Towards causal modeling of nutritional outcomes

Ksenia Gasnikova KSENIA.GASNIKOVA@INRIA.FR
Philippe Caillou CAILLOU@LISN.FR
TAU, CNRS – INRIA – LISN, Univ. Paris-Saclay, 91405 Orsay France

Olivier Allais OLIVIER.ALLAIS@INRAE.FR
ALISS, INRAE, ALISS, Univ. Paris-Saclay, 94205 Ivry France

Michèle Sebag SEBAG@LISN.FR
TAU, CNRS – INRIA – LISN, Univ. Paris-Saclay, 91405 Orsay France

Editor: Sisi Ma, Erich Kummerfeld

Abstract

This paper aims at observational causal modelling, investigating the causal relationships between food consumption and health status, exploiting the proprietary Kantar database. This database describes the socioeconomic characteristics and consumption habits of a few dozen thousands households; in particular, the consumed food items are documented almost at the level of precision of barcodes.

A first challenge for this observational causal study lies in the number of hidden confounders, ranging from genetic factors to life styles (i.e. smoking and sport habits), not documented in the data. Taking inspiration from the *Deconfounder* approach (Wang and Blei, 2019b), substitute hidden confounders based on dietary patterns – viewed as characteristics of the alimentary lifestyle – are extracted from the database and exploited to block the biases due to hidden confounders.

A second challenge lies in the fact that the data size hardly allows for investigating a number of fine-grained interventions. We thus define a new type of intervention, enabled by the data structure and referred to as *macro-intervention*, acting on the full basket of food items; an example of such macro-intervention is to replace every non-organic product in a household basket with its organic counterpart. The average treatment effect of this macro-intervention is assessed in the context of the substitute hidden confounders, using inverse propensity weighted estimates to control for covariates such as wealth or education.

1. Introduction

The presented study aims at assessing the causal impact of food on health. The long term goal of this study is to support recommendations meant to decrease health hazards, e.g. regarding Type 2 Diabetes.

The domain knowledge, based on experiments in vivo (e.g. concerning the "Western style" of nutrition (Hasegawa et al., 2020)), does not deliver fine-grained results, due to the combinatorial explosion of the food item ensembles involved in a diet. Multiple randomised controlled trials (Reeves et al., 2018) need to consider very large cohorts; additionally, the set of food items keeps increasing and evolving, severely limiting the building and maintenance of background knowledge regarding the causal impact of food on health. The

© K. Gasnikova, P. Caillou, O. Allais & M. Sebag.

alternative explored in the paper is thus based on observational causal modeling, exploiting the proprietary observational Kantar database. This database describes the sociological features and consumption habits of a few dozen thousands households over 20 years. A single feature related to individual health is available, the body mass index (BMI); this feature is commonly considered a relevant health risk indicator (McGee, 2005; Flegal et al., 2013).

With respect to observational causal modelling (Pearl and Mackenzie, 2018; Peters et al., 2017), the challenge is threefold:
A first challenge regards the number of potential causes. The state of the art in epidemiology suggests that similar food items might have very diverse impacts on health (Blundell et al., 2005; Stöger, 2008).[1] For this reason, a coarse-grained representation of consumption habits, e.g., in terms of amount of lipids, proteins and glucids ingested, is bound to miss the point. One must take into account to the best possible extent the level of detail available in the Kantar database, where purchased food items are represented almost at the level of precision of barcodes; eventually, the number of potential causes is circa four thousands (Section 2). On the top of this large number of potential causes, the Markovian assumption − that every potential cause acts independently of the others − hardly holds, as different food items might share same components, e.g. gluten.
The second challenge regards the plausible type of causal mechanisms. While linear mechanisms (Spirtes et al., 2000; Shimizu et al., 2011) are widely used when considering a high number of potential causes, the effects of food items on health hardly are linear in the consumption quantity; and "cocktail effects" are notorious, that is, the effects of joint causes differ from the sum of cause effects.
The third challenge concerns the presence of external hidden confounders. No assumption regarding the lack of hidden confounders with respect to the selected household cohort can be made: typically, the sport or smoking habits of the household members are unknown, though these factors notoriously impact both the individual food consumption and the BMI. The presence of unobserved confounders however makes it impossible in general to assess the effects of interventions (Pearl, 2009).

In order to address these challenges, the proposed approach referred to as *Deconfounded Macro-Interventions* (DeMaIn) takes inspiration from the *Deconfounder* approach proposed by Wang and Blei (2019a,b) (Section 3). So-called substitute hidden confounders (SHCs) are extracted and used to block the impact of the unobserved confounders. Informally, the *Deconfounder* states that the hidden confounders are constant when SHCs are constant; this makes it possible to locally assess the average treatment effects of interventions without incurring the biases due to the impact of the hidden confounders, conditionally to the stability of the SHCs. In the consumption context, the extracted SHCs are defined based on dietary patterns, representing at a high-level the diet "topics" involved in the nutritional household style. For instance, three (excerpts of) dietary patterns are: i/ wine, aperitives and biscuits; ii/ butter, cream, desserts; iii/ baby food (more in Appendix).

1. For instance, the *pizza* term encompasses 392 different items divided into 8 categories with highly diversified nutritional properties and dramatically different impacts on health.

A main contribution of DeMaIn lies in the type of interventions considered. The high number of potential causes and their nutrition-wise likely redundancy makes it unrealistic to make interventions on single food items, all the more so as the amount of data does not provide the statistical power to assess many such interventions. The proposed alternative is based on the definition of few, high-level interventions, made possible by the structured description of the food items (Section 2). Specifically, a macro-intervention globally modifies the whole food basket consumed by a household *while preserving the amount of calories ingested*. An example of such macro-intervention, that will be considered in the experimental study, is to replace all food items in the household baskets with their organic counterpart, by switching the "organic" property of the food items to *True* or *False*.

In order to estimate the macro-intervention impact on BMI, the methodology consists of defining treatment and control groups associated to the SHCs, and testing these groups for distribution biases w.r.t. covariates (e.g., wealth, education, urban/rural location). The average treatment effect (ATE) of the macro-intervention is eventually estimated using an Inverse Propensity weighted score (IPW) approach (Austin, 2011), based on a calibrated classifier among the treatment and control groups, to control for the sociological covariates.

