories - i.e. baby products, alcohols, etc. This gives one key to interpreting the LDA topics. Some of our topics consist of required food - e.g. water, baby food, that one needs to buy to satisfy some basic need. Figure 6 illustrates such topics consistent within one category of products.



(a) Geographic Distrib of Topic 2 cons.(cidre, cheese, etc.)



(b) Geographic Distrib of Topic 16 cons.(sausages, pastry paste, etc.)

Figure 7

Then, we have a wholly different class of topics. These topics consist of a subtle blend of products, that only make sense when one considers regional diets. This idea is further confirmed when one looks at where do households with this diet live. Figure 7 gives a plot of where in France are topics 2 and 16 most represented - these topics include products that are typically associated with the Normandie region (cheese, cider, galette) and Alsace-lorraine(saucisse, pate a tarte, etc.). This concurs to the idea that diets have a highly geographical aspect to them, and people in different parts of the country do not eat the same thing. We can therefore see that there is more than a generational twist to food. Regions also matter, and once we take into account food everybody has to buy - water, baby food, etc., one can see such regional trends emerge.

2.3 Which socio-demographic features influence BMI?

Now we turn to our second question: namely who in France has the best/ worst eating habits. Here, the BMI will be taken as a very crude proxy for the quality of the diet. This question arises naturally given the data with which we work. Or in other terms, do people with higher BMI, and thus health risks typically cluster in certain subspaces of the sociodemographic space. The approach we tried here is a multivariate regression, or in other words trying to predict BMI using socio-demographic features in a linear model. Let us note that within each household we only chose the two heads - man and woman, and discarded all information about other people. So each household can be cast as at most two data points. Furthermore, these two people from the same households will share a vast array of features, the only ones that will note be shared are education, gender, employment, age.

We ran regular Linear Regression. We obtained a score of explained variance of 6%, which was robust when run on a held out dataset. Although the score is low, the features that have significant coefficients remain highly informative and interesting. Figure 8 gives a summary of our results.

Negative coefficients BMI

variable	coefficient	std	T-stat	low	high
rve	-0.0981	0.021	-4.638	-0.139	-0.057
proprio	-0.4219	0.107	-3.941	-0.632	-0.212
etude_5.0	-0.751	0.143	-5.245	-1.032	-0.47
etude_4.0	-0.762	0.134	-5.69	-1.025	-0.5
etude_6.0	-0.9506	0.167	-5.676	-1.279	-0.622
dpts_44	-1.0705	0.28	-3.828	-1.619	-0.522
dpts_56	-1.1218	0.312	-3.601	-1.732	-0.511
etude_7.0	-1.151	0.15	-7.675	-1.445	-0.857
dpts_81	-1.3529	0.355	-3.809	-2.049	-0.657
etude_8.0	-1.3655	0.168	-8.116	-1.695	-1.036

Positive coefficients BMI

variable	coefficient	std	T-stat	low	high
csp_Personnessansactivitéprofessionnelle	3.2821	0.134	24.456	3.019	3.545
csp_Anciensemployésetouvriers	2.9923	0.132	22.663	2.734	3.251
csp_Ancienscadresetprofessionsinterm	2.9836	0.142	21.023	2.705	3.262
csp_Anciensartisans	2.9056	0.234	12.399	2.446	3.365
csp_Anciensagriculteursexploitants	2.1129	0.428	4.937	1.274	2.952
csp_Chômeursn'ayantjamaistravaillé	1.8122	0.446	4.067	0.939	2.686
csp_Chauffeurs	1.3864	0.336	4.132	0.729	2.044
csp_Elèves	1.107	0.218	5.076	0.68	1.534
en6	0.4043	0.076	5.329	0.256	0.553
aiur_espacedominanterurale	0.3363	0.074	4.528	0.191	0.482
tvc1	0.2784	0.024	11.481	0.231	0.326
chie	0.1556	0.031	4.959	0.094	0.217
cha	0.1104	0.023	4.825	0.066	0.155
age	0.0462	0.003	17.331	0.041	0.051

Figure 8: Coefficients output by a Linear Regression of BMI on socio-demographic variables

The coefficients that dominate are the ones heavily related to lifestyle in general - age, revenue, education. We can also see that some specific departments in France have an impact, this is related to our idea that some regional diets survive and heavily influences how people eat across age and class. Finally, we note some more peculiar variables - such as the presence of a TV-set. This is most probably not by itself causing a higher BMI, than the inactivity it is often correlated with.

