LETTER • OPEN ACCESS

Using social media audience data to analyse the drivers of low-carbon diets

View the article online for updates and enhancements.

ENVIRONMENTAL RESEARCH
LETTERS

CrossMark

LETTER

Using social media audience data to analyse the drivers of low-carbon diets

Sibel Eker[1,2,*], David Garcia[3,4,5], Hugo Valin[1] and Bas van Ruijven[1]

1   International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
2   Climate Interactive, Washington, DC, United States of America
3   Graz University of Technology, Graz, Austria
4   Complexity Science Hub, Vienna, Austria
5   Medical University of Vienna, Vienna, Austria
*   Author to whom any correspondence should be addressed.

E-mail: eker@iiasa.ac.at

Abstract

Low-carbon lifestyles are key to climate change mitigation, biodiversity conservation, and keeping the Earth in a safe operating space. Understanding the global feasibility and drivers of low-carbon lifestyles requires large scale data covering various countries, demographic and socioeconomic groups. In this study, we use the audience segmentation data from Facebook's advertising platform to analyse the extent and drivers of interest in sustainable lifestyles, plant-based diets in particular, at a global level. We show that formal education level is the most important factor affecting vegetarianism interest, and it creates a sharper difference in low-income countries. Gender is a strong distinguishing factor, followed by national gross domestic product per capita and age. These findings enable upscaling local empirical studies to a global level with confidence for integrated assessments of low-carbon lifestyles. Future studies can expand this analysis of social media audience data to other consumption areas, such as household energy demand, and can also contribute to quantifying the psychosocial drivers of low-carbon lifestyles, such as personal and social norms.

1. Introduction

Low-carbon lifestyles, comprised of sustainable choices in various consumption areas from food to energy, are considered a key mitigation option to tackle climate change [1, 2]. Besides lowering the resource demand and greenhouse gas emissions [3], lifestyle change has a strong potential to limit environmental pressure [4, 5], to create co-benefits for multiple sustainable development goals (SDGs) regarding public health, poverty and biodiversity [3, 6, 7], and to reduce the intensity of SDG trade-offs [8].

Achieving the full potential of lifestyle change requires widespread societal transformation. The feasibility of this transformation and how it can be facilitated is yet unknown, because lifestyle change is a complex phenomenon driven by various social, economic, cultural and psychological factors. Quantitative scenario analyses that explore the contribution

of lifestyle change to climate mitigation and sustainable development urgently need to address this complexity. However, the lack of large scale data about the societal heterogeneity of pro-environmental consumption behaviour hinders such quantitative integrated analyses on the feasible potential of lifestyle change.

Theoretical and empirical studies provide a growing understanding of pro-environmental behaviour [9–14], hence shed light on the bottom-up feasibility of lifestyle change. However, such empirical studies are limited temporally, geographically and contextually [10, 12]. In other words, they are based on case studies and surveys that are conducted in a limited number of countries at a particular time, for a particular lifestyle domain and from a particular disciplinary perspective such as behavioural economics or environmental psychology [15, 16]. Therefore, such empirical studies may not allow generalization and large-scale experimentation

to understand the feasible mitigation potential and key drivers of lifestyle change especially in interdisciplinary studies such as integrated assessment modelling [17]. Furthermore, most empirical studies are bounded by *self-reported data,* which may be biased by response styles such as socially desirable or acquiescent responding [18, 19], hence often differ from actual actions and consumption behaviour that can be better measured by observed data [12].

Big data sources, i.e. individuals' and households' online data footprint, can address these limitations of conventional data sources by helping to understand personal carbon footprints, lifestyle change tendencies and their drivers [20]. Online social media (OSM) data, as a publicly available big data source, is particularly promising since it can shed light on socioeconomic, demographic, cultural and even psychological drivers of consumption and lifestyle change. OSM data has been used to analyse several social phenomena such as epidemics [21, 22]; obesity prevalence [23], food choices [24], human migration [25–27], disaster damage and risk perception [28], and gender inequality [29, 30]. However, the use of OSM data for estimating the demand and understanding societal heterogeneity in key consumption sectors behind environmental degradation has been limited to a few studies in the transport sector [31, 32].

OSM data provides large-scale information harmonized across different countries that cannot be obtained from surveys with inevitably limited sample size. OSM data might reflect 'observed' data as opposed to self-reported, since it is based on users' posts, activities, purchases and other online behaviour. Still, social media data has several limitations. It is biased towards the users of these online platforms and may not represent the entire population. Accessing the commercial platforms for data collection may not be straightforward. The publicly available data is aggregated from individual data based on black-box algorithms, it may not fully and transparently represent the actual consumption behaviour, also because online and offline behaviour can still differ. Therefore, OSM data is a promising source to investigate and quantify global societal trends and heterogeneity behind lifestyle change for demand-side climate change mitigation, but its usability in this context should be investigated due to potential limitations.

The objective of this paper is to explore the usability of OSM data to analyse the drivers of low-carbon lifestyles and identify the relative impact of demographic factors such as age, gender and education level on population-wide lifestyle change interest. We particularly demonstrate how Facebook audience segmentation data published for advertising purposes can be used to quantify the societal heterogeneity of global interest in low-carbon diets. For this purpose, we created a dataset of daily and monthly active users (DAU and MAU) marked by pre-defined *interests* in

sustainable lifestyles, particularly vegetarianism. We retrieved publicly available and anonymous data from Facebook marketing application programming interface (API) as described in the section 2. We collected the audience size data at multiple points between September 2019 and June 2020, for each interest category, age, gender, education level and country. This dataset covers 131 countries and around 1.9 billion people as the total Facebook audience size in those countries, 210 million interested in vegetarianism, and 33 million interested in sustainable living (supplementary table 1 (available online at stacks.iop.org/ERL/16/074001/mmedia)).

2. Methods and data

To explore the usability of OSM data, we first collected the audience segmentation data from Facebook marketing API. The lack of large-scale and reliable survey data on the interest in low-carbon lifestyles impedes a precise validation. Still, as an initial validation step, we compare the Facebook audience size data to the limited empirical data from scientific and market-research surveys, to Google Trends data as another indicator of online interest, and to food consumption trends based on UN Food and Agriculture Organization's (FAO) statistics. We then analyse the relationship between the Facebook audience's interest in low-carbon diets, GDP per capita and mean years of schooling (MYS) at the country level using multiple linear regression. Lastly, we identify the key drivers of interest in low-carbon diets based on the granular Facebook data using machine learning (ML) techniques. We describe these data and methods below.

2.1. Data collection
2.2. Facebook
We collected the Facebook audience size data using a Python interface called pySocialWatcher [33] to the Facebook Marketing API [34]. The audience size data is freely available to any registered advertiser on Facebook, and Facebook Marketing API includes only aggregated and publicly available data. Therefore, we had no access to and we did not use any personal information in this study.

The Marketing API allows targeting specific population groups with queries on demographic factors such as age, gender, location, education level, and interest categories that refer to social, economic, and cultural interests like *soccer, yoga* or *agriculture.* While the demographic factors are mostly user-defined, interests are inferred by Facebook algorithms according to what people share on their timelines, apps they use, ads they click, pages they like and other activities related to things like their device usage and travel preferences [35].

In this study, we chose two interest categories relevant for low-carbon lifestyles, *vegetarianism* and *sustainable living.* We determined these interest IDs

based on a keyword search on the Marketing API for available interest categories. For instance, a query with the keyword 'vegetarian' returns the interest categories *vegetarianism, vegetarian cuisine, lacto vegetarianism* etc. We chose *vegetarianism* and *sustainable living* since they are the ones with the highest global audience size among the interest categories returned for respective searches. Supplementary table 1 shows the interest IDs and global audience sizes returned by a keyword search. In addition to the audience sizes of specific interests, we collected the total audience size data for each demographic group (age, gender, education, country) without any interest constraint so that the fractional interest of this demographic group in a subject could be calculated.

Our choice of the keyword *vegetarianism* was motivated by the breadth of the term compared to plant-based diets or sustainable diets, and its availability as a pre-defined interest on the Facebook advertising platform. Interest in vegetarianism can be motivated by different reasons, such as animal welfare, health and religion. Therefore, vegetarianism interest analysed in this study is an indicator of the spread of meat-free diets, more relevant for estimating food demand, but not an indicator of vegetarianism interest only for pro-environmental reasons.

The Marketing API returns two metrics for audience size: DAU and MAU. MAU can be a better estimate of the target group since not everyone uses Facebook every day. However, the Marketing API returns rounded numbers for MAU, for instance 1000 by default for very small audience sizes that have zero DAU. This potentially leads to an overestimate of the actual audience size. Therefore, we use DAU as the metric for audience size throughout this study.

The audience size returned by the Marketing API reflects the present use of Facebook and does not include a temporal dimension. To account for the changes over time either in the actual user interests or in the definition of these interests in the Facebook algorithms, we collected the audience size data in September 2019, January 2020 and June 2020. While the January dataset includes only the total audience sizes in each country, September and June datasets are disaggregated for age, gender and education. Supplementary figure 1 shows that both absolute and fractional audience size for vegetarianism increased from September to June in almost all 50 countries with the highest fractions of vegetarianism interest. Table 1 summarizes the dimensions and size of each audience size dataset. For instance, the June data contains the audience size for two interest groups, for 11 age cohorts, 2 genders, 6 education levels, and 132 countries. This corresponds to 17292 data points for each interest.

