Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-03890-w OPEN

Forecasting food trends using demographic pyramid, generational differentiation and SuperLearner

Daria Loginova [1✉] & Stefan Mann[1]

The objective of this paper is to predict food consumption patterns for future decades by different social groups while taking generational change into account in the modelling. Using over 20 million observations of households in Switzerland from 1990 to 2017, we develop and apply four forecasting techniques that shift from referenced linear forecasts to population-driven forecasts. Each method considers values of selected household characteristics to define a "social group", derives the proportion of each social group in society for the years 1990–2050, forecasts the future consumption of 75 food items in each social group in its unique way, and weighs these consumption patterns to obtain a future consumption for the total population. Although the results vary for each of the 75 food items and each method, altogether and in general, they define a narrow interval of future consumption development until 2050. All aspects of the approaches and the comparison of the outcomes contribute to knowledge about possible and nontrivial forecasting techniques on big data and foresight about the future of home food consumption.

[1] Socioeconomics, Agroscope, 8356 Ettenhausen, Tänikon 1, Switzerland. ✉email: daria.loginova@agroscope.admin.ch

Introduction

Studies on future food demand indicate that realistic projections of food demand are important and relevant in managing stocks, logistics, food production, supply chains, disease risk and the environmental footprint (Daniel et al., 2010; Food and Agriculture Organization of the United Nations (FAO), 2017; Flies et al., 2018; Smith, 2014; United Nations, 2019). Our contribution builds on the importance of generations in prognosing food trends (see Mann and Loginova, 2023; FAO, 2018) and the old tradition of finding better statistical solutions motivated by the demands of food research (e.g. Brown, 1954; Benjamin, 1992; Girshick and Haavelmo, 1947; Tintner, 1953). The main purpose and goal of our research is to develop and apply methods of predicting food consumption patterns for future decades and social groups while taking generational change into account.

Benefitting from the advantages of existing big data on Swiss food consumption and a ready demographic prognosis on population by age and gender (demographic pyramid) performed by the Federal Statistical Office (2020), we extend the range of ideas about studying consumption offered by Athey and Imbens (2019) by providing and performing the following applied techniques of long-term food trend forecasting:

- Model A: reference scenario, extrapolations of trends yearly;
- Model B: extrapolations of trends by age and gender and adding the information from the demographic pyramid;
- Model C: forecasting consumption using generational trends;
- Model D: general linear model (GLM) prediction with additional factors;
- Model E: a SuperLearner (Polley et al., 2023) that allocates weights for the predictions of powerful models to achieve even better explanatory power compared to the powers of the sole models included in the SuperLearner. The SuperLearner used in this work is a weighted combination of the logit model, XGBoosting and random forest applied to range normalised consumption data for each separate food and considering the best explanatories available in our data set.

The remainder of this article is structured as follows. Section "Literary review" provides a short literary review. Section "Data" describes our data and data processing. Section "Methods" describes five applied methods. Section "Results" presents and discusses our results and the limitations of our approaches. Section "Discussion and conclusion" outlines the conclusions reached in this study.

Literary review

Past approaches to food demand forecasting vary widely from early partial equilibrium modelling, extrapolation and linear regression (Bodirsky et al., 2015; Mackle and David, 1976; Wood, 1977) to relatively new machine learning, decision tree, neural network and boosting techniques (e.g. Bozkir and Sezer, 2011; Harshini et al., 2021; Iyer, 2020; Lime, 2022; Lutoslawski et al., 2021; Massachusetts Institute of Technology, 2020; Perego, 2019; Pujara et al., 2022). Reviews of these methods were presented by van Dijk et al. (2021), Petropoulos and Carver (2019) and Tarallo et al. (2019). Flies et al. (2018) noted that 'while estimates of future global kilocalorie demand have a broad range, they are not consistently dependent on model complexity or form'.

Food tastes and demands not only depend on age and other well-known factors, such as income and gender, but also change across generations and social groups over time. In this paper, we attempt to consider these culturally and demographically driven factors in prognosing future food consumption at the national level. Switzerland is a convenient case in point for a study on culturally driven trends because changes in consumption patterns are driven mostly by demand rather than by supply (Baur et al., 2022; Godin and Sahakian, 2018; Mann and Loginova, 2023; Sahakian et al., 2020).

Data

Household survey. We used disaggregated household-level data from the Swiss Federal Statistical Office (2022) from up to 12,000 households in Switzerland per year for the years 1990–2017. The basis of this study is a yearly random sample of households in Switzerland who reported their household characteristics and food purchasing diaries, specifically a list and the amount (in kilograms or litres) of food they bought over a one-month period. Therefore, the data we used were obtained from a randomised observational survey. The average numbers of consumption per person calculated for this database mainly match official statistics (AGRISTAT, 2023) and are smaller than those declared by some Swiss non-governmental organizations because the latter tend to use food balances and add restaurant food and foods eaten by tourists to their averages for the population. The data we used do not contain this kind of information, so our data do not account for a general shift from home to restaurant food consumption within households (Czarniecka-Skubina and Kowalczuk, 2015). Additionally, the weight declared by households is usually the weight of food before processing (e.g. cooking or peeling) and contains, for example, the weight of bones present in meat products. To the best of our knowledge, these shortcomings in food data have not yet been solved in any of the existing consumption databases, even though this bias is considered meaningful (see, e.g., potato weight shrinkage along the supply chain in Willersinn et al., 2015; vegetable waste in Bouclaous and Jaubert, 2015). Nevertheless, we have used the largest, most long-term and most reliable dataset that is available in Switzerland.

We employed 75 food categories, which exceed those in Mann and Loginova (2023), and also defined 16 generations with 10 consecutive birth years each, ranging from Generations 0 (born between 1896 and 1905) to 15 (born between 2046 and 2055). If all participants of the household belonged to the same generation, the household was assigned to this generation; that is, generations were assigned to singles and to households with people of similar age. The other households were distributed between the 'households with children' and 'mixed' groups, which were not used for generational analyses. In total, our analysis included 20 million observations representing consumption volumes per person for 75 comparable food categories for the years 1990–2017. The characteristics of the households are listed in Table 1.

In general, we used only balanced household characteristics in our estimations. The population and our data are balanced by members: one-half of all members belong to households with one or two members, and the other half belong to households with more than two members. Income distribution is close to the normal distribution in our sample, and the number of members in the income groups is close to each other. The age distribution is normal, but the older population has been slightly overrepresented since 2010. The sample was overweighted towards females. The regional language variable is balanced, with 51% of the population belonging to the German-speaking part of Switzerland and the French, Italian, retro-Romanian and other languages totalling 49% of the sample. Mixed households (by age and generation) and households with children were excluded from the analysis.

Table 1 Data observed for households.

#	Household characteristic	Levels	Format and meaning of the variable
1	Household ID	Character	Individual ID for each household and year
2	Year of observation	Years	1990, 2000–2017
3	Food	Character	Food Name
4	Household consumption	None	Numeric variable, household consumption in grams
5	Number of members in the household	None	Numeric variable, meaning the household population
6	Gross household income (monthly)	(1) <3999 CHF (2) 4000–7000 CHF (3) 7000–10,000 CHF (4) More than 10,000 CHF	Ordered numeric variable, meaning the gradual income class increase
7	Age interval	(1) 15–24 years (2) 25–34 years (3) 35–44 years (4) 45–54 years (5) 55–64 years (6) 65–74 years (7) 75 and over	The number that declares age group in the observed year; this characteristic maps the age between the participants of the household
8	Gender majority	If the number of female respondents relative to the number of total household respondents (%): (1) 50% or higher, then women (2) is strictly <50%, then men	Dummy variable is higher for a household with a higher share of females in their population; the 50% threshold is considered female, following Loginova and Mann (2024)
9	Landscape, culture, language	Regional language: (1) German and Swiss German (0) Other	Dummy that reflects the language of the region
10	Generations	(0–15) Generations 0–15	Generations, as defined in the main text of the article

Our data needed an outlier policy, so we considered as outliers 1% of households with the highest consumption and 1% of households with the lowest consumption per person. Any higher percentage of outliers would have reduced the information on rare food consumption habits (e.g. consumption of horse meat, in the case of a lower upper bond for horse meat, and consumption of vegetarian consumers, in the case of a higher bottom bond for meat).