Let us note that beside this approach, we tried a more holistic methodology, by projecting our households into a lower dimensional socio-demographic space, visualizing the households in 2-D and how they interact in this space, and then seek if there are clusters of high / low BMI within this socio-demographic space. Sadly, this approach did not yield very intuitive results. One explanation is that BMI is an imperfect measure of diet quality, with a lot of noise. Even the regression, only captures 6% of explained variance.Thus, any attemps at eye-balling structurein the relationship between BMI and socio-demography is poised to fail.

3 Phase II : Relationship between products and BMI

A natural subsequent question is to explore further how much of BMI we can explain. So far, we have used only socio-demographic data, but we also have data about eating habits. How informative is the conjunction of the two? How much of BMI can we explain? And is it possible to isolate effects of unique variables, i.e. products?

Before further delving into the issue, it would be useful to have a framework in mind of what we are trying to achieve, and of which variables do influence BMI. Figure 9 presents a very crude sketch of the causal relationships determining BMI.

The graph is straightforward, our BMI is determined by how we eat, our lifestyle, and our metabolism. If we had all of the relevant variables, it could be easy to isolate the effects of each by controlling for all other variables, and indicate its contribution to overall BMI. Here, it is not the case, we only have partially observed information for lifestyle and food, and no information about

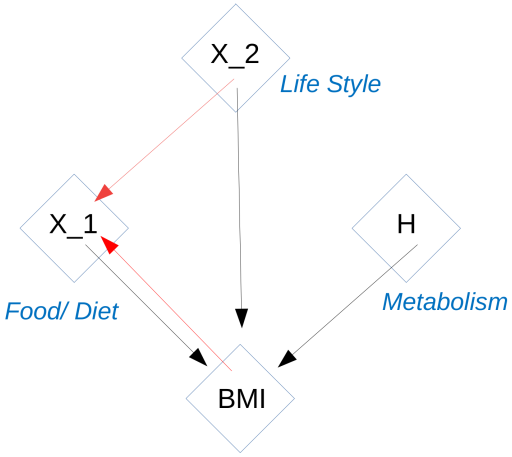


Figure 9: Causal graph of BMI

the metabolism.

This brings up confounder issues as well as inversed causality issues - red arrows. For the confounder issue, this is most represented by the red arrow going from lifestyle to eating habits. The idea is that some of the lifestyle variables that do influence the BMI and that we do not have also influence eating habits, and thus through eating habits will appear to influence BMI. For example, if all people who did sports bought sport drinks, sport's effect on BMI would be entirely captured by sports drinks, and it would seem that sports drinks have in themselves all the beneficial aspects sports brings to BMI.Another case that is of interest to us is reverse causality. Is it that fat-free yogurts cause obesity? One would argue no, it is the opposite way around. Because the only people that buy them are already overweight , it is overweight / BMI that causes their consumption.

How do we control for these two facts? We did not find a solution for the confounding issue. Since we do not have many relevant data concerning the lifestyle of people in our sample, the results we will obtain have to be taken with a grain of doubt, since always a confounder can be lurking in the back and explaining the entirety of our results. For the second problem, about reversed causality, it is possible to test statistically each direction of the causality, and see which one is most probable from a statistical point of view. For lack of time we did not explore this line of research very far, but we know effective models exist to do so and determine quite accurately the direction of causality.

What we can do, and we did is try to answer the following questions: 1- with our variables, and by training the best model we can, how much of BMI can we explain? 2- can we isolate effects of specific products, by controlling for the variables we do have?