Facebook data is biased towards the internet users, for instance young, urban, educated demographic groups, hence may not represent the entire population. In the social science studies where individual participant recruitment was done through Facebook questionnaires, this bias was found to be non-significant [36]. In recent studies that use the audience segmentation data, though, Facebook audience size is often corrected with the penetration rate, i.e. the fraction of actual population who is active on Facebook [37, 38]. We do not use a correction factor in this study since we do not aim at prediction and the metric we use is not the audience size but fraction with a specific interest.

However, to avoid overconfidence in Facebook data as a representative of offline behaviour, we exclude the countries where Facebook penetration is low. Figure 1 illustrates the distribution of total Facebook audience size across 131 countries, namely the daily active users (DAU_j) and the penetration rate (p_j) calculated as in equation (1). For the penetration rate, we take the population aged 15 and more (Pop_j) in equation (1) to correspond to the reported age cohorts of the Facebook audience. We exclude the countries where the penetration rate is below the 25th percentile (0.24) and also the total audience size is below the 25th percentile (1.6 million). With this choice, we leave out the countries where Facebook audience represents less than 24% of the population, yet we keep the ones where the audience size is still considerable (above 1.6 million) even though penetration is low. This choice leads to 16 countries being excluded (j*), and 115 being included (j) in our analysis. Equation (2) denotes the subset of chosen countries, where η symbolizes the percentile function.

$$p_{j'} = \frac{DAU_{j'}}{Pop_{j'}}; \quad j' : 131 \text{ countries} \qquad (1)$$

$$j = \{j' - j^*, \quad j^* : p_{j^*} < \eta_{.25}(p_{j^*}) \wedge DAU_{j^*} \\ < \eta_{.25}(DAU_{j^*})\}. \qquad (2)$$

Thus, the Facebook audience fraction for each interest, e.g. vegetarianism, in each country ($F_{i,j}$) is the DAU with this interest in each country ($DAU_{i,j}$) divided by the total DAU in that country (DAU_j) as denoted in equation (3). Equation (4) shows the audience fraction at higher granularity for each demographic group.

$$F_{i,j} = \frac{DAU_{i,j}}{DAU_j}; \quad i: \text{interests, j: 115 countries} \qquad (3)$$
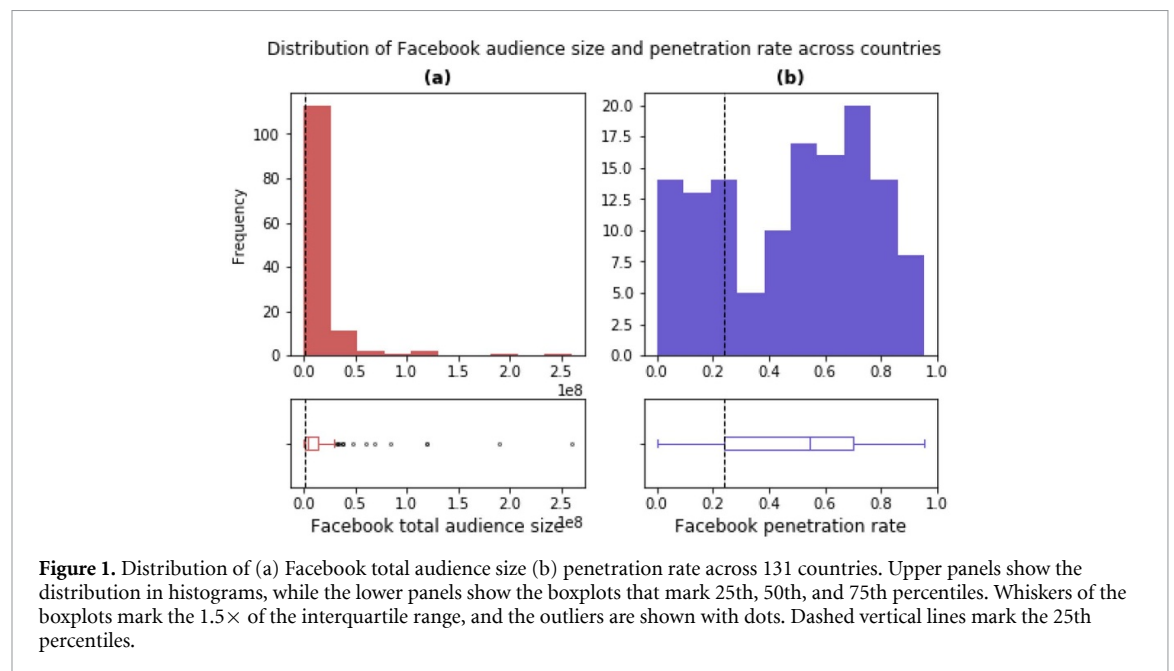
$$F_{i,j,k,l,m} = \frac{DAU_{i,j,k,l,m}}{DAU_{j,kl,m}}; \quad k: \text{age cohorts,} \\ l: \text{gender, } m: \text{scholarities.} \qquad (4)$$

2.3. Surveys

We compare the vegetarian interest of the Facebook audience to available empirical data about vegetarian population in 30 countries (figure 3(a)). We compiled

Table 1. Dimensions and the size of the Facebook audience datasets collected in September 2019, January and June 2020.

	September	January	June
Age	11 Cohorts: 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65+	15–65+	11 Cohorts: 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65+
Gender	Male, female	Both	Male, female
Scholarities	Six groups: in high school, High school grad, In college, College grad, Professional degree, Doctorate degree	—	Six groups: in high school, High school grad, In college, College grad, Professional degree, Doctorate degree
Countries	61 countries which have above average meat consumption per capita according to the UN FAO food supply statistics	132 countries which are common in the Marketing API and the FAO food supply statistics	132 countries which are common in the Marketing API and the FAO food supply statistics
Interests	Vegetarianism (IDs: 6003155333705 6003341096108 6003421091712)	Vegetarianism (IDs: 6003155333705 6003341096108 6003421091712) Sustainable living (IDs: 6003175412902)	Vegetarianism (IDs: 6003155333705 6003341096108 6003421091712) Sustainable living (IDs: 6003175 412902)
N	8052	132 (for each interest)	17292 (for each interest)

Figure 1. Distribution of (a) Facebook total audience size (b) penetration rate across 131 countries. Upper panels show the distribution in histograms, while the lower panels show the boxplots that mark 25th, 50th, and 75th percentiles. Whiskers of the boxplots mark the 1.5× of the interquartile range, and the outliers are shown with dots. Dashed vertical lines mark the 25th percentiles.

this survey data from online news, NGO or market research articles based on queries on Google search engine in English, following the references in the news articles with a snowball approach, and from scientific literature that cites these market research articles. When the original sources could not be reached, we repeated the search engine query in the local language of the corresponding country using online translation. As listed in supplementary table 2, the collection year and sources vary across countries, and often do not have a reliable citation in the online article. We did not leave out such sources but recorded the lack of reliable original sources. Therefore, while using this survey list as the best

available knowledge for comparison purpose, we note that it is not accurate and fully reliable for some countries, due to the differences in data collection time, method, and the lack of citations to actual data sources.

In figure 3(a) we use the Facebook audience data from September 2019—the oldest dataset—since the empirical data is relatively old (supplementary table 2). We report Spearman correlation coefficient as the *r*-value and the two-sided *p*-value for a hypothesis test whose null hypothesis is that the slope is zero, based on a Wald Test with *t*-distribution, calculated using SciPy's linear regression (*stats.linregress*) [39].

2.4. Google Trends

We check the consistency of Facebook data on vegetarianism interest with another online activity indicator, the Google Trends data (figure 3(b)). Google Trends reports the interest in a topic specified by a keyword globally over time, or across all countries at the present time. The interest value aggregates Google search volume, and is reported as an index relative to the categories in the inquiry. We downloaded the Google Trends data in January 2020 for the topics *Vegetarianism, Sustainable diet, Sustainable lifestyles, Sustainable living and Plant-based diet* for 126 countries using the Python package *pytrends* [40]. Global interest in vegetarianism, sustainable diet and plant-based diets have substantially increased in the last 10 years, whereas the interest in sustainable living and sustainable lifestyle have declined (supplementary figure 3). Since Facebook data measures the audience size and Google Trends data measures relative search volumes, in figure 3(b), we compare not these two metrics but the ranking of 126 countries in terms of them.

2.5. Socioeconomic indicators

We analyse the correlation of Facebook data to the country level socioeconomic indicators such as gross domestic product (GDP) per capita, MYS, and average meat consumption per capita. GDP data is obtained from the World Bank statistics [41] for the year 2018, and it is in real USD per capita. MYS data, specifically the MYS *by Broad Age* for the population aged 15 and older, is obtained from the Wittgenstein Centre Data Explorer [42] for the year 2015, the most recent available year. The average meat consumption per capita in 2017 is obtained from the FAO Food Balance Sheets [43], covering the domestic supply of all meat products. It corresponds to the element *Food supply quantity (*kg capita^{-1} yr^{-1}*)* and the aggregated item *Meat (Total)* in the FAOSTAT database. We tag each country in the Facebook dataset with its geographic region as defined in the MESSAGE integrated assessment modelling framework [44]. Table 2 lists the region acronyms used throughout this study, and supplementary figure 14 visualizes the regions.