The DeMaIn approach differs from the *Deconfounder* in two ways:
On the one hand, the *Deconfounder* aims to investigate the impact of intervening on a single variable (e.g. the impact of a single gene in a GWAS context, the impact of a given actor playing in a movie); it ideally aims to estimate the counterfactual effect of modifying this single variable (the individual treatment effect, ITE) in a given context (e.g. within a narrow region of the space defined after the SHCs). In contrast, DeMaIn estimates the counterfactual effect of intervening on a group of variables, focusing on macro-interventions defined from prior knowledge.
A second difference is that DeMaIn only aims at estimating the average treatment effect: the individual treatment effect is considered to be beyond reach given the *known unknowns*. The unbiasedness of the ATE estimate is sought *a priori* through the identification and exploitation of SHCs, and *a posteriori* using an IPW approach.

The paper is organized as follows. Section 2 describes the Kantar database and its specifics. For the sake of self-containedness, Section 3 presents the *Deconfounder* approach; its merits and limitations are discussed with respect to the state of the art and in the particular context of the Kantar database. The DeMaIn framework addressing these limitations is presented in Section 4. Empirical results are described in Section 5 and the paper concludes with a discussion and perspectives for further work.

2. The Kantar observational database

The data used in the study have been gathered since 2008 by Kantar Worldpanel. As might have been expected, these data have not been gathered for the purpose of the current study, relating food consumption and health status. The design of the panel (selection of the households and the features) has been motivated to conduct marketing studies, focusing on the identification of consumption patterns, of brand market shares, and of potential substitutions among products.

Three tables are considered in the following: the Household table includes circa 25,000 households; the Food items table contains 170,000 items, without any nutritional information; last, the Purchase table includes 10,400,000 transactions, where each transaction reports the date, the household, the food item purchased and the quantity. We restrict ourselves to the 2014 database; the exploitation of the longitudinal information is left for further work.

The outcome. As said, the outcome variable is the BMI, computed from the weight and height of each household member.

Potential causes (food items). Food items are described based on their nature, brand and packaging. In a pre-processing step, food items only differing in their packaging (e.g. Coca-Cola in 1 liter or quarters) are merged. A shortcoming of the data is that the purchase database only reports if and when a food item has been purchased. Whether the food has actually been consumed and by which household member, is unknown. How to handle this issue, e.g. using deconvolution operators to actually separate the male/female, adult/children consumption, is left for further work. For simplicity, we shall speak of *food consumption* instead of *food purchase* in the remainder of the paper.

The amount of consumption is normalized to handle the differences of scales, e.g. between the amounts of water and meat consumed in the household. Specifically, the yearly amount reported for a food item is divided by the number of persons in the household (where each adult counts for 1, a teenager for .7 and a child for .3 following (INSEE, 2019)) and the average amount of individual consumption for this food item in the database.

Structure on the potential causes. Depending on the category of food items, some continuous or categorical features are available, e.g. for most vegetables a feature indicates whether it is fresh or frozen; for most food items, a feature indicates whether it is organic or not; for wines the features include the type of wine (red, white or rosé), the degree of alcohol and whether it is organically grown or not; for cheese, the features include the provenance region and the degree of fat.

Covariates. Each household is described through 160 features, including the number of persons in the household, their age, weight, height, the salary of adults, their educational background, the urban/rural location of the household.

Hidden confounders. The BMI notoriously depends on many factors beside the food consumption, including (but not limited to) the level of physical activity, the smoking habits, and the genetics. These factors are *not* present in the covariates. The only assumption made in the remainder of the paper, following the *Deconfounder* (below) is the lack of confounders affecting both the BMI *and* a single potential cause, that is, a single food item.

3. Related work

Mainstream observational causal modelling approaches (Pearl, 2003; Pearl and Mackenzie, 2018; Peters et al., 2017; Colnet et al., 2020) appear to be ill-suited w.r.t. the considered observational dataset, due to the small number of samples (households) comparatively to the high number of potential causes (the food items), and the fact that linear causal mechanisms hardly reflect the impact of nutrition on health (e.g. due to cocktail effects).

As said, the proposed approach builds upon the *Deconfounder* approach (Wang and Blei, 2019a,b). This Section first presents the *Deconfounder* for the sake of self-containedness, and reports on its limitations and caveats discussed in the literature (D'Amour, 2019; Ogburn et al., 2020; Imai and Jiang, 2019; Grimmer et al., 2020). The limitations of the *Deconfounder* in the context of the Kantar observational data are thereafter discussed.

3.1. The *Deconfounder*

The *Deconfounder* aims to causal modelling in high dimensional domains, in presence of hidden confounders. For instance in genome-wide association studies, the outcome is affected by patterns involving many potential causes (the genes); furthermore, these causes are not independent, making it all the more difficult to establish causal models. Lastly, the outcome is known to depend also on (unknown) confounders: the lack of hidden confounders, a common assumption in the causal modelling literature, does not hold true in the considered application domains.

Formally, let Y, $\boldsymbol{X}$ and $\boldsymbol{U}$ respectively denote the outcome, the potential causes and the (unknown) multi-cause confounders (influencing several causes and the outcome).

The challenge is that these $\boldsymbol{U}$ cannot be uniquely identified. However, some approximations thereof, called *substitute hidden confounders* (SHC) and noted $\boldsymbol{Z}$, can be extracted. Specifically, the *Deconfounder* aims to find a probabilistic model explaining the patterns of co-occurrence of the causes and rendering the causes independent conditionally to $\boldsymbol{Z}$:

$$P(\boldsymbol{X}|\boldsymbol{Z}) = \prod_j P(X_j|\boldsymbol{Z}), \tag{1}$$

with X_j ranging in the n-size set of potential causes.

The standard ignorability assumption − the fact that the outcome only depends on the observed causes and treatments − is relaxed in the *Deconfounder* in the form of *single ignorability*, that is, the assumption that there exists no hidden confounder influencing both the outcome and a *single* cause.

Under the single ignorability and other mild assumptions (below), the *Deconfounder* proceeds by: i/ building the latent model M characterizing $\boldsymbol{Z}$ from the $\boldsymbol{X}$, enforcing Eq. 1; ii/ estimating the SHCs as a function of the $\boldsymbol{X}$s values:

$$\hat{\boldsymbol{z}}_i = \mathsf{E}_M[\boldsymbol{Z}|\boldsymbol{X} = \boldsymbol{x}_i]$$

iii/ estimating the average causal effect using the expectation of the substitute confounders. Letting $\boldsymbol{x}$ and $\boldsymbol{x}^{\emptyset}$ denote two causal patterns (n-dimensional vectors of $\boldsymbol{X}$ values), then:

$$\begin{aligned} \mu(\boldsymbol{x}, \boldsymbol{x}^{\emptyset}) &= \mathsf{E}_i[Y_i(\boldsymbol{x}) - Y_i(\boldsymbol{x}^{\emptyset})] = \\ &= \mathsf{E}_i[\mathsf{E}[Y_i|\boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{Z}_i = \hat{\boldsymbol{z}}_i] - \mathsf{E}[Y_i|\boldsymbol{X}_i = \boldsymbol{x}^{\emptyset}, \boldsymbol{Z}_i = \hat{\boldsymbol{z}}_i]], \end{aligned} \tag{2}$$

where i ranges over the set of samples.

