A Catalogue of UK household datasets to monitor transitions to sustainable diets

Remco Benthem de Grave, Institute of Neuroscience, Newcastle University, UK r.benthemdegrave2@newcastle.ac.uk

Niki A. Rust, Newcastle University, UK niki.rust@ncl.ac.uk

Christian J Reynolds, Department of Geography, University of Sheffield, and the Barbara Hardy Institute, University of South Australia c.reynolds@sheffield.ac.uk, ORCID 0000-0002-1073-7394 (Corresponding author).

Anthony W. Watson, Human Nutrition Research centre, School of Biomedical Sciences, Newcastle University UK. anthony.watson@newcastle.ac.uk

Jan D Smeddinck, Open Lab, Newcastle University jan.smeddinck@ncl.ac.uk

Diogo M. de Souza Monteiro, Centre for Rural Economy, School of Natural and Environmental Sciences, Newcastle University, diogo.souza-monteiro@ncl.ac.uk

Draft unformatted post-print (the version of the paper after peer-review, with revisions having been made), Accepted to Global Food Security, 9th of December 2019.

Please see Global Food Security website for final version:
https://www.sciencedirect.com/science/journal/22119124

Abstract: There is growing international consensus that current patterns of food consumption are not sustainable and global change is needed. Understanding the mechanisms for a transition towards more sustainable diets requires systematic temporal monitoring at the individual or household level. Whilst many countries collect panel data on food expenditure and dietary intake, these datasets are often not designed to monitor progress towards dietary sustainability, therefore using them to understand how or why diets are becoming more or less sustainable can prove challenging. What is also lacking is a curated dataset catalogue or a library where all relevant data could be easily accessible to enable such evaluation. Our aim was to identify, classify and describe existing food expenditure and diet datasets available in the UK and to assess the extent to which they can be used to monitor transitions to sustainable diets. We found that despite the large number of datasets tracking UK individual or household food purchases and consumption over time, these datasets are not suited to understand how and why individuals are transitioning to sustainable diets. With the exception of proprietary datasets, most datasets only collect data annually, making it challenging to understand fine-scale behavioural change over shorter timeframes. There is an opportunity to design and implement an open-access UK sustainable diets data collection effort at the individual and household level. These efforts can be complemented with recent innovations in data science methods and digital technologies – such as dietary intake trackers – that, along with supporting individuals in their dietary behaviour change, may enable the collection of high-quality datasets.

Keywords: Panel data; food consumption; sustainable diets; data science; digital technologies; review.

Highlights:

- The current open-access UK datasets have limited effectiveness to monitor fine-scale transitions to sustainable diets.
- No single dataset recorded purchased and consumed quantities, alongside attitudes/perceptions of dietary sustainability and food consumption or purchase.
- Multiple UK datasets can be used to collectively conduct analyses of general trends and to compare different cohorts regarding the changes toward sustainable dietary patterns.
- Not all UK datasets are linked to databases containing environmental impact information.

1. Introduction

Current food purchase and consumption patterns are leading to unhealthy diets (Kearney, 2010), which in turn are linked to increased prevalence of non-communicable diseases, such as obesity, type 2 diabetes and cardiovascular disease (Aston et al., 2012; Blundell and Cooling, 2000). Moreover, there is mounting evidence that the production, processing, transport and final preparation of food to support current dietary patterns have increasing environmental and social costs, creating an unsustainable food and agricultural system that leads to increasing eutrophication, greenhouse gas emissions, biodiversity loss, and food insecurity (Aleksandrowicz et al., 2016; Green et al., 2015; Poore and Nemecek, 2018; Tilman and Clark, 2014; Willett et al., 2019).

As the evidence of the contribution of food production and consumption to the deterioration of planetary health becomes clear, so too does the need to help consumers choose more sustainable diets (Willet et al, 2019). The Food and Agricultural Organisation of the United Nations (FAO) (Burlingame and Dernini, 2012) and the first and second US National Academies of Sciences, Engineering and Medicine Workshops on Sustainable Diets, Food and Nutrition (Institute of Medicine, 2014; National Academies of Sciences, Engineering, and Medicine et al., 2019) suggest that sustainable diets must be affordable and acceptable, as well as being healthy and nutritionally balanced with low environmental impact. Transitioning towards sustainable diets is directly related to all of the United Nations Sustainable Development Goals[1]. There is thus a clear need for rapid, international change in how we produce and consume food. Importantly, changes in demand patterns will help lead to changes in production (Horton, 2017); indeed, Ingram (2017) argues that we need to change the way we look at food systems and, rather than emphasizing the need to increase production, we should focus on managing demand.