2.6. Multiple linear regression of country-level indicators

We analyse the dependency between the Facebook audience fraction interested in vegetarianism, meat consumption and other socioeconomic indicators at the country level (figure 4) using a multiple linear regression model denoted in equation (5). We opt for linear regression over nonlinear due to the ease of interpretation of the results. The dependent variable $F_{\text{veg},j}$ is the audience fraction as defined in equation (3), while the independent variables are GDP per capita (GDP$_j$), mean years of schooling (MYS$_j$) and meat consumption per capita (M_j) of each country.

Table 2. Eleven geographic regions.

Acronym	Region	Acronym	Region
AFR	Sub-Saharan Africa	NAM	North America
CPR	Centrally planned Asia and China	PAO	Pacific OECD
EEU	Central and Eastern Europe	PAS	Other Pacific Asia
FSU	Former Soviet Union	SAS	South Asia
LAC	Latin America and the Caribbean	WEU	Western Europe
MEA	Middle East and North Africa		

β_0 denotes the constant term and $\in$ is the error term. We test an alternative model without a constant, and it leads to a slightly worse fit ($R^2 = 0.526$) than the model with constant ($R^2 = 0.54$) as supplementary figure 9 demonstrates. To fit this regression model to the data, we use an ordinary least squares (OLS) model using the Python Package StatsModels [45].

$$F_{\text{veg},j} = \beta_{\text{GDP}}GDP_j + \beta_{\text{MYS}}MYS_j$$
$$+ \beta_m M_j + \beta_0 + \in; \; j: \text{countries.} \quad (5)$$

2.7. Feature importance based on regression tree models

We employ ML models to identify the relative importance of demographic factors (age, gender, education, location) included in the granular Facebook audience segmentation data. This choice of ML models is motivated by the limitations of statistical models to address the collinearity between the input factors such as income and education. We use a regression tree model that can address this drawback of statistical models not for prediction purposes but to quantitatively link the input features (demographic and socioeconomic factors) to the output (audience fraction) and to use the interpretation functionality of this model. Below, we describe the XGBoost learning algorithm that we used to build a regression tree model, and Shapley additive explanation values that we used for calculating feature importance on this model.

2.7.1. XGBoost learning algorithm

XGBoost is an ensemble learning method based on gradient-boosted decision trees [46], meaning that the tree ensemble is formed by additive training where each new tree is fit to the data considering what has been learned in the previous steps, as opposed to random forests where each tree is fit by random bagging of the training data. XGBoost is shown to provide a robust performance, accuracy and computational efficiency on classification and regression tasks compared to linear regression and deep learning methods [46, 47]. XGBoost has been widely used

in scientific applications from disease diagnosis in healthcare [48, 49] to environmental pollution prediction [50, 51].

We choose to use tree-based methods in general and XGBoost in particular due to two reasons: first, the relationship between demographic factors and the audience fraction is nonlinear as supplementary figure 8 shows, and there are potential collinearities between these factors. For instance, education level is dependent on age to some extent since higher educational attainment takes time, or the country and education level may be dependent due to the GDP and MYS relationship shown in supplementary figures 6 and 7. Tree-based methods better address these non- and collinearities than linear regression. Second, we use these ML methods to identify the factor importance, not for prediction. Therefore, explainability of tree-methods provides an advantage for our purpose. XGBoost, in particular, is chosen for its superior performance over other tree-based methods. Supplementary figure 10 illustrates a comparison of linear regression, random forest and XGBoost on our dataset, with XGBoost leading to the lowest mean squared error (MSE), hence better accuracy.

For implementation, we use the Python implementation of XGBoost after splitting the data into a 75/25 training/test set with random shuffling. Although overfitting should be avoided in training tree-models since it can cause conservative predictions, we aimed at a low MSE between the test data and model predictions since we use the model for now-casting, i.e. explaining the present data. We iterated over different parameter values of the XGBoost algorithm, such as the learning rate, maximum tree depth, objective function and tree method. Supplementary figure 11 shows the model fit for the first two lowest MSE options. We obtained the lowest MSE with a learning rate of 0.5, a maximum depth of 9, an objective function based on logistic regression and the tree construction method is set to 'auto', which uses a heuristic to choose the fastest tree construction method from the available options. In feature importance analysis, we use this model with the lowest MSE resulting from the abovementioned specifications.

2.7.2. Shapley additive explanation (SHAP) values
To estimate the importance of demographic factors for the audience fraction interested in low-carbon lifestyles, we used Shapley additive explanation values [47]. Shapley values originate from game theory, where they are used to calculate the individual contributions to the cooperative payoff in an n-player game regardless of the order of coalition formation [52]. They are adopted in ML since they meet all desirable properties of an explanation model, i.e. a model that is used to explain the behaviour of a prediction model based on the individual contribution of input factors (features) [53]. Compared to other

metrics used in ML, Shapley additive explanation values provide more robust conclusions for feature importance [47], since they can better account for high order interactions, correlations and categorical features with highly imbalanced classes, as we have in our dataset.
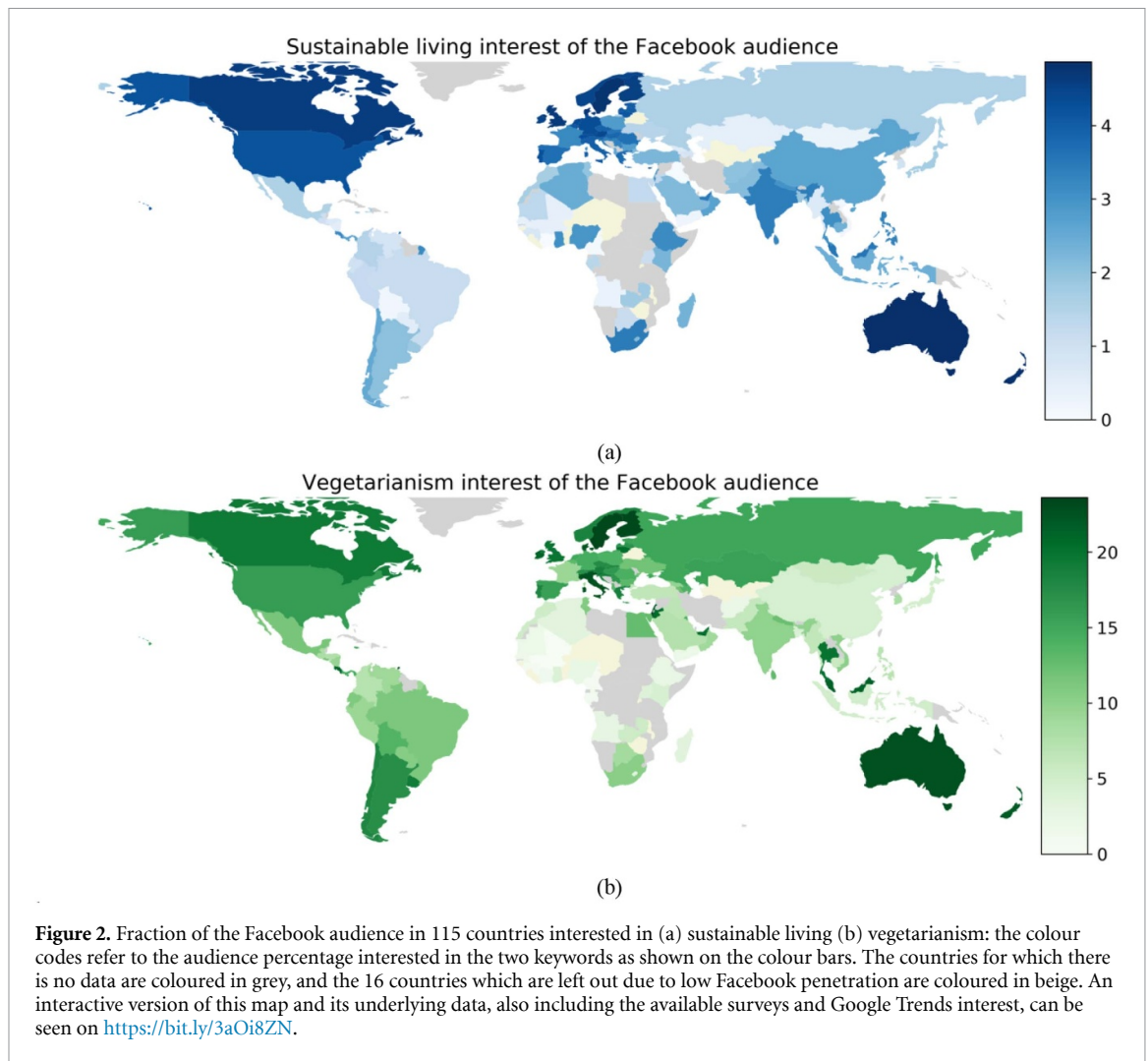
To calculate the Shapley values on the tree model we generated with XGBoost and visualize the results, we use the tree explainer feature of Python package SHAP [47, 53] and its supporting visualizations.