3.1 Multivariate regression

A natural next step is setting up a supervised task to explore the extent to which a model with our available variables can capture BMI. The features will be the consumption data, and in the target will be BMI. We tried also a classification - i.e. predict if a person is underweight, normal weight,

features	add socio	preprocessing	predict_worse	extreme_only	threshold	var_expl train	mse_train	var_expl test	mse test	examples
socio-eco			False	5	300	0.0663	19.40	0.0591	19.54	30383
socio-eco			True	4.6	300	0.0987	20.17	0.0890	20.39	19658
raw	False	normalize	False	5	3000	0.2434	16.43	0.1117	18.47	30383
raw	False	binarize	True	4.6	3000	0.3229	16.48	0.1357	19.42	19658
raw	True	normalize	False	5	3000	0.2612	15.97	0.1344	18.01	30383
raw	True	normalize	True	4.6	3000	0.3130	16.43	0.1641	18.84	19658

Figure 10: BMI Regression results

overweight, obese - as well as a ranking task - predict of two households which one has higher BMI. Here we only show the results for regression, since the two other tasks give similar results and did not bring much value to the analysis. We used two evaluation metrics to tune our model: explained variance – the ratio of variance that is explained by our model, and Mean Squared Error – how much our prediction is off from the true value of the BMI. We mainly used linear models with heavy regularization to predict BMI.

As a note, from a technical perspective we performed L-2 penalized regression, with 5-fold cross-validation, and we report here mean performance on the test sets with the best hyper-parameters.

Figure 10 presents our results. We tried to improve the performance of our model by solving problems we noticed with our data. The problems we have encountered are mainly about quality of the data : eliminating data points for which we do not have enough information, i.e. with not enough purchases, by using entropy as a measure of completeness of our data,the threshold column, meaning the cutoff threshold below which we discarded households as having not enough data information. Also because in households with more than on person, it is impossible to disambiguate what each person has effectively eaten, we only tried to predict the worse BMI in the household $here the column of interest is predict_worse$. This approach can certainly be criticized, but our objective is to try to detect mostly harmful products. Therefore, such an harmful product is more likely to have its impact on the person with the worse BMI.

Nevertheless, the overall performance remains low. Below are a few elements that could explain the low results we have obtained :

- the absence of important determinants of BMI – smoking status, sports, etc
- the quality of the data: again this dataset was not set up with a nutritional perspective in mind. We only have what people have purchased as food for home, not what food they have eaten at friend's, or in restaurants. Moreover, many households have only participated to part of the study, and thus we have somewhat unreliable data for them.

We remark nevertheless several things. First, diet information encapsulates more explanatory power than socio-demographics. Furthermore, we notice that the two information are not redundant, i.e. the whole of sociodemographic is not included in the diet, nor the opposite. But there

is some overlap over the two since the model only improves marginally when both type of features are combines. As a note, we can remark that the model behaves significantly better in this aggregated representation, than if we kept the full initial representation with 170,000 products. This is explained by the power gained from aggregating similar features in such a context of curse of dimensionality – where the number of dimensions is significantly higher than the number of samples.

3.2 Coefficient Analysis

It is further interesting to take a closer look at the coefficients in themselves. Which products most lower or higher BMI? Can a look at this products inform us on what our model learns?

Below are the top influencing products obtained from the regression that takes into account all relevant sociodemographic data. These results are supposed to control for confounding variables such as age, wealth, and education. Nevertheless we notice some results that seem old. A first example of such an odd result is Champagne, or LEffe Ruby beer. One can hardly argue that these products have an inherently lowering effect on BMI. What we notice here seems to be some residual correlation, between age and BMI - usually young people buy leffe ruby beer, and between wealth and BMI, that was not eliminated by our introduction of sociodemographic variables. This can be explained by several facts. Either our sociodemographic data is too noisy, or we do not have enough data, and our model still captures noisy information. We also have some other byproducts, for instance products that are specifically for newborns are reported to have a negative impact on BMI. One can not easily argue that these products have a direct impact on parents, since they obviously do not consume them, on the other hand, they are the sign of a new baby in the family, a phenomenon which is frequently associate with some gain in weight for the mother - what is also known as 'baby weight'. There are probably many more confounding variables of this sort that are at play here. We also find our previously mentioned fat losing products - weightwatchers in the positive side.

Figure 11 presents the value of the products with the highest coefficients. These results are intuitive and consistent with prior knowledge. But one fails to see any pattern in the data, in this form. Therefore, we would argue that we need to go further, and seek a more robust and wholesome analysis.