Considering social characteristics. We did not consider the smoking status of household residents, which was deemed significant for Swiss dairy consumption by Inanir et al. (2020) because the degree to which smoking contributes to consumption in most included nondairy food categories is not quantitatively proven for Swiss residents. Additionally, we did not use cultural variables, such as nationality, because the data set contains information about persons of more than 100 different nationalities who may live with a Swiss person within one household, which represents thousands of possible cultural combinations and food attitudes. Kearney (2010) defines other drivers of food consumption at a global level: marketing, urbanisation, retail spread, consumer attitudes, trade liberalisation and international corporations. Because Switzerland is a smaller territory than most countries, the global factors of food consumption change may be considered relatively homogeneous for most Swiss territories.

We list the value of each characteristic of a household as an ID, which we define as a 'social ID', which defines a 'socioeconomic group' or 'social group'. Social IDs vary between the studied models and contain information about combinations of the following:

- Model A: no social factors (the social ID is not used);
- Model B: household age and gender;
- Model C: household generation and gender;
- Models D and E: household age, gender, generation, gross income and region.

For example, in Model B, a group of households aged 35–44 and with more males than females among the members is assigned the social ID of '3|0'. In Model D, a group of households aged 25–34, with more male than female members, belonging to Generation 7, Income Group 2 (4000–7000 CHF) and the German-speaking region of Switzerland is assigned a social ID of '2|0|7|2|1'. Despite the data set containing over 20 million observations, we can calculate averages for almost any combination of the listed factors and even track the dynamics of consumption for any of these social groups.

We grouped the households in our data set for each studied year according to a social ID and averaged the consumption per person separately within each group and for each food. Each characteristic considered increased the number of socioeconomic groups of households and decreased the number of households in each group. As a result, our data became panelled, whereby the single sample was a social group, and the years 1990–2017 were the time period. We calculated the number of participants (people in households) in each social group, which thereafter served as the weight of the socioeconomic group. We used the weight of a socioeconomic group to calculate the share of the socioeconomic group in the total population.

Population dynamics. Information on population dynamics was obtained from statistical and demographic forecasting performed by the Federal Statistical Office (2020). These data contained the actual (since 1860) and forecasted (2020–2070) number of men and women in Switzerland in each age group. We used the actual numbers of men and women of each age for 1990–2022 and the forecasted numbers of men and women of each age for 2023–2050. These forecasts predict population numbers based on available demographic data, with the upper and lower limits defined by the cumulative standard error. We chose the average pyramid forecast to model the social group weighting.

The most convenient way to process these data was to switch from age to year of birth. After assigning generations and ages to cohorts by birth years, we established the group sizes by age and gender and by generation and gender. We used this information to calculate the shares of age groups by gender and the shares of generations by gender to weigh the final forecasts of Models B and C, respectively. However, we also used a data-driven approach to discover the social group sizes for Models D and E because the ready forecasts for the population groups defined at the level of detail used in our models do not currently exist (see the section "Sizes and shares of social groups").

Avoiding the weaknesses of methods. There are three main issues that complicate forecasting over time using trends. The first is the nonstationarity of time series, which challenges the validity of any prediction technique (see, e.g., Loginova and Mann, 2022). Moreover, there is a growing practice of achieving stationarity before performing causal and panel data analysis (e.g. Callaway and Sant'Anna, 2021; Mink et al., 2023); therefore, stationarity was ensured for the linear estimations in this study. The second issue is the possibility of occasionally assuming linearity for nonlinear data and estimates over time. When this assumption is made by researchers, machines, or functions, it may harm the prediction if the development of the dependent variable is exponential or polynomial. The large number of cases and variety of best-fitting trends may result in forecast challenges. The same principle applies to assumptions of normality for nonnormally distributed data. The third issue is the zero and negative predicted values for variables in negative dynamics and infinitely high predicted values for variables in positive dynamics, both of which are predicted in the long run. This issue is hidden in modelling for descriptive purposes but becomes apparent if the model is applied to predicting the values of many time units in the future.

Growth rates for Models A–C. When analysing our household consumption data with Models A–C, we faced the three above-mentioned issues. A well-known solution to avoid or at least minimise the consequences of such issues has been established in time-series analysis: the transition from values in levels to growth rates. Therefore, we finished the data preparation by calculating the growth rate values within each social ID (id). Formally, we took a variable of interest $\acute{c}_{i,\mathrm{id},t,h}$ in levels for each household $h \in [1\ldots H]$, food (i) at time (t), calculated the weighted mean of those across the households $\acute{c}_{i,\mathrm{id},t}$ and calculated growth rates $c_{i,\mathrm{id},t}$ as follows:

$$c_{i,\mathrm{id},t} = \left\{ \begin{array}{l} \frac{\acute{c}_{i,\mathrm{id},t}}{\acute{c}_{i,\mathrm{id},t-1}} - 1 = \frac{\sum_h^H \acute{c}_{i,\mathrm{id},t,h} * \acute{w}_{\mathrm{id},t,h}}{\sum_h^H \acute{c}_{i,\mathrm{id},t-1,h}, * \acute{w}_{\mathrm{id},t-1,h}} - 1,\ t \in [2001\ldots 2017] \\ \sqrt[n]{\frac{\acute{c}_{i,\mathrm{id},2000}}{\acute{c}_{i,\mathrm{id},1990}}} - 1,\ t \in [1991\ldots 2000],\ n = 10 \end{array} \right\} \tag{1}$$

We used the number of members in the households as the weight of a household in social group (id); therefore, $\acute{w}_{\mathrm{id},t,h}$ is a share of the population belonging to household (h) at time (t) in the total population with social group (id). Because the first year of observation was lost during the rate calculation and the values for the years between 1990 and 2000 were missing from the raw data, we shared consumption growth between n years from 1991 to 2000 equally. The equation for $t \in [1991\ldots 2000]$ was derived from $\acute{c}_{i,\mathrm{id},2000} = \acute{c}_{i,\mathrm{id},1990} * (1+x)^{2000-1990}$. In this manner, the speed of change (x) was the same for $t \in [1991\ldots 2000]$. Compared to the extrapolation of levels, when the change in levels is stable, the applied transformation does not result in negative growth dynamics, reflects the future estimations of trends in growth rates less and is more realistic and easier to achieve in practice. Transition to growth rates is a common practice in time-series analysis that makes obtaining stationary, normalised variables and fair estimates possible.

Range normalisation for Models D and E. Panel data modelling is mainly based on within-and-between effects in levels, which to some extent is the first differencing of the data. However, panel data can suffer from the other issues discussed in the section "Avoiding the weaknesses of methods". We decided to range normalise the data to compare the models between the products and each other.

To prepare the data for further estimation, we added a shorter ID to our database that excluded the generational factor from each social group (id) and denoted it thereafter as $\ddot{\mathrm{id}}$. Generations (and time) within the IDs reduced the opportunity to match social group sizes, maximum (and minimum) consumptions and other variables for future generations and years. Using a new $\ddot{\mathrm{id}}$, we performed range normalisation for the consumption growth of different social groups (id) to the maximum and minimum consumption growth ($c_{i,\ddot{\mathrm{id}},\max}$ and $c_{i,\ddot{\mathrm{id}},\min}$) between the social groups ($\ddot{\mathrm{id}}$):

$$\ddot{c}_{i,\mathrm{id},t,\mathrm{GLM}} = \frac{c_{i,\mathrm{id},t} - c_{i,\ddot{\mathrm{id}},\min}}{(c_{i,\ddot{\mathrm{id}},\max} - c_{i,\ddot{\mathrm{id}},\min})} \tag{2}$$

$$\ddot{c}_{i,\mathrm{id},t,\mathrm{SL}} = \left\{ \begin{array}{l} 1,\ 0.5 < \frac{c_{i,\mathrm{id},t} - c_{i,\ddot{\mathrm{id}},\min}}{(c_{i,\ddot{\mathrm{id}},\max} - c_{i,\ddot{\mathrm{id}},\min})} \\ 0,\ 0.5 \geq \frac{c_{i,\mathrm{id},t} - c_{i,\ddot{\mathrm{id}},\min}}{(c_{i,\ddot{\mathrm{id}},\max} - c_{i,\ddot{\mathrm{id}},\min})} \end{array} \right\} \tag{3}$$

$$c_{i,\ddot{\mathrm{id}},\max} = \max(c_{i,\ddot{\mathrm{id}},t}),\ c_{i,\ddot{\mathrm{id}},\min} = \min\left(c_{i,\ddot{\mathrm{id}},t}\right),\ t \in [1991\ldots 2017] \tag{4}$$

For GLM we used $\ddot{c}_{i,\mathrm{id},t,\mathrm{GLM}}$ as described in Eq. (2). For binary models in SuperLearner, we considered values of $\ddot{c}_{i,\mathrm{id},t,\mathrm{GLM}}$ that were over 0.5 as 1, and values <0.5 were assigned a 0 value, as $\ddot{c}_{i,\mathrm{id},t,\mathrm{SL}}$ in Eq. (3) describes. The approaches to feeding these data into the models are described in the section "Methods".