3. Results

We consider the fraction of Facebook audience interested in low-carbon lifestyles as a proxy for the spread of this phenomenon in each country where Facebook penetration is relatively high, and as a cross-country comparison indicator. Figure 2 visualizes the relative spread of *sustainable living* interest and *vegetarianism* interest of the Facebook audience in 115 countries in January 2020. Australia, New Zealand, Sweden and Denmark have the highest ranks for the *sustainable living* interest (4.9%–3.5%) with a mean value of 0.7% across all countries. Vegetarianism interest is most common among the Facebook audience in Singapore, Sweden, Finland and Israel (∼18%), with a mean value of 7.5% across all countries. Despite the countries that show relatively high interest in both sustainable living and vegetarianism, such as Scandinavian countries, the two phenomena do not strongly correlate (supplementary figure 2). In other words, the country-wide interest in sustainable living does not lead to a country-wide interest in vegetarianism as an indicator of sustainable diets, or vice versa.

3.1. Data consistency
We investigate the consistency of Facebook data with other offline and online sources through a series of comparisons shown in figure 3. There is no large scale data available about sustainable living interest to our knowledge. Therefore, we perform these comparisons only for the interest in vegetarianism. Figure 3(a) illustrates the Facebook audience fraction interested in *vegetarianism* in $N = 30$ countries with respect to the survey results about vegetarian population fraction in those countries. The statistical measures do not indicate a strong consistency between the two datasets, with a small and negative correlation ($r = -0.18$), a high p-value (0.335) for a null hypothesis that the slope of linear regression is zero, and a high mean absolute percentage error (181%). It must however be reminded that the available survey data is on average 4 years older than the Facebook data, based on limited sample sizes and different collection methods (see supplementary table 2). The deviation of Facebook data from the empirical data is smaller for a few countries with recent surveys. For instance, according to the

Figure 2. Fraction of the Facebook audience in 115 countries interested in (a) sustainable living (b) vegetarianism: the colour codes refer to the audience percentage interested in the two keywords as shown on the colour bars. The countries for which there is no data are coloured in grey, and the 16 countries which are left out due to low Facebook penetration are coloured in beige. An interactive version of this map and its underlying data, also including the available surveys and Google Trends interest, can be seen on https://bit.ly/3aOi8ZN.
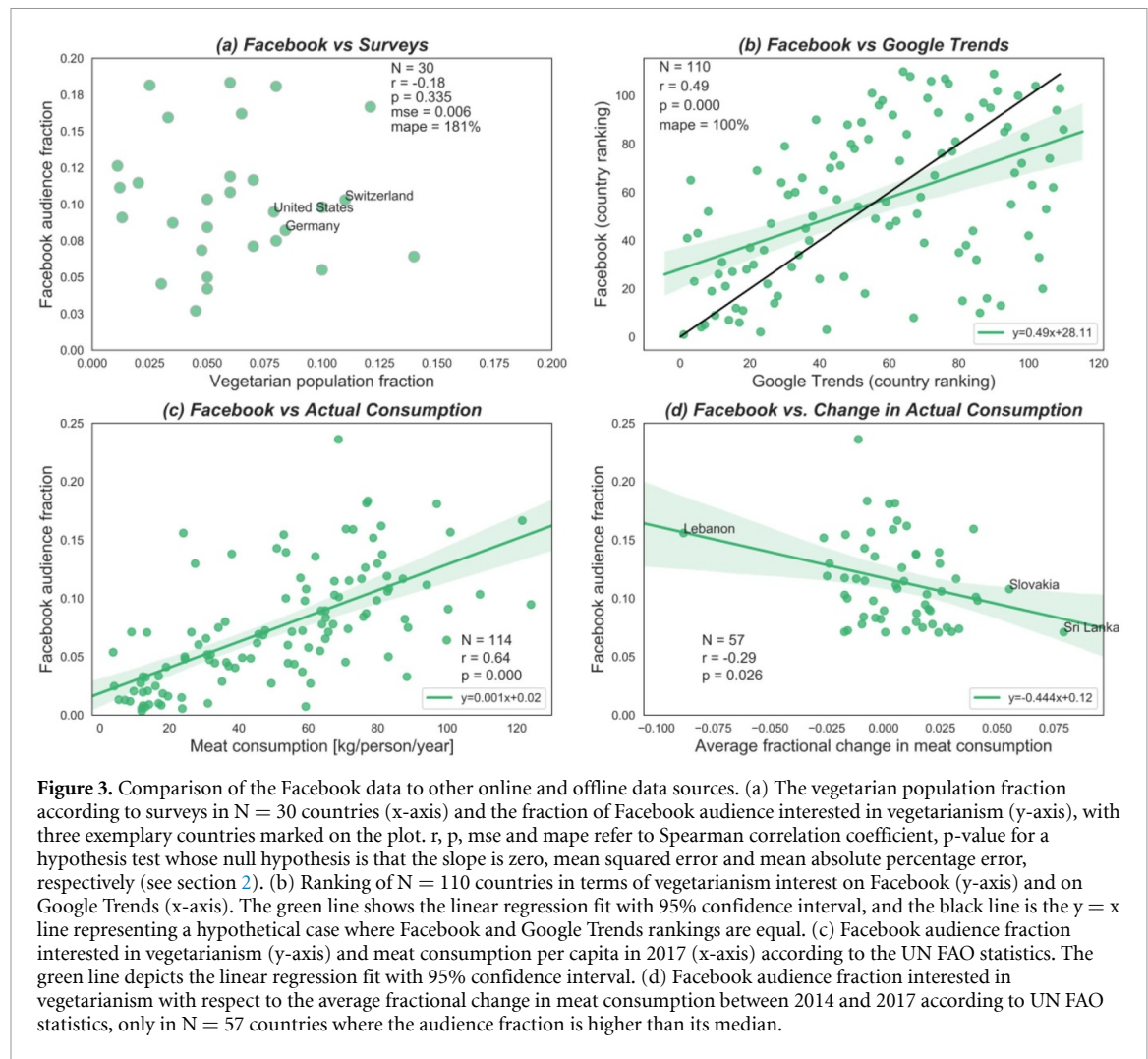
data downloaded in January 2020, the audience fraction interested in vegetarianism is 8%, 10% and 9% in Germany, Switzerland and the US, respectively. Empirical surveys reported the fraction of vegetarians as 7.6% in 2016 in Germany [54, 55], 11% in 2017 in Switzerland [56], and 7.9% in 2016 in the US [57].

Figure 3(b) compares the ranking of 110 countries that are common in both Facebook and Google Trends country lists with respect to vegetarianism interest in these two online platforms. Despite discordances between the two data sources, the results show a relatively strong positive correlation ($r = 0.49$), statistically significant linear relationship ($p < 0.001$) and a smaller mean percentage error (mape = 100%). Therefore, Facebook audience data is more coherent with Google Trends, another indicator of online activity, than it is with offline empirical data.

Figure 3(c) compares the Facebook interest in vegetarianism to meat consumption in 114 countries. The results show a strong positive correlation ($r = 0.64$) and a statistically significant positive linear

relationship ($p < 0.001$). This finding is counterintuitive, because if the Facebook interest in vegetarianism is an indicator of actual interest in vegetarianism, one could expect meat consumption to be low in the countries with high vegetarianism interest. However, meat consumption is stated to be highly dependent on income both at an individual and national level [58, 59], while vegetarianism interest is linked to high income and education at an individual level, too [57, 60, 61]. Therefore, the positive relationship in figure 3(c) is due to common underlying factors as we discuss in more detail below. Still, in the countries where the Facebook vegetarianism interest is high, we observe a negative relationship between the vegetarianism interest and the trend of meat consumption between 2014 and 2017 (figure 3(d)). In other words, in countries with high vegetarianism interest, meat consumption per capita has declined between 2014 and 2017. This negative relationship visualized in figure 3(d) is present even if Lebanon, the outlier country with the lowest average fractional change in meat consumption, is removed (see supplementary figure 4 for the correlation statistics when