3.3 Identification of outstanding cluster of products with a high influence on BMI

From a technique we will introduce later, we clustered products in consistent and intuitive categories - using the purchase patterns found. We then tried to see if the clusters of products we found from this technique had somehow a correlation with high or low BMI. And we found two outstanding such cluster of products: one going in each direction. Organic food consumers tend to have better BMIs than their counterparts. On the other hand, meat consumers tend to have higher BMIs. Below are cumulative BMI plots for each category, comparing the population at hand with a random sample of the complementary population to better understand the effect of each. Results can be found in figure 12.

These results have been controlled for age, education, and income : the major factors of influence we have noticed from the previous part. As a note, the difference in the curves are as important as

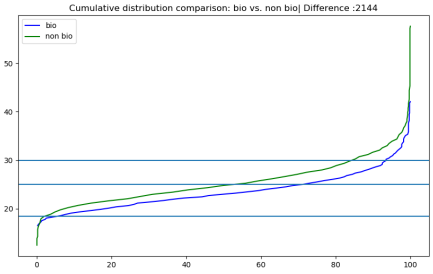
Negative coefficients BMI

groupe	sousgroupe	marque	count	coefficient
Plat	Plat Frais	SOJASUN	1576	-111
Boisson	CafÃ©	PLANTATION	500	-74
Plat	Plat SurgelÃ©	TANTE YVONNE	858	-73
Fruits et Legumes	LÃ©gumes Sec	NOTRE JARDIN	956	-72
Boisson	Champagne	Marque Non Trouvee	527	-69
Plat	Plat Frais	SOY	435	-69
Confit	Confit en conserve	LARNAUDIE	459	-67
Fruits et Legumes	Fruits Frais	Marque Non Trouvee	17545	-67
Aide Culinaire	Produit sucrant	CARREFOUR.DISCOUNT	4995	-65
Cereale A Cuire	CÃ©rÃ©ale A Cuire	BJORG	659	-65
Boisson	BiÃ¨re	LEFFE.RUBY	968	-64
Plat	Salades en conserve	PECHE OCEAN	557	-63
Fruits et Legumes	Legumes Frais	SAVEOL	3757	-61
Plat	Plat Frais	Marque Non Trouvee	907	-61
Boisson	CafÃ©	Nescafe special filtre	1054	-61
Boisson	Champagne	Marque Non Trouvee	1108	-56
Boisson	Infusion	Elephant nuit tranquille	1309	-56
Pain et viennoiserie	Panification SÃ¨che	BJORG	518	-56
Aide Culinaire	Lait De Coco	SUZI WAN	1580	-55
Boisson	Infusion	LEA NATURE.JARDIN BIO	754	-54
Fruits et Legumes	Legumes Frais	Marque Non Trouvee	14193	-54
Biscuits	Barre CÃ©rÃ©alière	BRIN DE JOUR	676	-53
Aide Culinaire	Vinaigre	ECO+	4625	-51