Methods

We aimed to forecast personal consumption for each of the 75 types of food. Model A in Eq. (1a), which is the reference scenario, presents a simple trend in personal consumption growth (per person per food):

$$c_{i,\mathrm{id},t} = \alpha_i + \beta_i t + \varepsilon_{i,\mathrm{id},t}, \tag{1a}$$

where $c_{i,\mathrm{id},t}$ is the growth in consumption per person of food (i) at time (t) in a social group (id), α_i are constants and $\varepsilon_{i,\mathrm{id},t}$ are the error terms. For linear models, we used the robust 'felm' function from the 'lfe' R package by Gaure et al. (2023) and the robust 'lm' function from the 'stats' R package. Model B, described in the section "Linear extrapolation by age and gender", used growth in consumption volume, food type and the variables used to define the demographic pyramid, namely age and gender. This model, therefore, considered only households of one gender. Model C, described in the section "Prognosing by decomposing the total trend into generational trends ", used only household generation and gender information. Models D and E, described in the sections "GLM model prediction" and "SuperLearner", respectively, used all available information, albeit under strong assumptions that the estimated contributions of all variables to consumption would remain unchanged and that the homogeneity assumption would hold within the social groups. Model A did not require social group weights, as it did not consider social change; however, Models B–E considered social characteristics and predicted the growth of consumption for all combinations of values in those characteristics, such that these outcomes needed to be weighted for a final forecast. We describe additional forecasts performed to assist the methods in the section "Additional calculations", the transition back from growth rates to levels in the section "Transition back to variables in levels", the weighting of the outcomes in the section "Weighting" and the limitations of our methods and research in the section "Limitations and assumptions".

Linear extrapolation by age and gender. To forecast food consumption using the population pyramid—that is, its two components, age and gender (Model B)—we employed a four-step procedure: (1) we employed linear models for growth in consumption of each studied food and social group defined by age interval and gender; (2) we extrapolated the estimated trends for each studied food and social group; (3) we conducted a transformation of consumption growth back to consumption in levels; and (4) we adjusted the forecasted values for demographic population changes by reweighting the final projections according to the new shares of social groups in the total population. Steps 1 and 2 occurred as follows: for each food (i), age (j) and gender (g) at time (t), we used the observations of growth in consumption per person in the social groups (id; age [j] and gender [g] combined) and defined consumption growth in percentage ($c_{i,j,g,t}$) and the slopes of the consumption growth trends ($\beta_{i,j,g}$). We then obtained the estimates of consumption growth trends per person ($\hat{\beta}_{i,j,g}$) and constant growth rates ($\hat{\alpha}_{i,j,g}$), using the following regression:

$$c_{i,j,g,t} = \alpha_{i,j,g} + \beta_{i,j,g} t + \varepsilon_{i,j,g,t}, \tag{1b}$$

where $\varepsilon_{i,j,g,t}$ are the error terms. Therefore, the trends in our study were measured by regressing the variable of interest on a time variable. We replaced insignificant estimates of $\beta_{i,j,g}$ and $\alpha_{i,j,g}$ (i.e. $\hat{\beta}_{i,j,g}$ and $\hat{\alpha}_{i,j,g}$ with a p-value of more than 0.01) with zeroes and calculated consumption growth rates until 2050 as follows:

$$\widetilde{c}_{i,j,g,\dot{t}} = \hat{\alpha}_{i,j,g} + \hat{\beta}_{i,j,g} * (t - 1990), t \in [2018 \ldots 2050] = \dot{t} \tag{2b}$$

Prognosing by decomposing the total trend into generational trends. When distinguishing generational trends (Model C), we employed a four-step procedure: (1) linear models of consumption growth for each studied food, generation, and gender, (2) the extrapolation of the trend estimates for future generations, (3) transformation back to consumption in levels, and (4) the calculation of consumption growth projections and adjustment based on generational population change. Steps 1 and 2 occurred as follows: for each food (i), generation (p) and gender (g) at time (t), we assigned the observations in the social groups (id; p and g combined) and defined consumption growth in percentage ($c_{i,p,g,t}$) and the slopes of the consumption growth trends ($\beta_{i,p,g}$). We then obtained the estimates of the consumption growth trends per person ($\hat{\beta}_{i,p,g}$) using the following regressions:

$$c_{i,p,g,t} = \alpha_{i,p,g} + \beta_{i,p,g} t + \varepsilon_{i,p,g,t}, g \in [6 \ldots 8], \tag{1c}$$

where $\alpha_{i,p,g}$ are constants and $\varepsilon_{i,p,g,t}$ are the error terms. For this method, we required explicit ways to deal with unborn generations. For this purpose, we extrapolated trend estimates ($\beta_{i,p,g}$), that is $\hat{\beta}_{i,p,g}$, on future generations as follows:

$$\hat{\beta}_{i,p,g} = \lambda_{i,p,g} + \gamma_{i,p,g} * p \tag{2c}$$

$$\widetilde{c}_{i,p,g,\dot{t}} = c_{i,p,g,2017} + \hat{\beta}_{i,p,g} * (t - 2017), t \in [2018 \ldots 2050] = \dot{t}, \tag{3c}$$

where $\lambda_{i,p,g}$ is a constant and $\gamma_{i,p,g}$ is a slope of a trend in $\hat{\beta}_{i,p,g}$ within scales $x = p$ and $y = \hat{\beta}_{i,p,g}$. The insignificant estimates were not replaced with zeroes because the hypothesis testing (H_0: zero coefficient, i.e. $H_0 : \gamma_{i,p,g} = 0$) was held on a small sample. Obviously, $c_{i,p,g,2017}$ for generations that do not yet exist was 'N/A' (not applicable, not available, not assessed, or no answer) and would become a number only once these generations are born. At the generation formation year (denoted as $\bar{\bar{t}}$), $c_{i,p,g,\bar{\bar{t}}}$ equals the average consumption growth of all previous generations at the same age and at any time ($\bar{c}_{i,p,g,t}$), as noted in Eq. (4c):

$$\widetilde{c}_{i,p,g,\bar{\bar{t}}} = \frac{1}{p-1} \sum_{p}^{p-1} \bar{c}_{i,p,g,t} \tag{4c}$$

GLM model prediction. After assessing any model with time (or time-fixed effects) and generational variables, time and generational range can be advanced to predict the dependent variable. However, we assumed that estimates of the factors' contributions would remain stable and robust over time. Using a GLM model, we (1) assessed the contributions of all factors (generations and time trend, as well as gross household income, age interval, gender and region) on normalised consumption per person; (2) considered the chronological change (i.e. we advanced the 1990–2017 range with 2020–2050) and the generational range (i.e. we advanced the generational range with three more generations); and (3) predicted future normalised consumption per person for all possible combinations of variables.

For each food (i), social ID (id; see the section "Considering social characteristics") and time (t), we defined consumption ($c_{i,\mathrm{id},t}$) and explained it with a time (t) and a set of other explanatory socioeconomic variables, each denoted as index $k \in [1 \ldots K = 5]$ – that is, $X_{i,\mathrm{id},t,k}$. For the generational variable, $k = \varphi$. The formula for the regressions we assessed is as follows:

$$c_{i,\mathrm{id},t} = \alpha_i + \beta_{0,i} * t + \sum_{k=1}^{K=5} \beta_{k,i} * X_{i,\mathrm{id},t,k} + u_{i,\mathrm{id},t}, \tag{1d}$$

where the regressions were assessed separately for each i, α_i was a constant and $\beta_{0,i}$ was the slope of the trend. We used the time trend to avoid predicting the time-fixed effects for the years after 2017. Estimation of Eq. (1d) served as a trained model to predict the same data set with $t \rightarrow t + 30$ and $X_{i,\mathrm{id},t,k=\varphi} \rightarrow X_{i,\mathrm{id},t,k=\varphi} + 3$. In the above description of the technique, we have omitted a discussion about testing the trained models because we relied mainly on the robust functions of the R packages used. However, it is important to mention that in the case of predicted values over 1 and below 0, we considered them as 1 and 0, respectively. This decision was made for less than 1% of predicted values to avoid vulnerable and negative consumption, respectively.