Figure 3. Comparison of the Facebook data to other online and offline data sources. (a) The vegetarian population fraction according to surveys in N = 30 countries (x-axis) and the fraction of Facebook audience interested in vegetarianism (y-axis), with three exemplary countries marked on the plot. r, p, mse and mape refer to Spearman correlation coefficient, p-value for a hypothesis test whose null hypothesis is that the slope is zero, mean squared error and mean absolute percentage error, respectively (see section 2). (b) Ranking of N = 110 countries in terms of vegetarianism interest on Facebook (y-axis) and on Google Trends (x-axis). The green line shows the linear regression fit with 95% confidence interval, and the black line is the y = x line representing a hypothetical case where Facebook and Google Trends rankings are equal. (c) Facebook audience fraction interested in vegetarianism (y-axis) and meat consumption per capita in 2017 (x-axis) according to the UN FAO statistics. The green line depicts the linear regression fit with 95% confidence interval. (d) Facebook audience fraction interested in vegetarianism with respect to the average fractional change in meat consumption between 2014 and 2017 according to UN FAO statistics, only in N = 57 countries where the audience fraction is higher than its median.
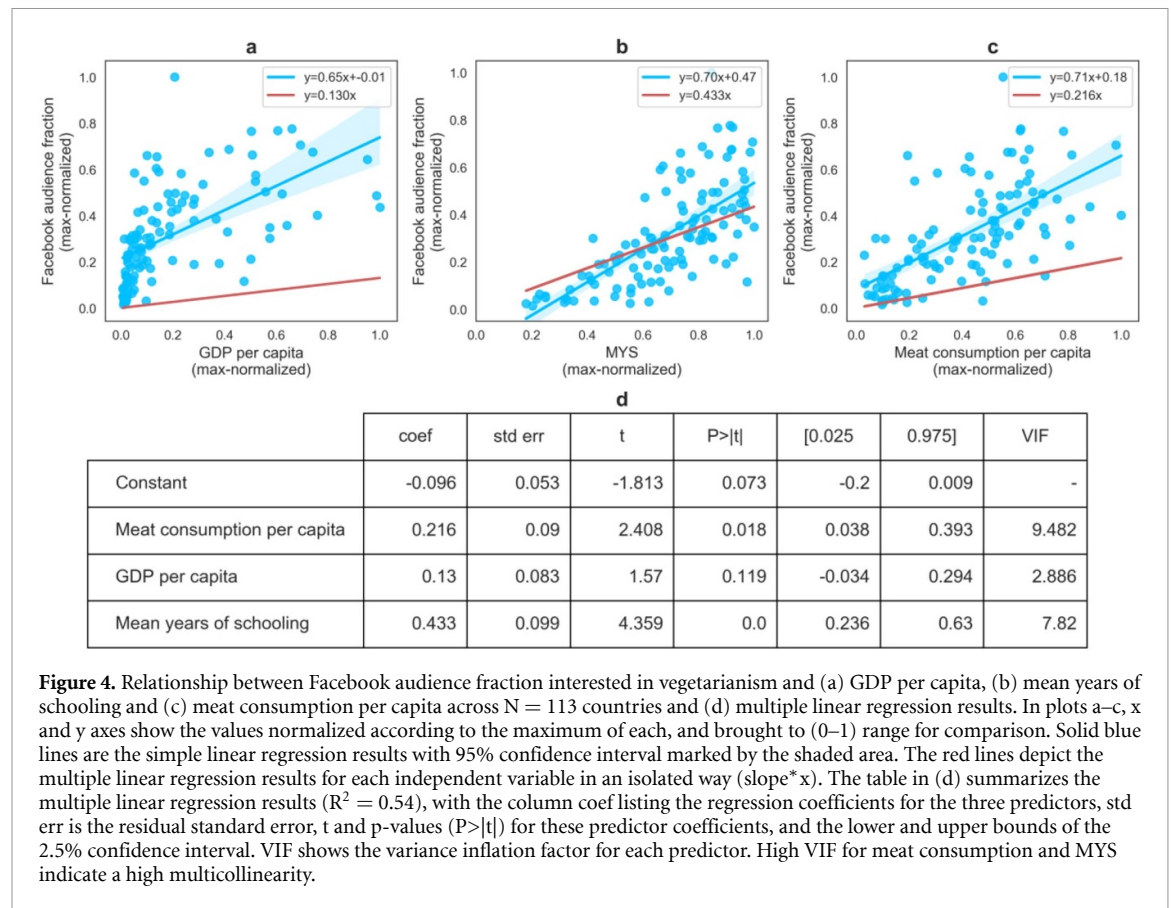
Lebanon and other outliers are removed, and supplementary figure 5 for the relationship between Facebook data and meat consumption trends in all countries). Therefore, even though the Facebook audience data do not fully align with the empirical surveys and actual consumption for the reasons we have discussed, it captures the consumption trends especially in countries where meat consumption has declined and Facebook vegetarianism interest has been high.

We test the relation between Facebook vegetarianism interest and its potential predictors at the country level—meat consumption, GDP per capita and education (MYS) —in a multiple linear regression model (see section 2). According to the results in figure 4, the three factors explain 54% of the variation in the Facebook vegetarianism interest ($R^2 = 0.54$). The relationship between the Facebook audience fraction and the three factors is significant ($p < 0.05$ for each except GDP per capita) and positive. Education (MYS) appears as a more important predictor than income (GDP per capita) and meat consumption. However, high correlation between MYS, GDP and meat consumption (supplementary figures 6 and 7) and a high variance inflation factor (VIF) for meat

consumption (figure 4(d)) indicate multicollinearity in this dataset, for instance the effect of education on income, hence meat consumption. Therefore, we conclude that the positive correlation between Facebook vegetarianism interest and meat consumption reveals the mutual underlying factors. In order to better understand the relationship between vegetarianism interest and the socioeconomic and demographic factors, and to derive a robust ranking of those factors despite their multicollinearity, we analyse the granular Facebook data for each audience group using ML techniques.

3.2. Importance of demographic factors for low-carbon lifestyle interest

The Facebook dataset includes audience groups defined by four factors: gender, age cohort, education level and country. No information on income is available on Facebook audience data, therefore we cannot include the income factor at the same granularity level. Still, due to the strong correlation of GDP per capita to meat consumption and vegetarianism interest at the country level (supplementary figure 7), we add GDP per capita as an additional factor to

Figure 4. Relationship between Facebook audience fraction interested in vegetarianism and (a) GDP per capita, (b) mean years of schooling and (c) meat consumption per capita across N = 113 countries and (d) multiple linear regression results. In plots a–c, x and y axes show the values normalized according to the maximum of each, and brought to (0–1) range for comparison. Solid blue lines are the simple linear regression results with 95% confidence interval marked by the shaded area. The red lines depict the multiple linear regression results for each independent variable in an isolated way (slope*x). The table in (d) summarizes the multiple linear regression results ($R^2 = 0.54$), with the column coef listing the regression coefficients for the three predictors, std err is the residual standard error, t and p-values (P>|t|) for these predictor coefficients, and the lower and upper bounds of the 2.5% confidence interval. VIF shows the variance inflation factor for each predictor. High VIF for meat consumption and MYS indicate a high multicollinearity.

account for the country level income. We also add the geographic *region* of each country, assuming cultural similarity among the countries in each region. We then identify the relative impact of each factor on audience fraction interested in vegetarianism by building a regression tree model on the dataset of N = 12884 audience groups, and computing shapley additive explanation values on this model (see section 2).

Education is the most important driver of vegetarianism interest in the Facebook audience followed by gender, GDP per capita, age and region (figure 5(a)). This finding resonates with the empirical studies which found that vegetarians tend to be more highly educated than meat-eaters [60–63]. The relationship between the impact of education and vegetarianism interest is nonlinear, though, high education levels leading to either a highly positive or highly negative impact on the vegetarianism interest. This is demonstrated by the bimodal distribution of individual importance metrics of each data point (individual SHAP values), with high education values on the two ends, in the top row of figure 5(b). Complementing this distribution of education impact, figure 6(a) shows the impact of each education level depending on the GDP per capita of the country. These figures highlight a dual effect of education. From high school to university graduates, the impact of education on the Facebook vegetarianism interest

is increasing, but it is much lower among the professional and doctorate degree holders (supplementary figure 8(b)). There are two possible explanations for this. First, this can be attributed to the increase in income as the education level increases, and income correlates with high meat consumption as it has long been known [58, 59]. Second, it can be due to the weak representativeness of these groups on Facebook audience data, since the doctorate graduates constitute a very low fraction of the population (1.1% in OECD countries on average [64]). Therefore, the vegetarianism interest in these small educational attainment groups should be further investigated.