Assumption 1: Consistent substitute confounder. The substitute confounder $\boldsymbol{Z}$ can be viewed almost surely as a deterministic function f_θ of the potential causes:

$$p(\boldsymbol{Z}|\boldsymbol{X} = \boldsymbol{x}) = 1_{\{\boldsymbol{Z} = f_\theta(\boldsymbol{x})\}} \tag{3}$$

Assumption 2: The overlap assumption reads:

$$p(\{\boldsymbol{X}_i\}|\boldsymbol{Z}_i) > 0, \tag{4}$$

for all subsets of causes $\{\boldsymbol{X}_i\}$ with non-zero probability.

3.2. Debate on the *Deconfounder*

The *Deconfounder*, generally viewed as a significant advance of the state of the art, generates an intensive debate about its assumptions. Several authors have been discussing its validity, with basically two types of arguments, related to the undeterminacy of the SHCs and the compatibility of both above assumptions.

Firstly, D'Amour (2019) notes that the non-uniqueness of the model M accounting for the causes distribution (Eq. 1) might result in an *ignorance region* where models fitting the observed data might yield different intervention estimate distributions. Another objection is that: i/ either the SHC $\boldsymbol{Z}$ is not determined from $\boldsymbol{X}$, and the undeterminacy of the intervention distribution might likewise follow; ii/ or, if $\boldsymbol{Z}$ is a function of $\boldsymbol{X}$, then the positivity assumption might be violated: in other words, Assumptions 1 and 2 above might exclude each other. Some ways to sidestep these limitations are through introducing proxy variables (also referred to as negative controls) in the estimation procedure, as shown by Louizos et al. (2017).

Athey et al. (2019) present empirical results suggesting that the *Deconfounder mostly helps when there are shared unobserved confounders*, and that instrumental variables should rather be used otherwise.

Imai and Jiang (2019) consider that the "single ignorability" assumption might require a significant expertise to be assessed; and that the overlap requirement can be stringent, suggesting that the use of parametric assumptions, or instrumental variables, or stochastic interventions, might be required to overcome this difficulty.

Grimmer et al. (2020) suggest that, in the linear-linear setting, a naive regression asymptotically produces the same result as the *Deconfounder*; in the finite sample case, the naive regression can be made to improve on the *Deconfounder* when some special care is taken to enforce the stability of the results.

The revised version of the *Deconfounder* discusses many of these critiques and take them into account (Wang and Blei, 2019b). In particular, the undeterminacy of the SHCs seems unavoidable; still, if the ignorance region (D'Amour, 2019) associated with the family of admissible SHCs results in a narrow confidence interval for the estimate of the intervention, then the approach does allow to overcome the obstacle of the hidden confounders (subject to the stability of the results w.r.t. the extracted SHCs).

3.3. The *Deconfounder* assumptions w.r.t. the Kantar data

The single ignorability assumption, requiring that no hidden confounder be controlling both a single cause X_i (here a food item) and the outcome Y, is considered plausible by domain experts.[2] Note however that this assumption cannot be tested empirically.

2. For instance, the smoking habit has an impact on the consumption of several food items (besides the BMI outcome), such as beer, wine, strong alcohol, chips and so on.

The overlap assumption, stating here that the probability of any food item is non 0 conditionally to a substitute confounder boils down to considering that any food item can possibly appear in any diet; likewise, this is considered to be acceptable by the domain experts.[3]

Another issue is whether the *Deconfounder* – meant to assess the individual treatment effect when replacing some conjunction of causes $\boldsymbol{x}$ with some $\boldsymbol{x}^{\emptyset}$ – addresses the dietary recommendation goals. The answer is *not exactly*, for two reasons. On the one hand, as said, the lack of critical confounding information (e.g. smoking or sport habits) precludes the estimation of individual treatment effects. On the other hand, such fine-grained interventions (changing $\boldsymbol{x}$ for $\boldsymbol{x}^{\emptyset}$) is beyond reach, for computational (considering any $\boldsymbol{x}^{\emptyset}$) and practical reasons (enforcing intervention $\boldsymbol{X} = \boldsymbol{x}^{\emptyset}$).

4. Overview of DeMaIn

This section presents the *Deconfounded Macro-Interventions* (DeMaIn) approach, designed to address the above mentioned limitations. The notion of *macro-interventions*, overcoming the low-granularity of the potential causes, is first defined. How to handle the hidden confounders in the same spirit as the *Deconfounder* is presented in Section 4.2. Last, the methodology proposed to build treatment and control groups associated to macro-interventions is detailed in Section 4.3.

4.1. Macro-Interventions

The original concept of macro-intervention (MI) is defined as follows, rooted in the definition of interventions as variables (Pearl, 2003)[chap 3.3.2]. Let σ denote an operator on the set of potential causes $\boldsymbol{X} = \{X_1, \ldots X_D\}$, mapping X_i onto $X_{\sigma(i)}$. Furthermore σ is assumed to be idempotent (if $\exists j$ s.t. $\sigma(j) = i$, then $\sigma(i) = i$). Intervention $do(\sigma)$ informally proceeds by transferring the information in X_i to $X_{\sigma(i)}$.
The definition of σ obviously relies on expert knowledge; it requires that X_i and $X_{\sigma(i)}$ be sufficiently similar for this transfer to make sense.

Definition (Macro-intervention).
The macro-intervention $do(\sigma)$ operates on $\mathbf{x} = (x_i)$ as follows: if σ effectively operates on X_i ($\sigma(i) \neq i$), then the value x_i is set to 0; else, x_i is set to the sum of x_j for all j s.t. $\sigma(j) = i$:

$$do(\sigma)(\mathbf{x}) = \mathbf{x}^{\emptyset} \text{ with } x_i^{\emptyset} := \begin{cases} 0 & \text{if } \sigma(i) \neq i \\ \sum_{j \text{ s.t. } \sigma(j)=i} x_j & \text{otherwise} \end{cases} \quad (5)$$

This definition stems from the structure of the set of potential causes in the Kantar database. As said, each potential cause is associated with a vectorial description, and some features are shared by a subset of potential causes. For instance, the boolean *organic* feature is defined for a large subset of food items; the boolean *fresh* feature is defined for all vegetables.