SuperLearner. SuperLearner approach weighs the predictions of several models to achieve the best receiver operating characteristic (ROC) curve or the highest ROC estimate (Sing et al., 2020). This estimate is analogous to the determination coefficient in regression models. This study follows the guidance of Lantz (2019) and uses the updated SuperLearner function from the SuperLearner package (Polley et al., 2023), which was available at the time the main calculation of this forecast was performed. The R libraries used to conduct the SuperLearner forecast were 'SL.xgboost', 'SL.glmnet', 'SL.lm', 'SL.mean' and 'SL.ranger'.

The binary dependent variable helped to maintain the most powerful models (binomial family) in SuperLearner at the cost of the variation loss during normalisation. As a benefit, the normalisation performed helped control the predicted consumption values within a finite range. In adopting this approach, we assumed that future social groups would not consume more than the maximum for social groups realised across years from 1990 to 2017 and predicted for years from 2018 to 2050. This assumption was reasonable because people in the highest-consuming social group seemed to have a set of socioeconomic characteristics suitable for consuming the most. As in GLM, we kept the characteristics but expanded the year range from 1990–2017 to 2018–2050 and advanced the generational range by three generations.

We tested the models packed in SuperLearner separately on a random portion (half) of the existing data to predict the other random portion (half) of the existing data. Using this method, we checked the model quality estimates (ROCs). More precisely, for the data for chicken, bread, milk and olive oil, we launched the GLM function from the 'stats' package, the 'ranger' function from the 'ranger' package (Wright et al., 2023), and the 'XGBoost' function from the 'XGBoost' package (Chen and Tong, 2023). These models delivered a ROC estimate ranging between 0.6 and 0.8, which was sufficiently high for each of the single techniques.[1] Since SuperLearners allocate the highest weight for the best models, we expected that the predictive power of the final models for different foods should be the highest that a machine can deliver for the available data.

We used all available data to produce the final trained model. Next, we applied the trained model to test data with extended year and generational ranges. The application of the model delivered the forecasted probability of $\widetilde{\pi}$ in each social group ($\widetilde{\pi}_{i,\mathrm{id},\dot{t}}$), this is, a value between 0 and 1.

Additional calculations

Sizes and shares of social groups. To discover the social group sizes for such a large set of variables, we calculated the total number of participants in households that had formed social groups in the past. Specifically, we calculated the social group size (total number of persons in households) in year t ($y_{\ddot{\mathrm{id}},t}$). We then used linear regression for each social group separately (205 extrapolations in our case) and obtained the estimates of the trends ($\hat{\beta}_{\ddot{\mathrm{id}}}$), using the following regressions:

$$y_{\ddot{\mathrm{id}},t} = \alpha_{\ddot{\mathrm{id}}} + \beta_{\ddot{\mathrm{id}}} t + \varepsilon_{\ddot{\mathrm{id}},t}, \tag{5}$$

where $\alpha_{\ddot{\mathrm{id}}}$ are constants and $\varepsilon_{\ddot{\mathrm{id}}}$ are the error terms. Further, we extrapolated the estimates of trends ($\hat{\beta}_{\ddot{\mathrm{id}}}$) on future social group sizes as follows:

$$\widetilde{y}_{\ddot{\mathrm{id}},\dot{t}} = y_{\ddot{\mathrm{id}},2017} + \hat{\beta}_{\ddot{\mathrm{id}}} * (t - 2017), t \in [2018 \ldots 2050] = \dot{t}, \tag{6}$$

We further derived new sizes to calculate the shares of social groups in the total population ($\widetilde{s}_{\ddot{\mathrm{id}},\dot{t}}$, where $\dot{t}$ is any t>2017), with the total size of the population in each year expressed as 100%:

$$\widetilde{s}_{\ddot{\mathrm{id}},\dot{t}} = \frac{y_{\ddot{\mathrm{id}},\dot{t}}}{\sum_{\ddot{\mathrm{id}}}^{\Omega} y_{\ddot{\mathrm{id}},\dot{t}}} * 100\%, \tag{7}$$

where $\widetilde{s}_{\ddot{\mathrm{id}}}$ are the shares of social groups in the total population taken as 100% and 205 types of shorter ID ($\ddot{\mathrm{id}}$) are denoted as a list Ω.

Shares of generations in population. We allocated the population between generations in the same way as in the survey observations (see the section "Household survey") and calculated shares of generations in the total population ($\widetilde{s}_{p,\dot{t}}$), based on the available statistics and predictions for the Swiss population (Fig. 1). The frequency of shares of generations is once per 10 years and restricts the frequency of the forecasted values. The number of participants in some generations increased due to migration.

Mixed-generation households and households with children may consume differently than households with defined generations (Mann and Loginova, 2023). However, there are no statistics on the number or share of mixed-generation households in Switzerland's total population. We cannot solely allocate any share to households with children because it is not clear which generation should change its share to compensate for this redistribution. Accordingly, we had to recognise the bias and assume that the error with regard to mixed-generation households and households with children may be better resolved in future studies with better statistics.

Transition back to variables in levels. The transition back to consumption in levels was performed before weighting the

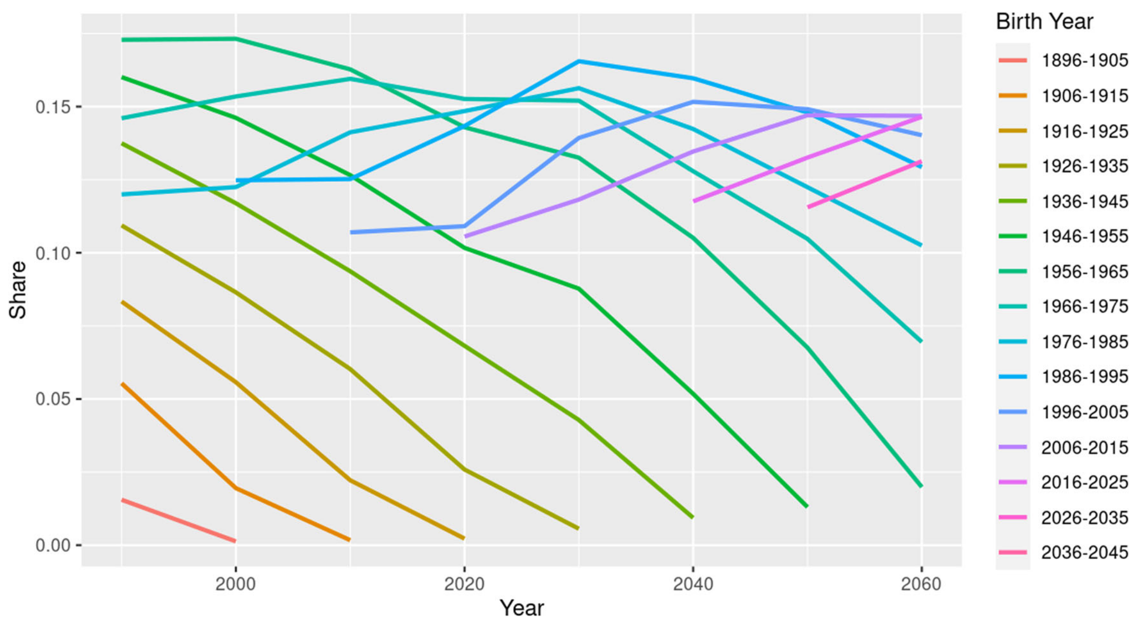

Fig. 1 Shares of generations in population.

outcomes (see the section "Weighting"), as shown in Eqs. (8)–(10):

$$\text{Growth rates } c_{i,\mathrm{id},t} \qquad \dddot{c}_{i,\mathrm{id},t} = \left\{ \begin{array}{ll} \acute{c}_{i,\mathrm{id},1990} * \left(1 + c_{i,\mathrm{id},t}\right)^{t-1990} & , t \in [1991\ldots 2000] \\ \acute{c}_{i,\mathrm{id},t-1} * \left(1 + c_{i,\mathrm{id},t}\right) & , t \in [2001\ldots 2017] \\ \acute{c}_{i,\mathrm{id},\dot{t}-1} * \left(1 + \widetilde{c}_{i,\mathrm{id},\dot{t}}\right) & , \dot{t} \in [2018\ldots 2050] \end{array} \right\} \tag{8}$$

$$\text{Normalised } c_{i,\mathrm{id},t} \qquad \dddot{c}_{i,\mathrm{id},t} = \widetilde{c}_{i,\mathrm{id},t} * \left(c_{i,\ddot{\mathrm{id}},\max} - c_{i,\ddot{\mathrm{id}},\min}\right) + c_{i,\ddot{\mathrm{id}},\min} \tag{9}$$

$$\text{Probability } \widetilde{\pi}_{i,\mathrm{id},\dot{t}} \qquad \dddot{c}_{i,\mathrm{id},t} = \widetilde{\pi}_{i,\mathrm{id},\dot{t}} * \left(c_{i,\ddot{\mathrm{id}},\max} - c_{i,\ddot{\mathrm{id}},\min}\right) + c_{i,\ddot{\mathrm{id}},\min} \tag{10}$$