Gender has a very distinctive impact on vegetarianism interest, with females leaning towards a higher interest in vegetarianism (figure 5(b), 2nd row). This finding is also supported by the available empirical studies [60, 62, 65]. Therefore, the Facebook audience data complements and supports the local empirical findings by covering a much larger population. The impact of age as a driver of vegetarianism interest is slightly lower than gender, and the distinction between young and old is not as clear. Empirical studies [57, 60] state that the youth have a wider interest in vegetarianism. In figure 5(b) (4th row), red points representing older age cohorts tend to accumulate around negative Shapely values hence lower vegetarianism interest, whereas the positive
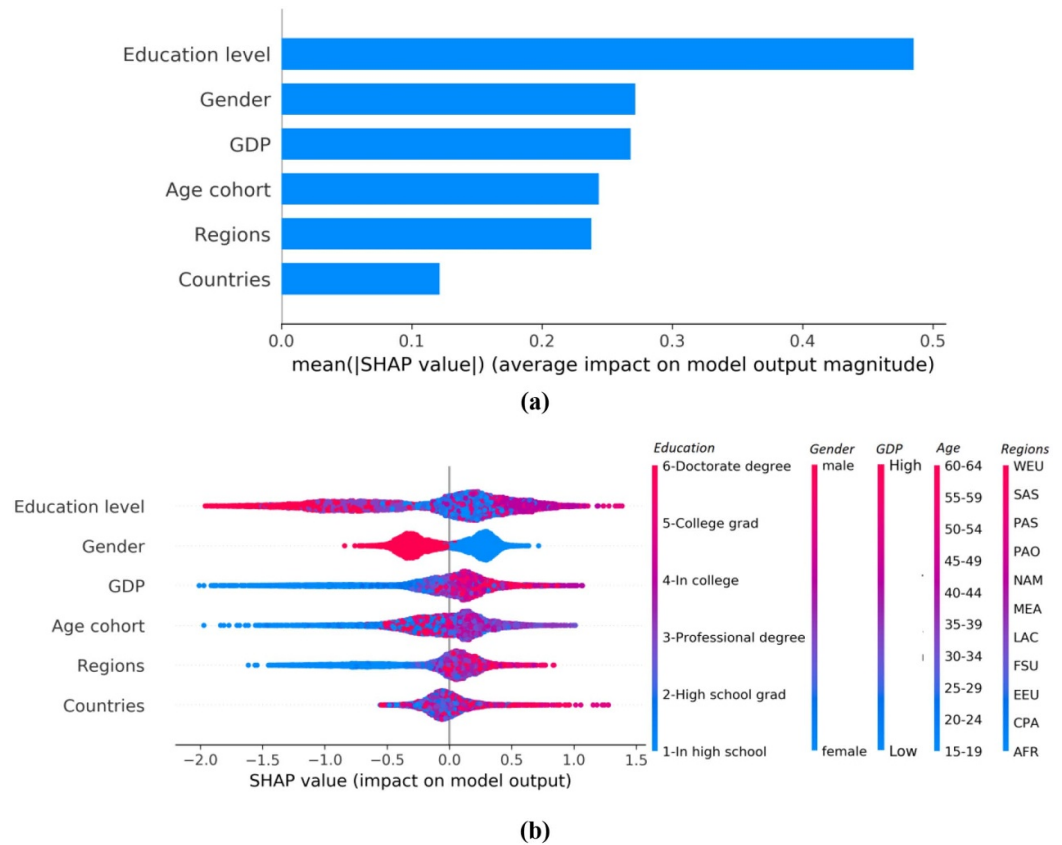
(a)

(b)

Figure 5. Relative importance of demographic factors for explaining the Facebook audience fraction interested in vegetarianism. (a) Mean absolute shapley additive explanation (SHAP) values estimated based on a regression tree model fit to the Facebook audience size data. (b) Individual SHAP values showing the distribution of the impacts of each demographic factor (feature) on the model output. The higher the SHAP value of a factor, the higher the vegetarianism interest. In (b), each dot refers to a data point, which is a demographic group defined by age cohort, gender, country and education level. The data points are stacked vertically to show the density, and coloured according to the feature value. For education level and age cohort, red points refer to higher education levels and older ages, respectively, whereas the blue is for lower education and younger age cohorts. For gender, red refers to males and blue refers to females. For countries, the countries that are last in the alphabetic order are marked with red. Similarly red-blue colour scale for GDP refers to the high-low spectrum. The 11 regions are shown on an additional colour bar on panel (b). See section 2 for the definition of regions. The figure is created using the Python package shap [47].
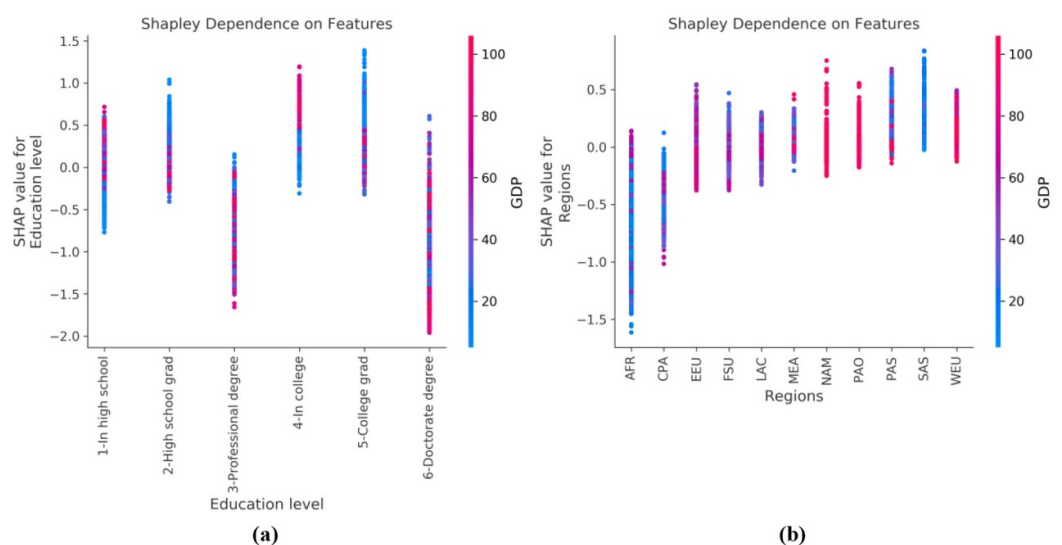

(a)   (b)

Figure 6. Shapley dependence plots (a) Impact of education level on Facebook vegetarianism interest depending on country's GDP per capita. The figure shows the individual SHAP values for each education level, coloured according to GDP per capita. (b) Impact of geographic region on Facebook vegetarianism interest depending on countries' GDP per capita. Individual SHAP values for each region coloured according to GDP per capita. These two plots can be considered as the first and 5th rows of figure 5(b) rotated and divided into their distinct values. The figure is created using the Python package shap [47].

Shapley values coincide with younger age cohorts. However, the youngest age cohort makes the most negative impact on vegetarianism interest, implying that vegetarianism interest is not high among the very young Facebook audience. Supplementary figure 8(a) supports this finding, as the age cohort 15–19 has the minimum average audience fraction interested in vegetarianism.

GDP per capita has a positive effect on vegetarianism interest on average, with high GDP values leading to a positive impact and low values leading to a negative impact (figures 5(a) and (b)). In other words, *GDP per capita* and the Facebook vegetarianism interest are parallel as demonstrated before with a few exceptions caused by regional differences, such as SAS countries having a low GDP but high vegetarianism interest. *GDP per capita* also interacts with the effect of *education level*, as illustrated in figure 6(a). Being a college or high school graduate in low-income countries has a higher impact on vegetarianism interest than it has in high-income countries, and than other education levels. In the doctorate and professional degree groups, living in a high-income country results in a negative impact on vegetarianism interest (see the lowest part of the doctorate degree column in figure 6(a)). These findings indicate that general assumptions, such as a steady positive relationship between vegetarianism and GDP per capita or education, do not globally hold. Heterogeneity across countries should be taken into account not only for individual effects, i.e. education on vegetarianism, but also for the interaction of effects.

Geographic regions also play a distinct role in the Facebook audience's vegetarianism interest. Western Europe (WEU), South Asia (SAS), Pacific OECD Countries (PAO) and other Pacific Asia (PAS) have a high vegetarianism interest, whereas Sub-Saharan Africa, Centrally Planned Asia and Eastern Europe are associated with low vegetarianism interest. While this can be explained by culture to some extent, i.e. low meat consumption in India in South Asia due to religious reasons, the similarity of *GDP per capita* within regions plays an important role in this distinction between high and low interest (figure 6(b)). For instance, in high-GDP regions (Western Europe, North America and Pacific OECD) the vegetarianism interest is high. Being in Latin America and Middle East makes a similar positive impact on the vegetarianism interest as in Pacific OECD countries, despite their lower GDP per capita values.

4. Discussion

This study showed that the audience data of OSM platforms can be a useful source to analyse the drivers of low-carbon lifestyles at the global level by taking the societal heterogeneity into account. Furthermore,

specifically in the case of meat consumption, it highlighted the complex interplay between income, education, meat consumption and the interest in plant-based diets, since the GDP per capita underlies both meat consumption and vegetarianism interest. Our findings showed that the fraction of Facebook audience interested in vegetarianism in a country positively correlates with the average meat consumption per capita, implying that a wider interest in vegetarianism in a country does not lead to a lower meat consumption. However, in the countries where Facebook audience's interest is high, there is a declining trend in meat consumption. In other words, Facebook data does not indicate a negative relationship between vegetarianism interest and meat consumption on a global scale, but it captures the trend of increasing vegetarianism interest and declining consumption.

The second main finding of this study is that education is the most important driver of vegetarianism interest of the Facebook audience among basic demographic factors such as age, gender, education level, country-level GDP per capita and geographic regions. High school and college graduates have a higher interest in vegetarian diets than others and education plays a distinctive role especially in low-income regions. Vegetarianism interest among the doctorate graduates on Facebook is low, indicating a non-monotonic positive relationship between education and vegetarianism interest. However, since the doctorate graduates constitute a very low fraction of the population and Facebook audience, the spread of vegetarianism and the representativeness of Facebook audience at this education level should be further investigated for a definite conclusion.