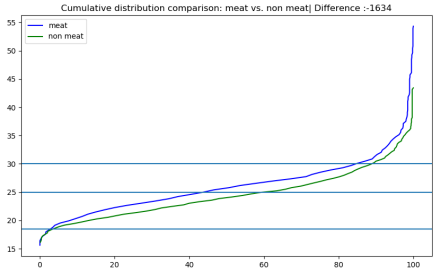
Positive coefficients BMI

groupe	sousgroupe	marque	count	coefficient
Aide Culinaire	Produit sucrant	CANDEREL	614	112
Charcuterie	Autre charcuterie	Marque Non Trouvee	28202	96
Jambon	Jambon Blanc	U	1496	95
Viande	Porc	Marque Non Trouvee	1669	95
Charcuterie	Lot Mixte PÃ¢tÃ©	MONTEROY	577	87
Boisson	Sirop	PULCO	2598	87
Charcuterie	Fromage De TÃªte	RANOU MONIQUE	245	86
Assaisonnement	Poivre	CIGALOU	258	86
Sauces	Sauces	RUSTICA	262	83
Aide Culinaire	Huile	BOUTON D OR	346	82
Aide Culinaire	Margarine	ST HUBERT.OMEGA 3	7475	76
Plat	Plat Frais	FLEURY MICHON	815	75
Fromage	Fromage camenbert	REO	972	73
Plat	Plat Frais	TANTE YVONNE	563	73
Boisson	Lait	NESTLE	959	73
Biscuits	Biscuit ApÃ©ritif	SUZI WAN	440	71
Viande	BÂ□uf	Marque Non Trouvee	4959	71
Pain et viennoiserie	Viennoiserie	PATIGEL	133	69
Charcuterie	Saucisse	Marque Non Trouvee	6021	69
Fruits et Legumes	Fruits et lÃ©gumes en Conserve	Marque Non Trouvee	274	68
Dessert	Dessert Frais	WEIGHT WATCHERS	1089	64
Boisson	Vins	Marque Non Trouvee	386	61

Figure 11: Coefficients of the socio-demographic regression



(a) BMI Cum. distrib. of bio vs non-bio



(b) BMI Cum. distrib. of meat vs non-meat

Figure 12

the one we observed for age, the single most influential feature in determining BMI from sociodemographic data. These results are in keeping with previous result linking organic products and lower BMI indexes [3]. Indeed, Kesse-Guyot et al have found even when controlling for a host of confounding variables – including smoking habits, physical activity, and quality of the diet. Other studies [**?**] have also confirmed higher BMIs associated with heavy meat eaters, the difference being attributed to macro nutrient intake differences.

3.4 Comments and further analysis

The first thing we have to note is the problem of confounding variables. Here we only controlled for age, education and wealth. But there are a host of other confounding variables which have a high impact on health and overweight. For instance smoking habits, or physical activity. Therefore, our results remain incomplete and cannot confirm or infirm anything - indeed it is possible that for instance organic food eating is highly correlated with high physical activity, and thus it is that physical activity that explains lower BMI, and not some intrinsic property of organic food. Our study can therefore only point into directions, and orientate further research. It is not possible for us to affirm that a product has intrinsically an effect on BMI, since we lack very important confounding variables, but we can nevertheless identify suspects, whose influence further research must confirm.

Finally, another thing we should note is that it seems possible to do much more if we are provided more data. Many clusters of products may have a significant impact on health, but this impact is dampened by problems inherent to our data: incomplete records of purchases, ambiguity of the diet in a household with more than one person. Luckily, one solution to all these problems is to increase the size of the data set. Thus, as we will handle more data, we will have more chances to identify clusters of suspect products.

4 Phase III : Exploring the possibility to build a recommender system for better eating

Our products are for the moment represented as one-hot vectors of dimension 4130. Although we have drastically reduced the space of products, we still have not captured all interactions among products. One senses that these products are not independent with respect to one another. Some of them share a lot of similarities, and can be considered synonymous, nearly replaceable. Salted butter versus unsalted butter, or blond beer versus brown beer. Others share a weaker link, but are still somehow related. Wine versus beer. Apples and oranges. Others are indeed not related at all. How to capture this tight web of relationships?

There are many ways to deal with this kind of problem. The main idea being to embed the products into a latent space, where closeness of products in this space – as measured by cosine distance, is tantamount to similarity of the products in prior knowledge.

In order to go about finding this latent space, one usually uses the insight that it is possible to infer the 'meaning' or the 'composition' of a products by seeing with which other products it co-appears. For instance fat-free milk and normal milk will both be consumed in combinations with cereals and cookies. Therefore, they should be close together in the embedding space. The context of consumption of products gives us a lot of information about the product.

Here we mainly used one technique that fits surprisingly well the context of our data- the Glove technique, whose characteristics we shall discuss next.

4.1 The Glove model

The Glove model [4] was introduced in order to improve word embeddings. It is based on this linguistic intuition of trying to know a word's meaning by the company it keeps. This model will endeavor to bring closer together products that appear in the same context. Two products which are often bought with the same other products must be close in meaning, or substance.

Glove works with the matrix of co-occurences. Given a matrix of co-occurences – of products x products, where each entry corresponds to how many times a product has been bought with another product. The Glove algorithm will try to find an embedding into a smaller dimensional space that best captures and reproduces the initial co-occurence matrix.