Weighting

Weighting the outcomes of Model B: age and gender. We used population pyramids—shares of men and women in the total population by age and year ($s_{j,\mathrm{woman},t}$ and $s_{j,\mathrm{man},t}$)—to adjust forecasts ($\dddot{c}_{i,j,g,t}$) on demographic change:

$$\dddot{c}_{i,\dot{t}} = \sum_j \dddot{c}_{i,j,\mathrm{woman},\dot{t}} * s_{j,\mathrm{woman},\dot{t}} + \sum_j \dddot{c}_{i,j,\mathrm{man},\dot{t}} * s_{j,\mathrm{man},\dot{t}} \tag{11}$$

The resulting values of $\dddot{c}_{i,\dot{t}}$ created a nonlinear line, which is a weighted version of a line from the reference Model A, such that Model B considered gender and age distribution over the years. The outcome of Model B can potentially differ significantly from that of Model A if, for example, the loss of men and women (e.g. due to war, illness and poor healthcare), changes in life expectancy, childbirth and mortality rates and sharp changes in migration happen compared to previous years.

Weighting the outcomes of Model C: generation and gender. We used forecasted population pyramids—shares of men and women in total population, predicted until 2050—by allocating people to generations and considering the shares of generations in particular years. Thus, we used population shares by generation, gender and year ($s_{\mathrm{p},\mathrm{man},t}$ and $s_{\mathrm{p},\mathrm{woman},t}$) to adjust forecasts ($\dddot{c}_{i,\mathrm{p},g,t}$) of demographic change in the population pyramid, as follows:

$$\begin{aligned} \dddot{c}_{i,t} &= \sum_{j,\mathrm{p}} \hat{\beta}_{i,\mathrm{p},\mathrm{woman}} * s_{\mathrm{p},\mathrm{woman},t} * \dddot{c}_{i,\mathrm{p},\mathrm{woman},t} \\ &+ \sum_{j,\mathrm{p}} \hat{\beta}_{i,\mathrm{p},\mathrm{man}} * s_{\mathrm{p},\mathrm{man},t} * \dddot{c}_{i,\mathrm{p},\mathrm{man},t} \end{aligned} \tag{12}$$

Weighting the outcomes by data-driven socioeconomic group and generation as in Models D and E: GLM and SuperLearner. In Models D and E, generations explained the predictions together with other socioeconomic explanatories that were not covered by population pyramids. With the consumption values predicted for each social group and generation, we obtained the final values, as described in Eq. (13):

$$\dddot{c}_{i,\dot{t}} = \sum_{\mathrm{id},\mathrm{p}} \dddot{c}_{i,\mathrm{id},\mathrm{p},\dot{t}} * \widetilde{s}_{\ddot{\mathrm{id}}} * \widetilde{s}_{\mathrm{p},\dot{t}} \tag{13}$$

Thus, we weighted the predicted consumption volumes according to social group (see the section "Sizes and shares of social groups ") and generation (see the section "Shares of generations in population") shares of the total population to shape the final forecast ($\dddot{c}_{i,\dot{t}}$). The results of Models D and E differed significantly from those of Model B in the case of structural changes in societal characteristics. In contrast to Model B, for Switzerland, Models D and E may differ strongly from reference Model A.

Limitations and assumptions. The list of limitations and assumptions for the proposed approaches includes, but is not limited to:

1. A bias towards mixed-generation and mixed-age households (those with children and migrant populations in all the studied models): We excluded these households from the sample to avoid distorting the results and assumed that these households did not crucially change the results. This limitation may be addressed by including more question variety in the surveys or cleverer approaches to weighting in future studies, for example, by using inverse probability weighting (Abadie, 2005) for the shares or weights of social groups. In our estimations, consumption was measured per person on the adult data, so that children got the adult estimate of food consumption.
2. Bias in our analysis compared to analyses conducted with a wider or more accurate list of explanatory variables: This limitation may be addressed with a higher number and variety of respondents and more detailed questions in the survey. For example, household gender was binary in our estimations because the respondents from the earliest years of our dataset could only choose between male or female. Likewise, the regional language of households in our study was binary; even for a small territory, such as Switzerland, this approximation can be difficult. Therefore, using a higher detail of the cultural background of the population is a priority for future studies. For future studies that take a wider range of characteristics, our methodological background fits and is adaptable.
3. A group homogeneity assumption: We assumed that respondents with the same sets of characteristics were equal enough to be represented with an average. We also kept an equal-share gender distribution in most of our calculations, despite the raw data being slightly unbalanced by gender. The potential impact of these assumptions and biases is that consumption of smaller social groups and of men is slightly overrepresented in the prognosis for the total population.
4. The reliability and precision of food data: This assumption may be avoided with more detailed and stably classified product data sets that consider energy content, processing, waste, new foods and the weight of packaging and bones.
5. The potential inaccuracy of any approach based on a linear forecast, which may be addressed by using Seasonal Autoregressive Integrated Moving Average (SARIMA) and Generalised Autoregressive Conditional Heteroskedasticity (GARCH) forecasts, rather than linear ones, in future studies.
6. The reliability and precision of the models themselves and between-models comparisons, which is solvable by using SuperLearners with more models and data in future studies.
7. The assumption that the contributions of the studied factors will remain the same in 1990–2050, despite the degree to which societal characteristics changed between 1990 and 2017: Our data showed that income inequality between social groups in Switzerland increased and the number of households with three or more members halved over the time studied. Among those without children, the share of people over 65 years of age increased from 25% to 37%. These changes were coupled with the generational change presented in Fig. 1. Nevertheless, we maintained this strong assumption because, to the best of our knowledge, it

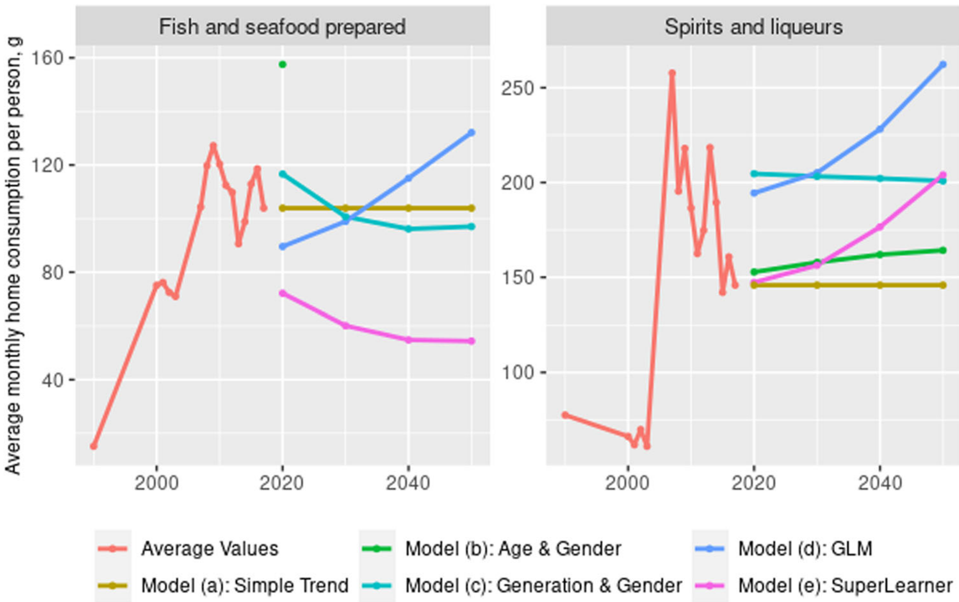

Fig. 2 Schematic visualisation of the results of projecting food consumption.

may be addressed only by methods with evolving estimates that are not yet developed, again to the best of our knowledge.