This study also showed that gender is a strongly distinguishing factor for vegetarianism interest on a global level, females having a significantly higher interest. The young and middle-aged (20–49) has a wider tendency for vegetarianism interest, yet the difference between age cohorts is not sharp, and the youngest cohort of the Facebook audience included in this study (15–19) has the lowest fraction interested in vegetarianism. Our findings at the global level about the effect of education, gender, age and income on plant-based dietary choices resonate with empirical findings from USA [57], Germany [60], Belgium [61]. Therefore, this analysis of Facebook market segmentation data complements empirical studies by extending their findings to a global scale with larger samples, and also highlights peculiar issues for instance regarding the youngest age cohort or highest education level.

GDP per capita is found to be one of the key factors that make a positive impact on the vegetarianism interest. However, while it enables distinguishing the countries with low and high income, it is not a precise indicator of personal income, hence both this study and Facebook audience data are limited

in investigating the effect of personal income on low-carbon lifestyle choices. This limitation can be addressed in future studies that focus on the income effect for instance by matching the social media audience data with country level personal income statistics based on demographic factors (e.g. education) and location, or by using proxies within the audience segmentation data such as interest in luxury as done in market research.

Another limitation that should be tackled while using Facebook audience segmentation data to analyse the drivers of low-carbon lifestyles is the availability of relevant interest categories and keywords. The interest group *vegetarianism* we used in this study is a relatively direct indicator of plant-based choices, whether the motivation behind this choice is animal welfare, religion or pro-environmental preferences. In other consumption areas such as heating and cooling demand, though, Facebook audience segmentation may not be as categorical as vegetarianism to represent the consumer preferences. Therefore, similar future analyses should be based on a representative relationship between the available interest categories and consumption areas.

5. Conclusion

Reduction of food and energy demand is often quoted as a highly promising climate change mitigation option. This requires widespread behavioural changes across the global population. Existing mitigation assessment frameworks, such as those used by IPCC, are limited in feasibility consideration since they lack such behavioural aspects of consumer response [66]. However, it is of crucial importance to include behavioural considerations in mitigation scenarios by bridging across disciplines in order to guide decision-making for a sustainable and healthy future [67, 68]. There are a few initial studies that incorporate behavioural factors into model-based integrated assessments of feasible mitigation potential [69–71], yet such quantitative analyses are bounded by data availability on a global scale.

This study addressed this data gap by investigating the drivers of low-carbon diets on a global scale based on the Facebook audience segmentation data. The conceptual agreement between the conventional empirical data and Facebook audience data shown in this study underlines the potential of combining these two sources for quantifying the trajectories of lifestyle change. In particular, while empirical studies and surveys shed light on the nuances of heterogeneity and provide a deeper understanding of low-carbon lifestyle choices, digital data, e.g. social media data, can extend the geographic, temporal and contextual scope of analysis and broaden the evidence.

The main policy implication of our findings is that education should be at the centre of policy design

for stimulating low-carbon diets. Other main demographic factors such as gender and age are also distinguishing, with females and younger people having a stronger interest in plant-based diets. Therefore, social heterogeneity in terms of these key factors should guide the assessment of any policy lever that aims to incentivize low-carbon diets. Education can also be a powerful lever itself, especially to counteract on the adverse effect of income. It is widely accepted that affluence has increased environmental degradation more than technological progress can prevent it, therefore the affluent citizens are central to reversing environmental degradation [72]. Although intervention studies report that targeted short education, such as those on multiple adverse consequences of eating meat, does not necessarily lead to behaviour change [73], our results show that formal education level is a strong determinant of interest in plant-based diets. Therefore, if the economic growth is to be continued, to make it 'green', school curricula can be instrumental to raising awareness of responsible consumption and sustainable choices among erudite and affluent citizens.

The main implication of our findings for further research is that societal heterogeneity should be at the forefront of quantitative scenario studies that evaluate the demand-focused mitigation and sustainability policies. Given that social-demographic factors such as education, gender and age are highly important in lifestyle change, hence demand, their future projections should guide the development of demand scenarios. Large scale audience data of social media platforms consistent across a large number of countries and large population groups can assist scenario development by quantifying the demand depending on social heterogeneity. It can provide insights about temporal trends of low-carbon lifestyle interest if the data is tracked over time. Therefore, it can help coupling of behavioural models of societal dynamics and integrated assessment models of environment and economy to ensure plausibility and feasibility of demand-focused mitigation scenarios.

Still, demographic and socioeconomic heterogeneity explored in this study through Facebook audience data is not sufficient to capture the psychosocial drivers of lifestyle change. In addition to the data on audience size used in this study, text analysis such as topic modelling and sentiment analysis on user-generated content [74, 75] can be useful to analyse the psychosocial drivers of lifestyle change. Social and personal norms, for instance, is a highly cited driver of dietary shifts and lifestyle change [69, 76]. Social media data can be useful especially to quantify and simulate the social norm effect in lifestyle change scenarios. This requires scientists to access anonymous data about social connections and diffusion that are currently not public. Therefore, the need for more and better data to analyse low-carbon lifestyles recalls the growing demand of scientists from technology

companies to publicize the user data for common interest [77, 78].

Acknowledgements

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/sibeleker/Facebook_Lifestyle.

References

[1] IPCC 2018 Special report on global warming of 1.5 °C: intergovernmental panel on climate change

[2] IPCC 2019 Special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems: intergovernmental panel on climate change

[3] Grubler A *et al* 2018 A low energy demand scenario for meeting the 1.5 °C target and sustainable development goals without negative emission technologies *Nat. Energy* **3** 515–27

[4] Tilman D and Clark M 2014 Global diets link environmental sustainability and human health *Nature* **515** 518

[5] Springmann M *et al* 2018 Options for keeping the food system within environmental limits *Nature* **562** 519–25

[6] Van Vuuren D P *et al* 2018 Alternative pathways to the 1.5 °C target reduce the need for negative emission technologies *Nat. Clim. Change* **8** 391–7

[7] Moallemi E A *et al* 2020 Global pathways to sustainable development to 2030 and beyond (arXiv:2012.04333)

[8] Obersteiner M *et al* 2016 Assessing the land resource–food price nexus of the sustainable development goals *Sci. Adv.* **2** e1501499

[9] Stern P C 2000 New environmental theories: toward a coherent theory of environmentally significant behavior *J. Soc. Issues* **56** 407–24

[10] Steg L and Vlek C 2009 Encouraging pro-environmental behaviour: an integrative review and research agenda *J. Environ. Psychol.* **29** 309–17

[11] Clayton S, Devine-Wright P, Stern P C, Whitmarsh L, Carrico A, Steg L, Swim J and Bonnes M 2015 Psychological research and global climate change *Nat. Clim. Change* **5** 640

[12] Wynes S, Nicholas K A, Zhao J and Donner S D 2018 Measuring what works: quantifying greenhouse gas emission reductions of behavioural interventions to reduce driving, meat consumption, and household energy use *Environ. Res. Lett.* **13** 113002

[13] Van Valkengoed A M and Steg L 2019 Meta-analyses of factors motivating climate change adaptation behaviour *Nat. Clim. Change* **9** 158–63

[14] Abrahamse W and Steg L 2013 Social influence approaches to encourage resource conservation: a meta-analysis *Glob. Environ. Change* **23** 1773–85

[15] Creutzig F *et al* 2018 Towards demand-side solutions for mitigating climate change *Nat. Clim. Change* **8** 260–3

[16] Creutzig F, Fernandez B, Haberl H, Khosla R, Mulugetta Y and Seto K C 2016 Beyond technology: demand-side solutions for climate change mitigation *Annu. Rev. Environ. Resour.* **41** 173–98

[17] Trutnevyte E, Hirt L F, Bauer N, Cherp A, Hawkes A, Edelenbosch O Y, Pedde S and Van Vuuren D P 2019 Societal transformations in models for energy and climate policy: the ambitious next step *One Earth* **1** 423–33

[18] Paulhus D L and Vazire S 2007 The self-report method *Handbook of Research Methods in Personality Psychology* ed R W Robins, R C Farley and R F Krueger (London: Guilford Press) pp 228–33

[19] Chan D 2009 So why ask me? Are self-report data really that bad? *Statistical and Methodological Myths and Urban Legends* ed C E Lance and R J Vandenberg (New York: Routledge) pp 309–36

[20] Rolnick D *et al* 2019 Tackling climate change with machine learning (arXiv:1906.05433)

[21] Lazer D, Kennedy R, King G and Vespignani A 2014 The parable of Google flu: traps in big data analysis *Science* **343** 1203–5

[22] Butler D 2013 When Google got flu wrong *Nature* **494** 155–6

[23] Chunara R, Bouton L, Ayers J W and Brownstein J S 2013 Assessing the online social environment for surveillance of obesity prevalence *PLoS One* **8** e61373

[24] Abbar S, Mejova Y and Weber I 2015 You tweet what you eat: studying food consumption through twitter *Proc. 33rd Annual ACM Conf. on Human Factors in Computing Systems 2015* (ACM) pp 3197–206

[25] Yildiz D, Munson J, Vitali A, Tinati R and Holland J A 2017 Using Twitter data for demographic research *Demogr. Res.* **37** 1477–514

[26] Zagheni E, Weber I and Gummadi K 2017 Leveraging facebook's advertising platform to monitor stocks of migrants *Popul. Dev. Rev.* **43** 721–34