3. One further notes that the overlap assumption is not in the hypotheses of Thm 8, (Identi cation of the conditional mean potential outcome) in (Wang and Blei, 2019a), when limiting oneself to interventions that preserve the substitute confounder estimate, as will be the case for the interventions proposed in Section 4.

By setting such a feature to e.g. true, one defines a mapping on the set of potential causes: if X_i bears the feature, $\sigma(i) = i$, otherwise X_i is mapped onto $X_{\sigma(i)}$ where the vectorial descriptions of X_i and $X_{\sigma(i)}$ only differ regarding the considered feature. The macro-intervention thus simultaneously operates over all food items for which this attribute is relevant.

In the following, we shall consider the macro-intervention associated with the boolean *organic* attribute. This macro-intervention noted *do(organic)* intervenes on each variable X_i involving the *organic* feature, turning all bought food items of type X_i into organic ones (thus with type $X_{\sigma(i)}$). Note this intervention operates on many but not all food items; for instance the *organic* attribute is not defined for water, or sheep yogurt.

This definition calls for two observations. Firstly, the difference w.r.t. standard intervention (e.g. turning the only bread into organic bread) is that macro-interventions will expectedly have a more visible effect than a single variable-based intervention.[4] Intuitively, the domain expert defining any $do(\sigma)$ interventions is supposed to define a mapping where all interventions go in the same direction.

A second remark is that the macro-intervention preserves the structure of the sample. It is easy to see that the L_1 norm of each sample is preserved under the intervention, as every x_i is positive and σ is idempotent.

4.2. Dimensionality Reduction and Substitute Hidden Confounders

The main two challenges posed by the Kantar database, namely the high number of potential causes and the presence of hiden confounders, are handled by taking inspiration from the *Deconfounder* and from the domain of Natural Language Processing.

Basically, the consumption of each household is viewed as a document, where each food item is viewed as a word. The search for a probabilistic model accounting for the consumption over all households (akin set of documents) is achieved using Latent Dirichlet Allocation (Blei et al., 2003), extracting the "topics" involved in these documents[5] where a topic is a distribution defined on the food item dictionary. Analogous to the NLP topics present in e.g. a set of journal articles (e.g. "politics", "culture", "events"), a dietary topic, referred to as *dietary pattern* in the following, is a distribution on the food items. The consumption of a household is thus viewed as a mixture of dietary patterns, e.g. 25% continental breakfast, 30% baby food, 10% wines and aperitives... Note that the attribute involved in the definition of the structured intervention is not taken into account to achieve the identification of the dietary patterns.

Formally, the consumption $\boldsymbol{x}^{(k)}$ of the k-th household is a vector in $\mathbb{R}^D$, with D circa 4,000. The LDA-based change of representation maps the household consumption vectors onto the latent space of dietary patterns; $d = 16$ dietary patterns are considered in the following and detailed in Appendix. LDA is used to define a mapping ϕ from $\mathbb{R}^D$ onto $\mathbb{R}^d$, where $\phi(\boldsymbol{x}^{(k)}) = (a_j^{(k)})$ for $j = 1$ to d and $a_j^{(k)}$ the mixture weight of the j-th dietary pattern

4. Actually, given the level of the known unknown, the impact of a single food item-based intervention is expected to be hardly signi cant.
5. The extension of the approach using e.g. Variational Auto-Encoder (Kingma and Welling, 2014), in the spirit of the CEVAE (Louizos et al., 2017) is left for further work.

for the k-th household with $a_j^{(k)}$ summing to one:

$$\phi(\boldsymbol{x}_k) = (a_j^{(k)}), \quad \sum_{j=1}^{d} a_j^{(k)} = 1$$

Representation changes, mapping initial causes onto compound causes, are notoriously disliked in the context of causal studies (Pearl, 2003). Indeed, causal relations among compound causes are brittle, as the sense of the causality arrow often depends on the way the initial factors are aggregated to form compound factors. In the DeMaIn context, it must be emphasized that dietary patterns are not used *in lieu* of causes for the BMI outcome. Quite the contrary, the dietary patterns are used to define the substitute hidden confounders (below), with the idea that in a region where the SHCs are constant, the true, hidden, confounders are constant as well. The average treatment effect of the interventions can thus be measured locally in each such region, avoiding the biases due to the hidden true confounders.

4.3. Treatment and Control Groups

The dietary patterns are used to partition the households, with cluster C_j composed of the households where the majority pattern is j:

$$C_j = \{k \ s.t. \ \phi(\boldsymbol{x}_k)_j > \phi(\boldsymbol{x}_k)_\ell, \ for \ all \ \ell \neq j\} \tag{6}$$

where $\phi(\boldsymbol{x}_k)_j$ denote the weight of the j-th dietary pattern for the k-th household.

The treatment and control groups are first defined with respect to the overall population. For each household, the percentage of the organic food items in the yearly consumption is computed and households are ordered accordingly (Fig. 1). The top 20% of the households form the overall treatment group; the bottom 20% form the overall control group. It is noted that the lower threshold, defining whether a household participates in the treatment group is quite low (less than 5%; remind that these data have been collected in 2014).

The treatment and control groups are partitioned among the clusters (Eq. 6), defining the treatment and control groups within each cluster. The fact that the treatment groups in all clusters have the same minimum threshold of organic consumption level enables their fair comparison.

4.4. Controlling for covariates

Food consumption as well as the BMI are known to depend to some non-negligible extent on socio-economic features. The treatment and control groups are thus controlled for the covariates involved in the household description (including wealth, urban/rural location and educational background of the household members).