One advantage with the Glove technique is that it is highly scalable. One can easily increase the granularity of the data without losing to exploding memory or computing complexity – the algorithm being linear in its input. So whereas before we have worked only at the aggregated yearly level, where one row in our matrix was the yearly purchases of a household, now we shall work at the basket level. Each line will be a basket of goods. This purchase matrix will give way to a co-occurence matrix, that will count in effect how many times each product has co-occured with each other product in a basket. And it will be this resulting matrix that we will try to approximate using Glove.

4.2 Illustration of results

Using the visualization technique that projects high dimensional data into a 2-dimensional plot – TSNE [1], we can see the results more clearly. Below is the resulting plot, giving intuitive results. We commented the plot lightly to give an idea of how the products relate to one another in this model. These results can be seen inf figure 13.

We can see that this embedding gives the most intuitive representation of products. Product categories are vastly respected, products close to one another tend to belong to the same category - here have the same color. This is confirmed by a closest analysis we ran on random products, see in figure 13 for a random subset of products and their closest neighbor in the glove space. The 5-closest products are very intuitive and confirm our intuition that GloVe space has managed to capture a great deal of the meaning of a product. It is important to note that beforehand no notion of product category or any other product attribute has been fed into Glove. The only thing the model knows is with which other products one product tends to co-occur.

There are a few zones of interest in figure 12. There is a sea of spaced products in a diagonal ray going through the center of the figure: these are products for which we do not have sufficient data, i.e. products that have not been bought a sufficient number of times. Furthermore, the products of all colors beneath this ray are all organic products. The common characteristic that binds all these seemingly different products together is the fact that they are organic. Which makes sense since often organic products are sold together, and since people who buy organic tend to only buy organic.