8. The assumption that social structure will develop between 2020 and 2050 as it did in the years before 2017: This assumption may be avoided by undertaking separate studies on demographic changes using the same methodology that was used for consumption in this study. We attempted to fit such an estimation into this study in the earlier stages of research but were forced to leave this idea to further demographic studies because our data were appropriate only for calculating average consumption and were not rich enough to predict population dynamics using growth rates. Nevertheless, the methodology provided in this study may be used to study population dynamics with better data.
9. Neglecting biological cycles in generations, for example, changes in consumption due to ageing. As shown in Eq. (1c), the 30 years of data and the ability to define all generations within this time span is a minimum requirement for building generational trends on three points ($g \in [6 \ldots 8]$ in Eq. (1c)) when using growth rates. This limitation may be addressed with longer-term data and a variety of respondents, which will only become available in the future.
10. Finally, the constant growth rates may not reflect real historical change. Sensitivity analyses or alternative imputation methods could be considered to validate this approach. The research could benefit from sources that explore multidisciplinary approaches to food consumption forecasting, in particular, behavioural economics, sociology, food sustainability, environmental science, as well as food security and food safety. Adding global and regional comparisons in food consumption trends could provide a broader context for the research findings. Including recent advances in big data analytics and AI for predictive modelling can strengthen its basis in future research.

Results

Approaches that use generations as explanatory or weights have the potential to predict changes in future consumption dynamics. As demonstrated in Fig. 2, the lines for projections that consider generations are convex or concave because the shares of generations in the population are used to weigh the outcomes of these models. The lines for simple forecasts were constant, as there were no *significant* levels or trends in growth rates. These results apply to all the studied products.

Table 2 presents a summary of the results for all 75 foods. At this stage, we considered all studied changes, contributions and their significances and compare only the outcome projections from the section "Weighting" for the years 2020 and 2050. The average annual projected consumption changes between 2020 and 2050 make it possible to compare the outcomes of different models for the variety of products and their consumption dynamics.

Generational and SuperLearner approaches did not fail as fast and as often as comparable (by a set of included factors) forecasts. For example, SuperLearner helped to avoid zero and negative consumption projections from GLM with many factors for at least 16 foods: canned meat, cream, jam, leafy vegetables, milk, mineral water, mushrooms, nuts, potatoes, stone fruit, sugar, tea and herbs, veal, vegetarian soy products, wild and rabbit meat and yogurt. In addition, apples, butter, coffee, margarine, root vegetables, potatoes, and tomatoes receive SuperLearner forecast when GLM fails.

Generational forecasts require longer data; therefore, they cannot be calculated more often than age and gender forecasts. However, generational forecasts that are possible to calculate are rarely implausible compared to age and gender forecasts. Examples illustrating this are dried vegetables and mushrooms, prepared fish and seafood, honey, jam and olive oil. Considering that consumption of most listed foods is not ever likely to be zero (on average), the methodological achievement of SuperLearner compared to GLM and the generational approach compared to age and gender approach is worth noticing. Additionally, SuperLearner and generational projections did not change as sharply as linear and GLM forecasts in many cases (e.g. honey, olive oil and prepared fish and seafood).

The dynamics of the forecasts differed over time. In the long run, bananas, beer, bread, cheese and curd, poultry, wines and nonalcoholic drinks received diverse forecasts according to the simplest and most complicated forecasting and weighting techniques.

Table 2 The results for all projectable food items and all described methods.

Food group	Food (i)	Average annual changes in consumption in 2020–2050 (%) in Models			
		B	C	D	E
Meat	Beef	−0.2	0.07	0.19	0.13
	Ham and bacon	0.05	−0.01	−1.59	−0.15
	Horse meat	0.02	0.08	-	-
	Pork	0.08	0.04	−4.6	−1.19
	Poultry	−0.08	0.01	0.76	0.8
	Sausages	−0.03	0.01	−3.61	−0.31
	Seafood	0.1	−0.1	1.44	0.46
	Sheep and goat meat	0.04	0.18	−7.28	−0.65
	Veal	0.29	−0.04	lost	−1.45
	Wild and rabbit meat	0.29	−0.11	lost	−0.35
	Meat, canned	0.37	−0.11	lost	−0.47
	Meat, other	−0.87	0.09	−0.91	−0.23
Fish	Fish	−0.12	0.06	0.46	0.49
	Fish and seafood prepared	lost	−0.61	1.3	−0.94
	Fish, canned	0.12	lost	1.19	−0.03
Dairy	Milk	−0.94	0.05	lost	−2.2
	Mixed milk-based products	0.04	-	1.06	0.26
	Cream	−0.39	0.04	lost	0.72
	Butter	0.26	0	-	−0.44
	Ice cream	−0.08	-	-	-
	Yoghurt	−0.1	0.11	lost	−0.69
	Cheese and curd	0.04	−0.03	1.03	0.03
	Egg	0.19	-	1.07	0.38
	Honey	lost	−0.18	1.23	0.07
Foods that may use animal food components	Baby food	0.12	lost	1.25	0.37
	Bread	3.47	0.01	−1.81	−0.07
	Cocoa and chocolate	−0.04	-	1.08	0.32
	Confectionery	0.05	0.04	1.61	−0.31
	Margarine	0.15	-	-	−1.65
	Oils and fats (except olive oil)	−0.41	0.03	0.24	−0.27
	Pasta	−0.06	−0.05	0.8	0.53
	Pastry	0.03	-	0.25	0.12
	Ready meals	−0.25	-	-	-
	Soups	-	-	-	-
Vegetables, beans, peas and mushrooms	Beans and peas	−0.07	−0.1	0.69	0.04
	Cabbage vegetables	0.01	−0.46	1.17	0.07
	Dried fruits	−0.02	−0.19	0.21	−0.59
	Dried vegetables and mushrooms	excluded	−0.67	-	
	Grapes	0.04	−0.02	−0.66	−0.3
	Kitchen herbs	0.08	−0.01	1.13	−0.21
	Leafy vegetables	0.08	0	lost	−0.41
	Mushrooms fresh	−0.04	0.06	lost	−2
	Onions and garlic	0.05	−0.06	1.17	0.06
	Potatoes	−0.03	0.17	lost	0.03
	Root vegetables	−0.13	0.01	-	−0.04
	Tomatoes	0.17	0.01	-	−0.31
	Vegetables (stem and fruit)	0.01	-	lost	0.23
	Vegetables and mushrooms, canned	0.2	−0.55	1.83	0.82
Fruits, Nuts, Jam	Apples	0.16	−0.04	-	−0.05
	Bananas	0.1	−0.03	1.29	0.9
	Berries	0.22	−0.09	0.95	0.43
	Canned fruit	0.05	−1.1	0.58	−0.07
	Citrus (except lemons)	0.05	0.01	−3.39	−0.92
	Jam	lost	−0.35	lost	−0.04
	Lemons	0.13	−0.26	-	-
	Nuts	0.09	0.16	lost	−0.49
	Pears and quinces	0.91	−0.06	1.32	0.07
	Stone fruit	0.07	−0.03	lost	−1.05
	Other fruits	0.01	0.01	1.37	0.69
Liquids	Beer	−0.14	0.08	1.06	0.39
	Coffee and substitutes	0.06	0.25	-	0.08
	Mineral water	−0.14	0.17	lost	−0.45
	Nonalcoholic drinks	−0.99	0.06	0.15	−0.07

Table 2 (continued)

Food group	Food (i)	Average annual changes in consumption in 2020–2050 (%) in Models			
		B	C	D	E
	Olive oil	excluded	−0.44	1.47	0.9
	Spirits and liqueurs	0.24	−0.06	1	1.09
	Syrups	−0.14	lost	-	-
	Wines	0.11	−0.02	−0.42	−0.11
Granular products	Aroma and taste essences	−0.03	-	-	-
	Cereal products	−0.07	0.01	0.37	−0.22
	Flours	0.06	−0.04	−1.79	−0.66
	Rice	−0.22	0.04	−0.88	−0.43
	Sugar	0.08	−0.02	lost	−1.3
	Tea and herbs	0.2	0.01	lost	−0.09
Other	Vegetarian soy products	−0.11	-	-	-
	Other foods	−0.2	-	0.51	0.11

The values are calculated as $\left(\left(\frac{c_{i,2050}}{c_{i,2020}}\right)^{\frac{1}{30}} - 1\right) * 100$ for each model separately. We do not apply margins of statistical error and significance levels because each number in Table 2 is a prognosed change in 2020–2050 distributed across years. Empty values indicate the forecasts that cannot be calculated, 'lost' indicates negative or zero forecast consumption, and 'excluded' indicates unreliably high forecasted growth. The results for Model A, 'simple trends', demonstrated no change in future food consumption; therefore, we omitted them for table consistency. 'B'–'E' indicate 'Model B'–'Model E', respectively.