The confounding effects of the household covariates are controlled using an Inverse Propensity method (Austin, 2011) as follows. A sufficiently powerful classifier is trained to discriminate among the treatment and control groups, based on their covariate description; Xgboost is used in the experiments (Chen and Guestrin, 2016; Prokhorenkova et al., 2019). The prediction accuracy on test data determines whether the covariates actually enable to

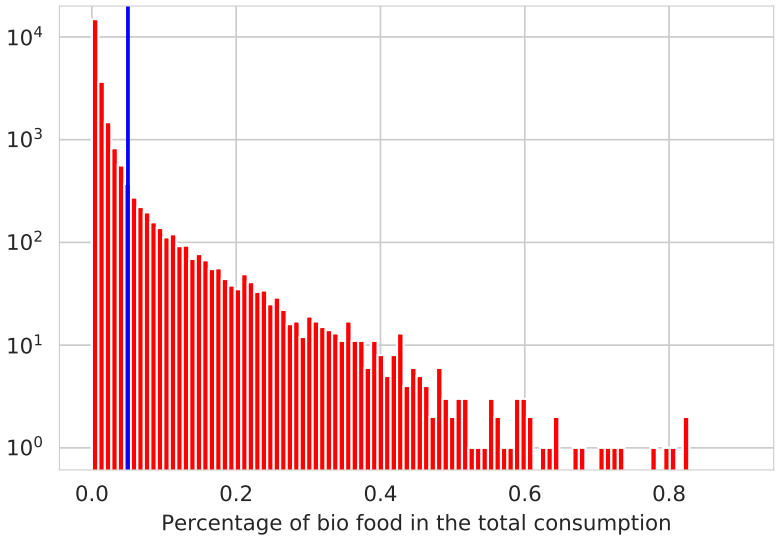


Figure 1: Kantar dataset: Histogram (in log scale) of the households ordered by percentage of organic food in household consumption in 2014.

predict the belonging of a household to the treatment or the control group, that is, if the covariates invalidate the unconfoundedness assumption.

In most cases, the predictive accuracy of the classifier is circa 60% good predictions on a balanced two-class sample, thus significantly higher than chance prediction. The prediction of the classifier is then firstly calibrated to deliver an estimate of the probability for a household member x to belong to the treatment or control group (Vaicenavicius et al., 2019; Alasalmi et al., 2020).

This estimated probability is used to support an inverse propensity score methodology (Austin, 2011), with

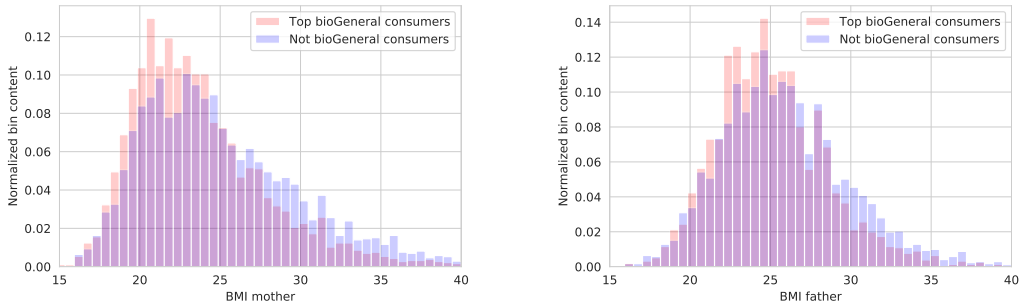
$$ATE(\mu) = \frac{\sum_{x2T} W(x,T)Y(x)}{\sum_{x2T} W(x,T)} - \frac{\sum_{x2C} W(x,C)Y(x)}{\sum_{x2C} W(x,C)} \tag{7}$$

where x is an adult individual in a household pertaining to T (respectively C) the treatment (resp. control) group, $W(x,T)$ (resp $W(x,C)$) is the inverse propensity score for x to belong to the treatment (resp. control) group, and $Y(x)$ is the BMI of the individual.

For numerical stability, only individuals with $W(x,T) < 5$ are considered in the treatment group (resp. with $W(x,C) < 5$ are considered in the control group).

5. Empirical study: The *do(organic)* macro-intervention

This section reports on the analysis conducted to estimate the average treatment effect of the *do(organic)* macro-intervention.



a) Women b) Men

Figure 2: The *do(organic)* macro-intervention: BMI of the overall treatment group (in pink) *vs* that of the general population (in blue) for Women (left) and Men (right).

5.1. Naive analysis

As shown in Fig. 1, the vast majority of households are not much interested in organic food, with the treatment group consisting of 2,000 households consuming more than 5% organic products.

The BMI histogram in the overall treatment group is compared to that of the whole population in Fig. 2, showing a clear difference of circa -.8 for men (Fig. 2, right) and about twice as high for women (Fig. 2, left):

$$\begin{aligned} \mathsf{E}_{emp}[BMI|treatment, men] - \mathsf{E}_{emp}[BMI|men] = & \quad 25.16 - 25.94 = -0.78 \\ \mathsf{E}_{emp}[BMI|treatment, women] - \mathsf{E}_{emp}[BMI|women] = & \quad 23.78 - 25.2 = -1.43 \end{aligned}$$

5.2. Controlling for covariates

The confounding effect of covariates (wealth, urban/rural location, education) is controlled using an Inverse Propensity correction based on a calibrated classifier (section 4.3). This correction, as could be expected, results in a significantly decreased, though still significant, ATE.

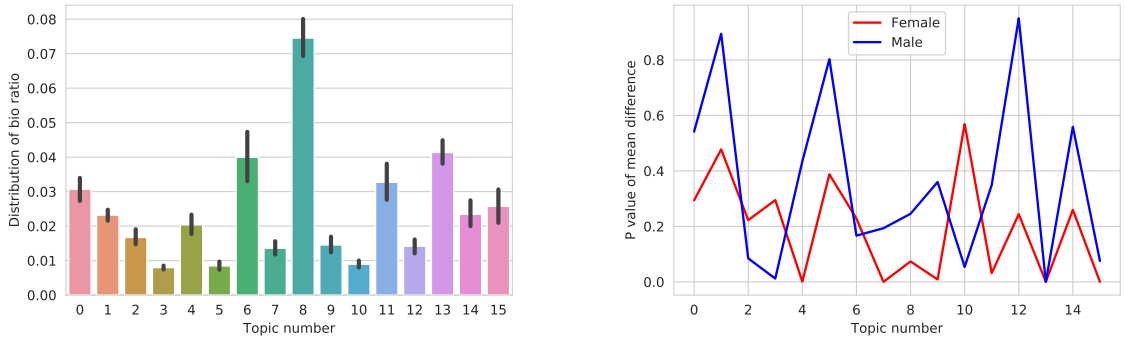
$$\begin{aligned} ATE_{men}(organic) = & \quad 25.16 - 25.54 = -0.38 \\ ATE_{women}(organic) = & \quad 23.78 - 24.36 = -0.58 \end{aligned}$$

5.3. Average Treatment Effect conditionally to dietary patterns.

The ATE is thereafter examined within each cluster. Remind that the dietary patterns are extracted from the household consumption baskets without considering the organic feature. The *do(organic)* intervention thus preserves each cluster, satisfying the assumptions of Thm 8 (Wang and Blei, 2019a), enabling to identify the mean potential outcome conditionally to the cluster.