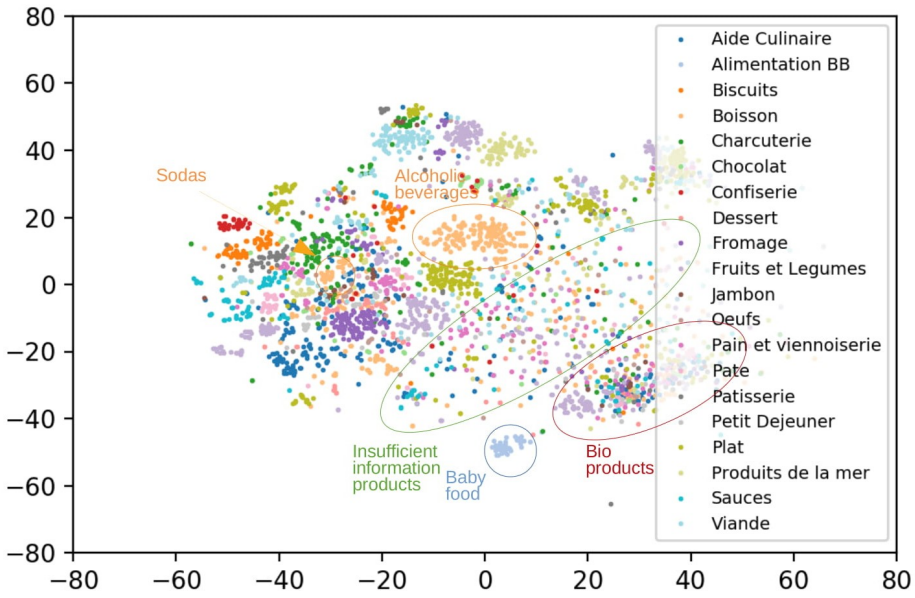


Figure 13: 2D representation of the Data in the GloVe Space

Product	SEVEN UP| Brsa|Non|Non| Non|nan| Limonade-Lime|	CASINO| Viennoiserie|Oui| Non|Croissant|	LA ROCHE MAZET| Vins|Non|Non|nan| Bouteille|	SAINT JEAN.TRADITION GOURMANDE| Pate_pate|Non| Non|Raviole|	NESTLE.ECLAT NOIR| Tablette|Non| Non|Noir| SucrÃ©|
Neighbour 1	JOKER.LE FRUIT|Brsa| Non|Non|Non| A Base De ConcentrÃ©| Nectar-Jus de Fruits|	COTE DOR.MIGNONETTE|Tablette|Non|Non| Noir|SucrÃ©|	Marque Non Trouvee|Vins|Non| Non|Autre|Bouteille|	SAINT JEAN.TRADITION GOURMANDE| Pate_pate|Non| Non|Raviole|	NESTLE| Tablette|Non| Non|Lait| SucrÃ©|
Neighbour 2	HERTA| Lardons|Non| Non|Frais| FumÃ©|	CASINO|Pain|Oui| Non|Viennois|	Marque Non Trouvee|Vins|Non| Non|Bordeaux Aoc| Bouteille|	ST JEAN| Pate_pate|Non| Non|Raviole|	PURE VIA| Produit sucrant|Non| Non|StÃ©via|
Neighbour 3	COCA COLA| Brsa|Non|Non| Non|nan|Cola|	Marque Non Trouvee|Whisky-Bourbon|Non|Non| Pur Malte|Whisky|	Marque Non Trouvee|Vins|Non| Non|Autre|Bouteille|	ECOCHARD|Plat Frais|Non|Non| nan|Quenelle|	FLEURIER LE|Margarine| Non|Non|nan|
Neighbour 4	ORANGINA| Brsa|Non|Non| Non|nan| Boisson Aux Fruits|	LEADER PRICE| Plats cuisine en conserve|Oui|Non| Corned Beef|	CARREFOUR|Vins| Oui|Non|Autre| Bouteille|	SAINT JEAN| Pate_pate|Non| Non|Raviole|	KAMBLY.FLUTES|Biscuit ApÃ©ritif|Non| Non|FlÃ»tes|
Neighbour 5	FINLEY|Brsa| Non|Non|Non| nan|Boisson Aux Fruits|	CARREFOUR.KIDS|Biscuit SucrÃ©| Oui|Biscuits|	CARREFOUR|Vins| Oui|Non|Autre| Bouteille|	CARREFOUR| Fromage|Oui|Non| ComtÃ©|	SOIGNON| Dessert Frais| Non|Non|nan| nan|Yaourt|

Figure 14: Sample of products and their 5-closest neighbors

Thus, building on this very interpretable representation of products we tried the following experiments : given a clustering of products in this space, a clustering which is bound to give meaningful categories, are there clusters that have a more significant impact on BMI? I.e. do people who tend to purchase more products from one of these categories have significantly higher or lower BMIs than the mean population? By running this analysis, we found the categories we discussed before: meat eaters and organic food eaters, along with other categories that may or may not deserve further inspection - again a confounder issue.

4.3 Caveats and further directions

A first caveat is that our representation is highly dependent on our clustering of products. The more clusters we have, the more diluted our data is. In our current representation, we have 1400 clusters out of the 4130 total clusters - so approximatively one third that are in a lost sea inside our space - that are close to no other products and seem lost in a sea of missing information. A more stringent clustering, regrouping possibly products which are infrequent more thoroughly can help us gain insight into a wider range of products. Another problem we had is that some clustering of products which were less stringent - with approx. 22, 000 clusters of products created Distributor universes, i.e. products that were close to one another only because they were sold by only one retailer. This is not desirable however in our study, since we want results that are robust to such noise.

In addition, a nice future perspective for this study that we did not have time to implement is the idea of a Gaussian LDA [**?**]. The idea is to perform LDA not anymore on one-hot vectors - i.e. the products as they are represented by an index, but on the sparse representation of products in some latent space - here the Glove space. This would help the model generalize better, and have more leeway in the topics it chooses. We feel that this approach would greatly help improve power of LDA, which currently suffers from generalization issues.

5 Conclusion and further directions

5.1 Results of the analysis

This analysis has already been able to provide insightful results. The first direction we explored in this study is the relationship between social condition and eating habits / general health. We have found a couple of factors influencing eating habits, with a major factor being geography. As for bad eating habits/ or bad BMI, we have found that Age, education, and wealth constitute important predictors for BMI. One can hardly argue that as we age we become more sloppy in the way we eat, this part being probably more due a slow-down of the metabolism. Nevertheless the effects for Wealth and Education remain relevant.