Regarding the growth shown in Table 2 the most diverse forecasts occurred for sheep and goat meat, ready meals, bread, pork, other foods, sausages and citrus (except lemons), mostly due to the vulnerability of GLM forecasts. Demands for aroma and taste essences (−0.03%), ice cream (−0.08%), vegetarian soy products (−0.11%) and ready meals (−0.25%) will implausibly decline according to age and gender projections, and the data for these products is insufficient to validate alternative estimates using other techniques. Although this negative scenario is possible if interest in these foods declines and if these foods are replaced by vegan products under aggressive labelling, the issue of data shortage for these foods remains. Therefore, these foods demand further research in food sectors, more data, improvement of statistical evidence with regard to weighting techniques, narrowing the standard errors of forecasts.

The reason for the differences in forecasts may lie in the predicted preferences of future social groups. We hope that this is true because most of the forecasts indicate an average movement towards healthier diets. However, the more technical reason may be that the normalisation, binarisation and stationarisation of a dependent variable did not make it possible to obtain a sufficiently balanced dependent variable with which to train the models for these foods. This issue might have been solved by stricter data cleaning, which we did not perform for reasons discussed in the section "Household survey", and by other means of improving the methodology in future studies, as discussed in the section "Limitations and assumptions".

Discussion and conclusion

Based on 20 million consumption observations across 46,456 households in Switzerland, 75 foods, and the years 1990–2017, we forecasted food demand for the Swiss population until 2050 for each food separately. In our case, the predicted period has the same length as the available time series of the past. Nevertheless, we performed these calculations to monitor the behaviour of our forecasts in the long run and to track implausible values during the prediction horizon. Only Models D and E consider the classification of key factors of individual consumption by Asp (1999), namely, cultural and lifestyle factors (e.g. region), psychological factors (e.g. aggregated gender, generation), and barriers (e.g. income); however, Models B and C already consider significantly more consumption factors compared to previously applied techniques.

The main contribution of this paper is its demonstration of (1) how to integrate generational changes into forecasts and predict turning points in consumption dynamics, as well as (2) how to act if simple linear forecasts fail to predict a change. Many foods were forecasted consistently by all the techniques, which suggests that this proposition is not only possible but that generations and other socioeconomic characteristics are influential in future food consumption. Accordingly, this paper contributes to the debate on food consumption, food consumption prediction and statistical approaches to considering the changing tastes of population groups alongside changing population group distributions in general.

For socioeconomists, the paper presents an additional, beautiful concept related to the study of food and generations. This study is the first to attempt to predict the (eating) behaviour of future generations using big data collected from previous generations. To achieve this, we ordered generations by birth year (a common technique), assigned them ordered numbers and conceptualised this order as a new scale for analysing generational trends. The most interesting cases included foods with intergenerational trends either in the observed data or between the forecasted values. In any case, this paper suggests the first quantitative technique to shed light on how common human eating behaviour trends can be used to forecast consumption decades ahead, even for people who are not yet born. Of course, the approaches to creating a line between generations may vary, and our contribution is ultimately a suggestion to consider consumption from the perspective of ordered slopes of generational trends.

This paper also makes a relatively small contribution to demographic science, as we use not only an age–gender pyramid, but also generational and social group forecasting. The last two techniques may be of interest to demographers if they are willing to strengthen future population forecasts and, more importantly, future population structures. Again, the linearity of our predicted future values is an obvious simplification of the task of predicting the dynamics of hundreds of social groups in the future. Future population forecasting techniques may rely strongly on population groups and become a strong alternative to linear predictions in current use.

A further contribution of this paper is the application of binary SuperLearner prediction techniques to volume data by scaling it to the [0;1] interval and exploiting the prediction of probabilities as a prediction of shares. This hook might be obvious for mathematicians and data analysts, but in food consumption volume studies, this is novel and will perhaps serve as a convenient guide for future studies of volumes and their development over time. The main reason for taking this approach is the rapid development of models with binary outcomes. The predicting power and variety of models developed in recent years have increased tremendously; thus, we decided to adopt this positive development rather than developing another big model for the numeric dependent variable. This adoption was a success for many foods, but there are various opportunities for improvement, which have been discussed throughout this paper.

Finally, our study has practical and theoretical implications for policymakers involved in analysing and shaping food trends. As mentioned above, food demand plays an important role not only in the sale of food but also in the management of stocks, logistics, food production, supply chains, disease risks and the environmental footprint. The food sector is interlinked with other sectors and requires resources. Therefore, our projections are of practical relevance for most actors in the food sector. We also introduce the idea that the demand for food changes as new generations replace older ones. This idea should serve as food for thought for stakeholders and policymakers, for example, if their policies and actions are geared towards sustainability and "do not compromise the needs of future generations" (Foresight, 2011, p. 204).

Time will tell which of our predictions, if any, will align with reality; however, the range of values that we predicted significantly limited the possible outcomes, so that it is, for example, very likely that our demand for eggs will grow, whereas our demand for jam will shrink. This work may provide a positive impulse for further statistical and applied studies. We considered as many approaches as possible to the best of our knowledge, and the limitations section will be useful for launching new research from a position of partial development rather than from scratch. By comparing the outcomes of various prediction techniques, we have facilitated a better understanding of these models and have hopefully assisted many researchers in their work worldwide.

Data availability

The data that support the findings of this study are available from the Swiss Federal Statistical Office, but restrictions apply to the availability of these data, which were used under licence for the current study and so are not publicly available. The data are, however, available from the authors upon reasonable request and with the permission of the Swiss Federal Statistical Office. The authors confirm that the codes supporting the findings of this study are available within the supplementary material.

Received: 21 December 2023; Accepted: 7 October 2024;
Published online: 28 October 2024

Note

1 We also checked the decision tree (the 'trainControl' function from the 'caret' package; Kuhn et al., 2023) and the gradient boosting machine (GBM; the 'gbm' function from the 'gbm' package by Greenwell et al., 2022). The decision tree delivered a low ROC estimate 0.5 and was thus excluded from SuperLearner, and GBM delivered a small training error on 10,000 trials, but we did not include this technique due to a mismatch between variable R formats and the other models, which made it difficult to run all the models together.

References

Abadie A (2005) Semiparametric difference-in-differences estimators. Rev Econ Stud 72(1):1–19. http://www.jstor.org/stable/3700681

AGRISTAT (2023) Agristat—statistics on Swiss agriculture. https://www.sbv-usp.ch/de/services/agristat-statistik-der-schweizer-landwirtschaft/agristat-aktuell

Asp EH (1999) Factors affecting food decisions made by individual consumers. Food Policy 24(2–3):287–294. https://doi.org/10.1016/S0306-9192(99)00024-X

Athey S, Imbens GW (2019) Machine learning methods that economists should know about. Annu Rev Econ 11(1):685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Baur I, Stylianou KS, Ernstoff A, Hansmann R, Jolliet O, Binder CR (2022) Drivers and barriers toward healthy and environmentally sustainable eating in Switzerland: linking impacts to intentions and practices. Front Sustain Food Syst 6. https://doi.org/10.3389/fsufs.2022.808521

Benjamin D (1992) Household composition, labor markets, and labor demand: testing for separation in agricultural household models. Econometrica 60(2):287. https://doi.org/10.2307/2951598

Bodirsky BL, Rolinski S, Biewald A, Weindl I, Popp A, Lotze-Campen H (2015) Global food demand scenarios for the 21st century. PLoS ONE 10(11):e0139201. https://doi.org/10.1371/journal.pone.0139201

Bouclaous C, Jaubert R (2015) The use of household budget surveys to estimate the availability of fruits and vegetables for consumption in Swiss households after deduction of food waste. Glob J Med Pub Health 4(5):1

Bozkir AS, Sezer EA (2011) Predicting food demand in food courts by decision tree approaches. Procedia Comput Sci 3:759–763. https://doi.org/10.1016/j.procs.2010.12.125

Brown JAC (1954) The Consumption of food in relation to household composition and income. Econometrica 22(4):444. https://doi.org/10.2307/1907436

Callaway B, Sant'Anna PHC (2021) Difference-in-Differences with multiple time periods. J Econ 225(2):200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Chen T, Tong H (2023) Xgboost: eXtreme gradient boosting, package version: 1.7.5.1. https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf

Czarniecka-Skubina E, Kowalczuk I (2015) Eating out in Poland: history, status, perspectives and trends. Serv manag 16:75–83. https://doi.org/10.18276/smt.2015.16-08

Daniel CR, Cross AJ, Koebnick C, Sinha R (2010) Trends in meat consumption in the USA. Public Health Nutr 14(4):575–583. https://doi.org/10.1017/s1368980010002077

Federal Statistical Office (2020) National projections. https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/zukuenftige-entwicklung/schweiz-szenarien.html

Federal Statistical Office (2022) Household budget survey. https://www.bfs.admin.ch/bfs/en/home/statistics/economic-social-situation-population/surveys/hbs.html#:~:text=The%20purpose%20of%20the%20Household,conducted%20on%20a%20continuous%20basis

Flies EJ, Brook BW, Blomqvist L, Buettel JC (2018) Forecasting future global food demand: a systematic review and meta-analysis of model complexity. Environ Int 120:93–103. https://doi.org/10.1016/j.envint.2018.07.019

Food and Agriculture Organization of the United Nations (2017) The future of food and agriculture. Trends and challenges. Food and Agriculture Organization of the United Nations, Rome

Food and Agriculture Organization of the United Nations (2018) Sustainable food systems. Concept and framework. Food and Agriculture Organization of the United Nations, Rome

Foresight (2011) The future of food and farming: challenges and choices for global sustainability. Final project report. Government Office for Science, London https://assets.publishing.service.gov.uk/media/5a7bf9f840f0b645ba3c5efe/11-546-future-of-food-and-farming-report.pdf

Gaure S, McDermott G, Vargas M, Werner KD, Stigler M, Lüdecke D (2023) 'lfe' Linear group fixed effects. https://rdocumentation.org/packages/lfe/versions/2.9-0

Girshick MA, Haavelmo T (1947) Statistical analysis of the demand for food: examples of simultaneous estimation of structural equations. Econometrica 15(2):79. https://doi.org/10.2307/1907066

Godin L, Sahakian M (2018) Cutting through conflicting prescriptions: how guidelines inform "healthy and sustainable" diets in Switzerland. Appetite 130:123–133. https://doi.org/10.1016/j.appet.2018.08.004

Greenwell B, Bradley B, Cunningham J, GBM Developers (2022) Package 'gbm', Version: 2.1.8.1, Generalized boosted regression models. https://github.com/gbm-developers/gbm

Harshini K, Madhira PK, Chaitra S, Reddy GP (2021) Enhanced demand forecasting system for food and raw materials using ensemble learning. In: 2021 International conference on artificial intelligence and machine vision, IEEE. https://doi.org/10.1109/aimv53313.2021.9671005

Inanir D, Kaelin I, Pestoni G, Faeh D, Mueller N, Rohrmann S, Sych J (2020) Daily and meal-based assessment of dairy and corresponding protein intake in Switzerland: results from the National Nutrition Survey menuCH. Eur J Nutr 60(4):2099–2109. https://doi.org/10.1007/s00394-020-02399-7

Iyer N (2020) Forecasting food demand applying neural networks to the meal kit industry. Towards Data Sci. https://towardsdatascience.com/forecasting-food-demand-applying-neural-networks-to-the-meal-kit-industry-6f1e3b2207aa

Kearney J (2010) Food consumption trends and drivers. Philos Trans R Soc B: Biol Sci 365(1554):2793–2807. https://doi.org/10.1098/rstb.2010.0149
Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al (2023) package 'caret' 6.0-94, classification and regression training. https://github.com/topepo/caret/
Lantz B (2019) Machine learning with R, 3rd edn. Packt Publishing Ltd
Lime M (2022) Food demand forecasting R2 0.883. Kaggle Notebook. https://www.kaggle.com/code/mattialime/food-demand-forecasting-r2-0-883/notebook
Loginova D, Mann S (2022) Measuring stability and structural breaks: applications in social sciences. J Econ Surv 37(2):302–320. https://doi.org/10.1111/joes.12505
Loginova D, Mann S (2024) Sweet home or battle of the sexes: who dominates food purchasing decisions? Humanit Soc Sci Commun 11(261). https://doi.org/10.1057/s41599-024-02745-8
Lutoslawski K, Hernes M, Radomska J, Hajdas M, Walaszczyk E, Kozina A (2021) Food demand prediction using the nonlinear autoregressive exogenous neural network. IEEE Access 9:146123–146136. https://doi.org/10.1109/access.2021.3123255
Mackle M, David BD (1976) Developing a demand forecasting system for a foodservice operation. J Am Diet Assoc 68(5):457–460. https://doi.org/10.1016/s0002-8223(21)06591-3
Mann S, Loginova D (2023) distinguishing inter- and pangenerational food trends. Agric Food Econ 11(10). https://doi.org/10.1186/s40100-023-00252-z
Massachusetts Institute of Technology (2020) Food demand forecasting challenge. https://github.com/Erdos1729/food-demand-forecasting
Mink S, Loginova D, Mann S (2023) Wolves' contribution to structural change in grazing systems among Swiss alpine summer farms: the evidence from causal random forest. J Agric Econ. https://doi.org/10.1111/1477-9552.12540
Perego L (2019) Food demand forecasting with machine learning: the prospects of cross-sectional training. Doctoral dissertation, Politecnico Milano
Petropoulos F, Carver S (2019) Forecasting for food demand. In: Sustainable food supply chains. pp. 237–248. Elsevier. https://doi.org/10.1016/B978-0-12-813411-5.00016-8
Polley E, LeDell E, Kennedy C, Lendle S, van der Laan M (2023) CRAN—Package SuperLearner. https://CRAN.R-project.org/package=SuperLearner
Pujara A, Pattabiraman V, Parvathi R (2022) Food demand forecast for online food delivery service using CatBoost model. EAI/Springer innovations in communication and computing. https://doi.org/10.1007/978-3-030-78750-9_9
Sahakian M, Godin L, Courtin I (2020) Promoting 'pro', 'low', and 'no' meat consumption in Switzerland: the role of emotions in practices. Appetite 150. https://doi.org/10.1016/j.appet.2020.104637
Sing T, Sander O, Beerenwinkel N, Lengauer T, Unterthiner T, Ernst FGM (2020) CRAN—Package ROCR. https://CRAN.R-project.org/package=ROCR
Smith P (2014) Malthus is still wrong: we can feed a world of 9–10 billion, but only by reducing food demand. Proc Nutr Soc 74(3):187–190. https://doi.org/10.1017/s0029665114001517
Tarallo E, Akabane GK, Shimabukuro CI, Mello J, Amancio D (2019) Machine Learning in predicting demand for fast-moving consumer goods: an exploratory research. IFAC-PapersOnLine 52(13):737–742. https://doi.org/10.1016/j.ifacol.2019.11.203
Tintner G (1953) The definition of econometrics. Econometrica 21(1):31. https://doi.org/10.2307/1906941
United Nations (2019) Growing at a slower pace, world population is expected to reach 9.7 billion in 2050 and could peak at nearly 11 billion around 2100. UN Report - United Nations Sustainable Development
van Dijk M, Morley T, Rau ML, Saghai Y (2021) A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. Nat Food 2(7):494–501. https://doi.org/10.1038/s43016-021-00322-9
Willersinn C, Mack G, Mouron P, Keiser A, Siegrist M (2015) Quantity and quality of food losses along the Swiss potato supply chain: stepwise investigation and the influence of quality standards on losses. Waste Manage 46:120–132. https://doi.org/10.1016/j.wasman.2015.08.033
Wood SD (1977) A model for statistical forecasting of menu item demand. J Am Diet Assoc 70(3):254–259. https://doi.org/10.1016/s0002-8223(21)06418-x
Wright MN, Wager S, Probst P (2023) Package 'ranger'. https://cran.r-project.org/web/packages/ranger/ranger.pdf

Author contributions

DL was responsible for conceptualisation, software, data curation, validation and analysis. SM performed conceptualisation, supervision and reviewing. Both authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Ethical approval

The authors confirm that this article does not contain any studies with human participants performed by any of the authors. The authors confirm that the data used in this research are not identifiable.

Informed consent

The authors confirm that this article does not contain any studies with human participants performed by any of the authors. The authors confirm that the data used in this research are not identifiable.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-024-03890-w.

Correspondence and requests for materials should be addressed to Daria Loginova.