The average value of the organic feature in each cluster is shown in Fig. 3.a. The p-value of the BMI difference between the treatment and control group in each cluster according to

a Student t-test is shown in Fig. 3.b. After correcting for the multiple hypothesis testing, few patterns present a statistically significant BMI difference between the treatment and the control group for men and women. The detail of the average BMI for men and women, in each pattern, in the treatment and control groups of each pattern, computed with inverse propensity weights (IPW) or without (Average) and the associated p-value is reported in Table 1.



(a) Average organic consumption

(b) P-values of the BMI difference

Figure 3: Organic macro-intervention, per dietary pattern.

Topic	Overall	Treatment		Control		Pvalue	
		Average	IPW	Average	IPW	Average	IPW
Women	24.8	23.8	24.1 ± 0.7	25.2	24.7 ± 1.1	$6 \cdot 10^{-32}$	$1 \cdot 10^{-20}$
4	24.1	23.0	23.5 ± 1.0	25.2	25.8 ± 1.8	$1 \cdot 10^{-2}$	$2 \cdot 10^{-8}$
7	24.7	22.9	22.9 ± 0.1	24.8	24.6 ± 1.5	$2 \cdot 10^{-4}$	$2 \cdot 10^{-14}$
9	24.8	23.3	23.4 ± 0.7	25.3	24.6 ± 1.4	$1 \cdot 10^{-1}$	$3 \cdot 10^{-2}$
11	25.1	24.5	25.0 ± 0.4	26.4	26.8 ± 1.5	$3 \cdot 10^{-1}$	$1 \cdot 10^{-2}$
13	24.6	24.1	24.7 ± 0.7	25.3	26.0 ± 1.3	$7 \cdot 10^{-3}$	$3 \cdot 10^{-8}$
15	24.4	23.0	23.0 ± 0.3	24.9	24.3 ± 1.2	$5 \cdot 10^{-3}$	$1 \cdot 10^{-4}$
Men	25.8	25.2	25.4 ± 0.5	25.9	25.8 ± 0.5	$5 \cdot 10^{-3}$	$4 \cdot 10^{-15}$
7	25.7	26.1	26.4 ± 0.7	25.5	25.4 ± 1.0	$9 \cdot 10^{-1}$	$6 \cdot 10^{-5}$
10	25.7	25.0	25.1 ± 0.6	26.1	26.2 ± 1.2	$5 \cdot 10^{-1}$	$4 \cdot 10^{-3}$
13	26.1	25.4	25.6 ± 0.4	26.8	26.7 ± 0.5	$1 \cdot 10^{-4}$	$2 \cdot 10^{-9}$

Table 1: The *do(organic)* macro-intervention per cluster, for Women and Men, showing the empirical BMI average and the inverse propensity weighted average (noted IPW), together with their p-value. Only patterns with statistically significant BMI difference are reported.

6. Discussion and Perspectives

This paper proposes a first step toward the observational causal modeling of the relations between food consumption and Body Mass Index, exploiting the wealth of Kantar observational data.

The lesson learned is twofold. Firstly, in this "many causes" landscape, the definition of the so-called macro-interventions was instrumental to actually detect significant effects. A limitation of the approach is that macro-interventions must be designed to preserve key properties according to the domain knowledge. For instance, a macro-intervention operating on the degree of alcohol in the alcoholic beverages would *not* preserve the caloric intake of the household, thus hindering the interpretation of its ATE.

Secondly, the construction of substitute hidden confounders enables to detect significantly different effects in different household clusters. Conditioning on these clusters suggests specific causal relations that are hardly visible when considering the overall population. Of course, these results need be confirmed by further experimental studies.

This work opens to several research perspectives. A first one focuses on considering other macro-interventions, such as based on the *fresh / frozen* feature defined for vegetables. Another perspective is to extend the extraction of the substitute hidden confounders to handle also the sociological covariates, and control *de facto* for their impact without requiring the IPW-based correction.

Acknowledgments

The study was conducted in the context of the Nutriperso *Initiative de Recherche Stratégique*, that funded the first author. The authors wish to thank Andrei Constantinescu, for many discussions on the topic.

References

Tuomo Alasalmi, Jaakko Suutala, Juha Röning, and Heli Koskimäki. Better classifier calibration for small datasets. *ACM Trans. Knowl. Discov. Data*, 14(3):34:1–34:19, 2020.

Susan Athey, Guido W. Imbens, and Michael Pollmann. Comment on: "the blessings of multiple causes" by Yixin Wang and David M. Blei. *Journal of the American Statistical Association*, 114(528):1602–1604, October 2019.

P.C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res.*, 46:399–424, 2011.

David Blei, Andrew Ng, Michael Jordan, and John Lafferty. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, pages 993–1022, 02 2003.

John E. Blundell, R.J. Stubbs, Cheryl Golding, Fiona Croden, Rahul Alam, Stephen Whybrow, J. Le Noury, and C.L. Lawton. Resistance and susceptibility to weight gain: individual variability in response to a high-fat diet. *Physiology & behavior*, 86(5):614–622, 2005.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD*, pages 785–794. ACM, 2016.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *ArXiv2011.08047*, 2020.

Alexander D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and A promising alternative. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 3478–3486. PMLR, 2019.

Katherine M Flegal, Brian K Kit, Heather Orpana, and Barry I Graubard. Association of allcause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. *Jama*, 309(1):71–82, 2013.

J. Grimmer, D. Knox, and B.M. Stewart. Naive regression requires weaker assumptions than factor models to adjust for multiple cause confounding. *arXiv preprint arXiv:2007.12702*, 2020.

Y. Hasegawa, SY. Chen, and L. Sheng. Long-term effects of western diet consumption in male and female mice. *Sci Rep*, 10(14686), 2020.

Kosuke Imai and Zhichao Jiang. Comment: The challenges of multiple causes. *Journal of the American Statistical Association*, 114(528):1605–1610, October 2019. doi: 10.1080/01621459.2019.1689137. URL https://doi.org/10.1080/01621459.2019.1689137.

INSEE. Comprendre le calcul du pouvoir d'achat : perceptions individuelles et mesure statistique. Technical report, 2019. Available online at https://www.insee.fr/fr/information/3707563, accessed september, 2021.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2014.

Christos Louizos, Uri Shalit, Joris M. Mooij, David A. Sontag, Richard S. Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *NIPS*, pages 6446–6456, 2017.

Daniel L. McGee. Body mass index and mortality: a meta-analysis based on person-level data from twenty-six observational studies. *Annals of epidemiology*, 15(2):87–97, 2005.

Elizabeth L. Ogburn, Ilya Shpitser, and Eric J. Tchetgen Tchetgen. Counterexamples to "the blessings of multiple causes" by Wang and Blei. *Arxiv2001.06555*, 2020.