[27] Palotti J, Adler N, Morales-Guzman A, Villaveces J, Sekara V, Garcia Herranz M, Al-Asad M and Weber I 2020 Monitoring of the Venezuelan exodus through Facebook's advertising platform *PLoS One* **15** e0229175

[28] Kryvasheyeu Y, Chen H, Obradovich N, Moro E, Van Hentenryck P, Fowler J and Cebrian M 2016 Rapid assessment of disaster damage using social media activity *Sci. Adv.* **2** e1500779

[29] Fatehkia M, Kashyap R and Weber I 2018 Using Facebook ad data to track the global digital gender gap *World Dev.* **107** 189–209

[30] Garcia D, Kassa Y M, Cuevas A, Cebrian M, Moro E, Rahwan I and Cuevas R 2018 Analyzing gender inequality through large-scale Facebook advertising data *Proc. Natl Acad. Sci.* **115** 6958–63

[31] Liao Y, Yeh S and Gil J 2021 Feasibility of estimating travel demand using geolocations of social media data *Transportation* (https://doi.org/10.1007/s11116-021-10171-x)

[32] Liao Y, Yeh S and Jeuken G S 2019 From individual to collective behaviours: exploring population heterogeneity of human mobility based on social media data *EPJ Data Sci.* **8** 34

[33] Araujo M, Mejova Y, Weber I and Benevenuto F 2017 Using Facebook ads audiences for global lifestyle disease surveillance: promises and limitations *Proc. 2017 ACM on Web Science Conf. 2017* pp 253–7

[34] Facebook 2020 Marketing API

[35] Facebook 2020 About detailed targeting (available at: www.facebook.com/business/help/182371508761821?id=176276233019487)

[36] Rife S C, Cate K L, Kosinski M and Stillwell D 2016 Participant recruitment and data collection through Facebook: the role of personality factors *Int. J. Soc. Res. Methodol.* **19** 69–83

[37] Ribeiro F N, Benevenuto F and Zagheni E 2020 How biased is the population of Facebook users? comparing the demographics of Facebook users with census data to generate correction factors (arXiv: 2005.08065)

[38] Kashyap R, Fatehkia M, Al Tamime R and Weber I 2020 Monitoring global digital gender inequality using the online

populations of Facebook and Google *Demogr. Res.* **43** 779–816

[39] Virtanen P *et al* 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python *Nat. Methods* **17** 261–72

[40] General Mills 2020 pytrends: pseudo API for google trends (https://pypi.org/project/pytrends/)

[41] World Bank 2019 GDP per capita (current US$) *World Bank national accounts data 2020* (available at: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD) (cited 20 February 2020)

[42] Wittgenstein Centre for Demography and Global Human Capital 2018 Wittgenstein centre data explorer version 2.0 (beta) [cited 20/ 02/2020] (available at: www.wittgensteincentre.org/dataexplorer)

[43] United Nations Food and Agriculture Organization 2019 New food balances. FAOSTAT statistical database (available at: www.fao.org/faostat/en/#data/FBS) (cited 20 February 2020)

[44] Krey V *et al* 2016 Message-globiom 1.0 documentation (Laxenburg: International Institute for Applied Systems Analysis)

[45] Seabold S and Perktold J 2010 Statsmodels: econometric and statistical modeling with python *Proc. 9th Python in Science Conf. 2010* (*Austin, TX*) p 61

[46] Chen T and Guestrin C 2016 Xgboost: a scalable tree boosting system *Proc. 22nd Acm Sigkdd Int. Conf. on Knowledge Discovery and Data Mining 2016* pp 785–94

[47] Lundberg S M, Erion G, Chen H, DeGrave A, Prutkin J M, Nair B, Katz R, Himmelfarb J, Bansal N and Lee S-I 2020 From local explanations to global understanding with explainable AI for trees *Nat. Mach. Intell.* **2** 2522–5839

[48] Gao C *et al* 2018 Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease *Sci. Rep.* **8** 1–21

[49] Lin H, Zou W, Li T, Feigenberg S J, Teo B K K and Dong L 2019 A super-learner model for tumor motion prediction and management in radiation therapy: development and feasibility evaluation *Sci. Rep.* **9** 1–11

[50] Chen J, Yin J, Zang L, Zhang T and Zhao M 2019 Stacking machine learning model for estimating hourly PM2. 5 in China based on Himawari 8 aerosol optical depth data *Sci. Total Environ.* **697** 134021

[51] Chen S, Liang Z, Webster R, Zhang G, Zhou Y, Teng H, Hu B, Arrouays D and Shi Z 2019 A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution *Sci. Total Environ.* **655** 273–83

[52] Shapley L S 1953 A value for n-person games *Contrib. Theory Games* **2** 307–17

[53] Lundberg S M and Lee S-I 2017 A unified approach to interpreting model predictions *Proceedings of the 31st International Conference on Neural Information Processing Systems* (*December 2017 Red Hook, NY, USA*) **2017** 4765–74

[54] Statista 2018 Vegetarismus und Veganismus [Vegetarianism and Veganism]. (available at: https://de-statista-com.pxz.iubh.de:8443/statistik/studie/id/27956/dokument/vegetarismus-und-veganismus-statista-dossier/)

[55] Asano Y M and Biermann G 2019 Rising adoption and retention of meat-free diets in online recipe data *Nat. Sustain.* **2** 621–7

[56] Swissveg 2017 Veggie survey 2017 (available at: www.swissveg.ch/veggie_survey?language=en)

[57] Lusk J L 2017 Consumer research with big data: applications from the food demand survey (FooDS) *Am. J. Agric. Econ.* **99** 303–20

[58] Milford A B, Le Mouël C, Bodirsky B L and Rolinski S 2019 Drivers of meat consumption *Appetite* **141** 104313

[59] Godfray H C J, Aveyard P, Garnett T, Hall J W, Key T J, Lorimer J, Pierrehumbert R T, Scarborough P, Springmann M and Jebb S A 2018 Meat consumption, health, and the environment *Science* **361** eaam5324

[60] Paslakis G *et al* 2020 Prevalence and psychopathology of vegetarians and vegans–results from a representative survey in Germany *Sci. Rep.* **10** 1–10

[61] Mullee A *et al* 2017 Vegetarianism and meat consumption: a comparison of attitudes and beliefs between vegetarian, semi-vegetarian, and omnivorous subjects in Belgium *Appetite* **114** 299–305

[62] Pfeiler T M and Egloff B 2018 Examining the 'Veggie' personality: results from a representative German sample *Appetite* **120** 246–55

[63] Allès B, Baudry J, Méjean C, Touvier M, Péneau S, Hercberg S and Kesse-Guyot E 2017 Comparison of sociodemographic and nutritional characteristics between self-reported vegetarians, vegans, and meat-eaters from the NutriNet-Santé study *Nutrients* **9** 1023

[64] OECD 2019 Education at a Glance 2019

[65] Moser S and Kleinhückelkotten S 2018 Good intents, but low impacts: diverging importance of motivational and socioeconomic determinants explaining pro-environmental behavior, energy use, and carbon footprint *Environ. Behav.* **50** 626–56

[66] Nielsen K S *et al* 2020 Improving climate change mitigation analysis: a framework for examining feasibility *One Earth* **3** 325–36

[67] Editorial 2019 Drivers of diet change *Nat. Sustain.* **2** 645

[68] Gilligan J M 2019 Modelling diet choices *Nat. Sustain.* **2** 661–2

[69] Eker S, Reese G and Obersteiner M 2019 Modelling the drivers of a widespread shift to sustainable diets *Nat. Sustain.* **2** 725–35

[70] Niamir L, Ivanova O and Filatova T 2020 Economy-wide impacts of behavioral climate change mitigation: linking agent-based and computable general equilibrium models *Environ. Model. Softw.* **134** 104839

[71] Van Den Berg N J, Hof A F, Akenji L, Edelenbosch O Y, Van Sluisveld M A E, Timmer V J and Van Vuuren D P 2019 Improved modelling of lifestyle changes in integrated assessment models: cross-disciplinary insights from methodologies and theories *Energy Strategy Rev.* **26** 100420

[72] Wiedmann T, Lenzen M, Keyßer L T and Steinberger J K 2020 Scientists' warning on affluence *Nat. Commun.* **11** 3107

[73] Bianchi F, Dorsel C, Garnett E, Aveyard P and Jebb S A 2018 Interventions targeting conscious determinants of human behaviour to reduce the demand for meat: a systematic review with qualitative comparative analysis *Int. J. Behav. Nutr. Phys. Act.* **15** 102

[74] Saura J R and Bennett D R 2019 A three-stage method for data text mining: using UGC in business intelligence analysis *Symmetry* **11** 519

[75] Pellert M, Schweighofer S and Garcia D 2020 The individual dynamics of affective expression on social media *EPJ Data Sci.* **9** 1

[76] Bouman T, Steg L and Dietz T 2020 Insights from early COVID-19 responses about promoting sustainable action *Nat. Sustain.* **4** 194–200

[77] Shah H 2020 Global problems need social science *Nature* **577** 295

[78] Lazer D M J *et al* 2020 Computational social science: obstacles and opportunities *Science* **369** 1060–2