Then we tried to explore the relationship between eating habits and Body Mass Index. Are certain types of food bad for health? Are certain diets increasing BMI? And finally how much of BMI can be explained by our eating and social condition. BMI is a complex and often inaccurate metric, which depends on a host of factors and can be quite volatile across time. Nevertheless we found that with our imperfect data - low quality data, low number of data points, and ambiguity in the diets of each household, we could explain 13% of the variance of BMI with only purchase data, and 17% with purchase data and demographic data. Furthermore we found that certain clusters of types of food have more influence on BMI that others. We found that organic food eaters tend to

have lower BMI than the rest of the population, and that meat eaters tend to have higher BMI - data that remains true when we control for the confounders we have. This data seems to confirm previous studies which go in the same direction [3] [**?**].

Finally we tried to build an ersatz of recommender system. The idea here is to recommend to people products that fit into their diets, yet that are healthier. The need is here to build an approach which takes head both of the impact of the product on BMI, and also of the taste of the people buying it, and of the place of the product in the consumption space of the population. To achieve such a goal we implemented a Glove algorithm [4] on our purchase data. This algorithm provided us with a powerful embedding, which yielded highly interpretable and intuitive results. Combined to our previous results on BMI sparse coefficients, and possibly prior knowledge, one can possibly start to see such a recommender system be put into place. Let us note that this embedding played a key role in identifying the organic and meat clusters as having a notable influence on BMI.

5.2 Further directions

This data poses many interesting questions from a Machine Learning perspective. These questions arise from the quality of the data itself - which is suboptimal for the problem at hand.

The first question is about the clustering of products. Is it possible to achieve a clustering of products, fusing several products into one, because of some commonality of composition to achieve better BMI predictions. Our clustering is mostly hand-made, and uses external data. Although we tried, we did not find an algorithm capable of bettering this hand made clustering.

Another question the data asks is: is it possible to disambiguate diets within a household. As such we have the purchase data at the granularity of the whole households. We do not know what each member of the household has eaten individually. This most likely hindered our performance in the BMI prediction task. One approach we tried to disambiguate diets is to use domain adaptation, and feature augmentation, a technique introduced by Daume III [**?**]. The idea is to augment the feature space, introducing copies of each feature, making it possible to determine is the purchase of one product has a bigger influence on males or females. If it has more influence on one that the other it can mean two things 1) the product is typically consumed by one gender only, and thus it affects only the member of that gender in the household 2) the product affects people differently according to gender. If one dismisses the second case as peripheral, this can be a nice approximation of diet disambiguation. We leave it to further studies to explore more in depth this question.

5.3 Data it would be nice to have in the future

5.3.1 BMI over time

One thing that could help us override a lot of confounders in our analysis of impact of specific products on BMI is to study rather than raw BMI, deltas in BMI. The idea is therefore to have for each person two data points of BMI. One at the beginning of the year, and one at the end. This would help us better dampen the effect of metabolism, by accounting for some base BMI level.

5.3.2 Nutritional information for products

A big handicap of this data is that it does not provide nutritional information about the products it features. This is understandable when one bears in mind the goal Kantar Worldpannel had in

mind when collecting the data. Nevertheless nutritional information about the products would be of great help 1- to better cluster products 2- to run an analysis on the influence of diet from a nutrient perspective on health 3- to control for possible confounding variables in our findings.

5.3.3 More data

We understand that the above points may be hard to put in practice. Nevertheless the following one seems straightforward. One problem we ran into when performing our analysis what that the data at hand was not big enough. Our algorithms require a lot of data to be able to detect patterns, and as such would benefit from an increase of data. Having the same information, i.e. socio-demographic and purchase for other years can only improve our analysis, and yield better and more precise results.

References

[1] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[3] E. Kesse-Guyot, J. Baudry, K. Assmann, M. d. P. Galan, S. Hercberg, and D. Lairon, "Prospective association between consumption frequency of organic food and body weight change, risk of overweight or obesity: results from the nutrinet-sante study," *British Journal of Nutrition*, pp. 325 – 334, 2017.

[4] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.