J. Pearl and D. Mackenzie. *The book of why.* Basic Books, 2018.

Judea Pearl. Causality: models, reasoning, and inference. *Econometric Theory*, 19(46): 675–685, 2003.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press, 2017.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features, 2019.

D. Reeves, K. Howells, M. Sidaway, A. Blakemore, M. Hann, M. Panagioti, and P. Bower. The cohort multiple randomized controlled trial design was found to be highly susceptible to low statistical power and internal validity biases. *Journal of clinical epidemiology*, 95: 111–119, 2018.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *J. Mach. Learn. Res.*, 12:1225–1248, 2011.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Reinhard Stöger. The thrifty epigenotype: an acquired and heritable predisposition for obesity and diabetes? *Bioessays*, 30(2):156–166, 2008.

Juozas Vaicenavicius, David Widmann, Carl R. Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B. Schön. Evaluating model calibration in classification. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 3459–3467. PMLR, 2019.

Yixin Wang and David M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019a.

Yixin Wang and David M. Blei. The blessings of multiple causes: Rejoinder. *Journal of the American Statistical Association*, 114(528):1616–1619, October 2019b.

Appendix: Clusters and Dietary patterns

The set of clusters used as SHCs is described in Table 2.

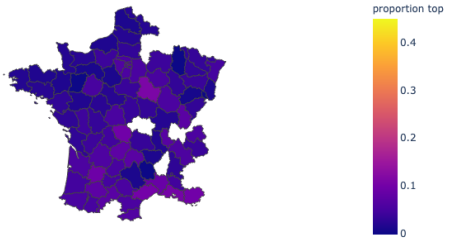
The i-th cluster is formed of the households for which the majoritary dietary pattern is the i-th one.

Each dietary pattern is described from its main 10 food items, represented as a word cloud, and the geographical location of the households falling in this cluster. Note that some dietary patterns but not all are attached with a given region, although the geographic location was not taken into account to define the dietary patterns.

Cluster	# Households	# Women	# Men	# Couples	Fraction
0	933	777	667	511	0.047
1	1804	1640	1328	1164	0.090
2	1170	843	891	564	0.058
3	4811	4334	4169	3692	0.238
4	729	673	538	482	0.037
5	860	666	711	517	0.043
6	521	453	415	347	0.027
7	1042	991	911	860	0.052
8	1784	1641	1381	1238	0.088
9	887	626	754	493	0.044
10	1301	1224	1023	946	0.065
11	420	344	229	153	0.022
12	824	702	665	543	0.041
13	1900	1671	1305	1076	0.094
14	674	539	473	338	0.034
15	595	593	575	573	0.030

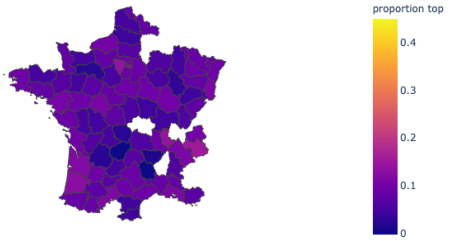
Table 2: Distribution of households among the 16 clusters

fresh vegetables	0.153
ham	0.076
sweet biscuits	0.029
milk	0.029
fish	0.028
sauces	0.027
bread	0.026
pastry	0.025
fresh dessert	0.023
cream	0.022



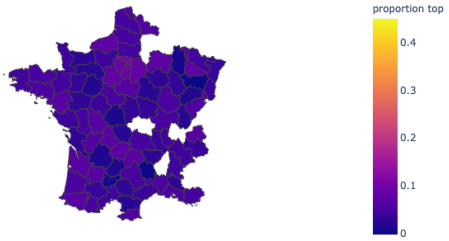
(a) Topic 0, mostly fresh vegetables

fresh dish	0.120
fresh dessert	0.118
frozen dishes	0.059
pastry	0.056
sodas	0.052
sweet biscuits	0.041
canned ready-made meals	0.036
bonbon	0.028
icecreams	0.028
fresh vegetables	0.022



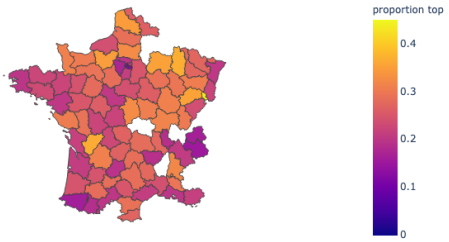
(b) Topic 1: fresh, frozen and canned dishes + fresh desserts

butter	0.066
sweet biscuits	0.040
fresh dessert	0.037
fresh vegetables	0.036
cream	0.035
pastry	0.033
beef	0.031
canned ready-made meals	0.030
charcuterie	0.029
sauces	0.029



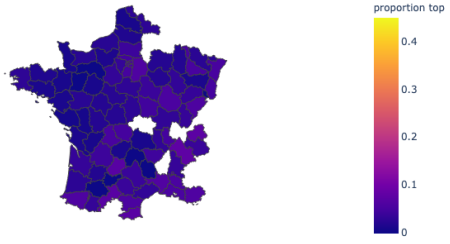
(c) Topic 2: butter, desserts, fresh cream

baby food	0.242
baby dessert	0.167
baby drinks	0.050
fresh dessert	0.049
fresh vegetables	0.039
baby breakfast	0.0212
sweet biscuits	0.020
pastry	0.019
sauces	0.018
canned dessert	0.017



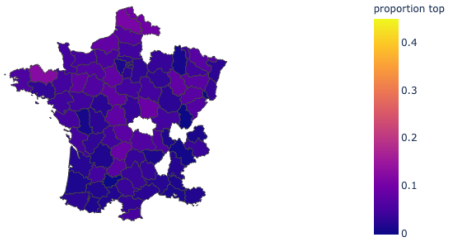
(d) Topic 3: food and drinks for babies

fresh vegetables	0.263
beef	0.054
fish	0.049
ham	0.036
pork	0.031
poultry	0.029
veal	0.022
sausage	0.022
sweet biscuits	0.021
fresh dish	0.019



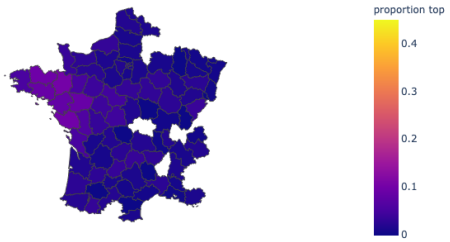
(e) Topic 4: fresh vegetables + meat (beef, poultry, pork, ham) and fish

fish	0.066
beef	0.050
shellfish	0.034
wine	0.031
pastry	0.027
fresh dessert	0.027
pork	0.026
soup cans	0.026
snacks	0.023



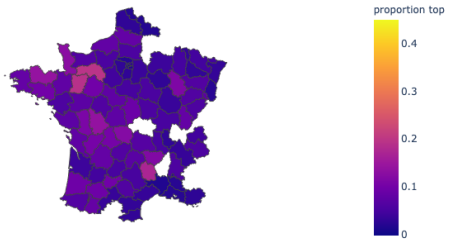
(f) Topic 5: fresh vegetables + fish + shellfish + beef

sweet biscuits	0.091
fresh dessert	0.080
cereals	0.070
canned dessert	0.053
pastry	0.050
sodas	0.048
bonbon	0.036
sauces	0.032
cream	0.028
icecreams	0.026



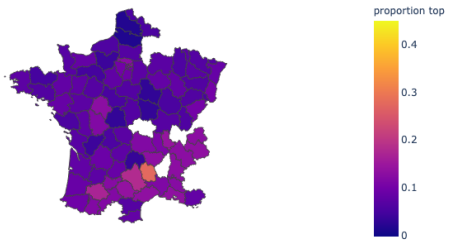
(g) Topic 6: sweet biscuits + dessert + cereals − mostly in Brittany and Normandy

fresh vegetables	0.117
dry fruits and seeds	0.044
sweet biscuits	0.036
fish	0.035
bread	0.035
snacks	0.033
cream	0.032
sauces	0.031
cereals	0.030
fresh dessert	0.028



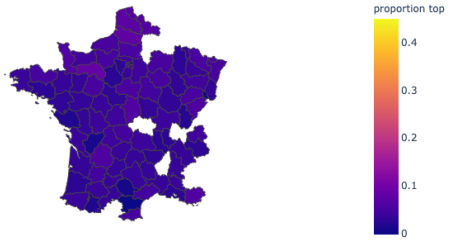
(h) Topic 7: fresh vegetables + dry fruits + desserts

fish	0.066
fresh vegetables	0.062
jam	0.053
soup cans	0.046
margarine	0.044
sweet biscuits	0.042
bread	0.035
fresh dish	0.033
coulommier cheese	0.032
milk	0.030



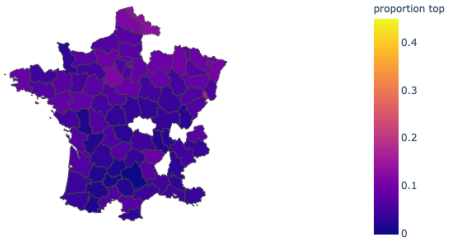
(i) Topic 8: fish, vegetables, ham, soups − mostly in South

fresh dish	0.242
fresh vegetables	0.067
beef	0.038
canned ready-made meals	0.035
fresh dessert	0.028
pastry	0.028
sauces	0.027
ham	0.026
sweet biscuits	0.024
sodas	0.022



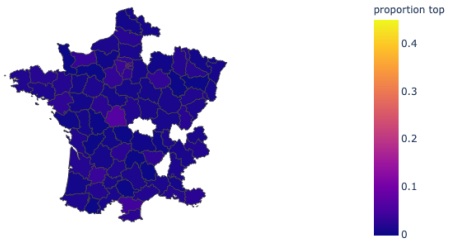
(j) Topic 9: fresh dishes and vegetables + ready meals

fresh vegetables	0.049
sauces	0.047
charcuterie	0.035
bread	0.034
sodas	0.033
cream	0.031
sweet biscuits	0.030
pastry	0.030
icecreams	0.027
frozen dishes	0.026



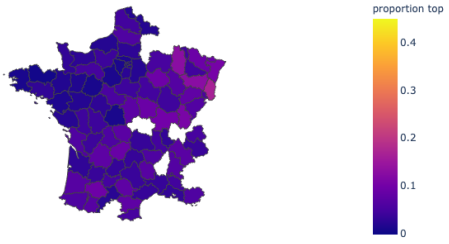
(k) Topic 10: vegetables, sandwiches, sodas − mostly North

sweet biscuits	0.071
sodas	0.070
fresh dessert	0.058
pastry	0.050
sauces	0.044
bonbon	0.038
beef	0.031
fresh dish	0.029
canned ready-made meals	0.027
cream	0.026



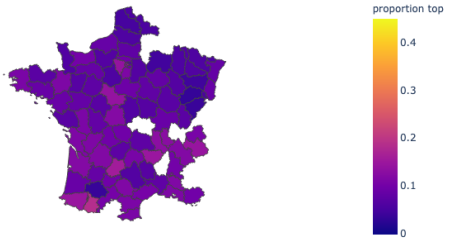
(l) Topic 11: sweet food

cream	0.066
confectionary flavouring	0.065
fresh vegetables	0.063
milk	0.057
sweet biscuits	0.035
fresh dessert	0.031
sauces	0.026
dry fruits and seeds	0.0263
fish	0.025
flour	0.022



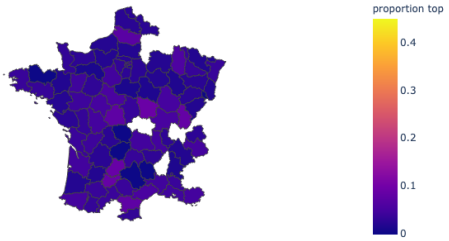
(m) Topic 12: cream, milk, flour, confectionary flavourings – mostly East

fresh vegetables	0.096
beef	0.084
pork	0.076
poultry	0.047
sausage	0.036
pastry	0.036
fresh dish	0.033
fish	0.026
wine	0.026
salami	0.025



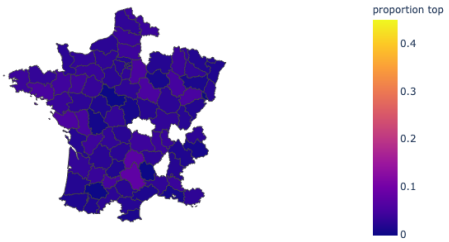
(n) Topic 13: fresh vegetables + beef, pork, poultry

fresh dessert	0.203
ham	0.061
fresh vegetables	0.052
bread	0.032
pastry	0.029
sodas	0.028
fish	0.028
sauces	0.026
cream	0.023
beef	0.022



(o)Topic 14: fresh dessert + ham

wine	0.223
fresh vegetables	0.050
aperitives	0.038
snacks	0.030
fish	0.026
sauces	0.026
sparkling wine	0.025
fresh dish	0.024
cream	0.020
dry fruits and seeds	0.019



(p) Topic 15: wine, aperitives,