While there is increasing scientific consensus over the need to shift to more sustainable diets, there is less clarity on how to implement that transition. Food choices are complex and have numerous determinants. They are influenced by geographical, economic and social factors along with a mix of local, regional and national government policies, as well as business strategies. Due to these multiple influences, it is vital to systematically monitor the effectiveness of different interventions and assess if, how, where and which dietary transitions are occurring. To understand these trends, it is important to identify what datasets are currently in the public domain that monitor individual or household food expenditure and consumption both at home and away from home with a regular frequency to determine micro-level change.

In many developed countries there are both private and public data collection efforts collecting information on food expenses, consumption patterns and nutrition[2]. However, to the best of our knowledge, these data sets are not curated or catalogued in a systematic way and then made available to the policy or research community to conduct further analysis that can be used to inform policy and practice.

Having a catalogue or list of datasets can be beneficial for undertaking future research in this area. Examples of such research is De Keyzer et al. (2015), Perignon et al. (2017), as well as Bandy et al.

[1] https://www.stockholmresilience.org/research/research-news/2016-06-14-how-food-connects-all-the-sdgs.html

[2] For example the World Bank Global Consumption Database compiles food expenditures across food and drinks expenses from a nationally representative sample of developing countries households (http://datatopics.worldbank.org/consumption/sector/Food-and-Beverages). Similar datasets are available from international organizations like the OECD, the European Union and the national statistics of all high-income countries.

(2019), who conducted systematic reviews of food consumption datasets investigating evidence of progress to health or sustainable diets. They found important gaps and limitations regarding the applicability of datasets for monitoring transitions to sustainable food consumption behaviour. The reviews indicate that there does not appear to be a systematic data collection effort capturing all the dimensions of sustainability. However, databases currently exist that allow estimating nutritional values, greenhouse gas emission (GHGE) and cost from purchased or consumed products (see for instance www.ggdot.org, (Hobbs et al., 2015; Horgan et al., 2016; Monsivais et al., 2013)).

This paper aims to start filling this gap by:

1. providing an overview of existing private and public datasets on food purchases, as well as consumption and dietary intake patterns in the UK and,
2. discussing their suitability to assess transitions and changes towards sustainable diets, to offer a catalogue of existing data sources enabling a monitoring or assessment of transitions to sustainable diets in the UK, as well as highlighting the limitations and opportunities of available datasets.

The UK is an interesting starting point and case for observation with respect to this topic because the sustainability of current UK diets has been questioned (Reynolds et al., 2019a,b; Reynolds et al., 2015; Wrieden et al., 2017). The UK is committed to meeting the UN Sustainable Development Goals, recently also declaring the goal of reaching 'net zero' carbon emissions by 2050 (Pye et al., 2017; Walker et al., 2019), and has an actively engaged political and civil society in developing approaches to improving the current dietary and environmental situation. Moreover, the UK has a strong tradition and capacity to collect data on food purchase and consumption (Oddy, 2003; Orr, 1937) and is in the process of developing a National Food Strategy[3] that focuses on human and environmental health, to which an understanding of current food consumption trends will be integral.

Along with providing a comprehensive overview and discussion of available datasets on UK food expenditure and consumption patterns to support future data collection efforts, we also provide suggestions for approaches to improving the completeness, quality, and linking of existing datasets, and we discuss the potential for improved data collection and monitoring with digital technologies. As such, next to informing further research, this work provides guidance and evidence on improving data collection that can lead to better monitoring and understanding of transitions towards more sustainable diets. The outcomes can therefore be helpful to policy makers, research an industry alike.

3. **Methods**

We aimed to identify datasets that researchers can use to assess trends in individual and household dietary behaviour that can be used to track the level of sustainability in their food consumption, expenditures, purchases and dietary intakes. We therefore searched for UK datasets that had temporal information on individuals or household food purchases/expenditure –i.e. panel data– or consumption –i.e. dietary surveys. From these initial criteria, we added a second layer, using a range of sustainability dimensions as the second criterion for selection, i.e. we identified in the panel datasets whether they contained measures of:

1. healthiness of diets /purchases (estimated using nutrition profile tables);

2. affordability (using price and income information);

[3] https://consult.defra.gov.uk/agri-food-chain-directorate/national-food-strategy-call-for-evidence/

3. environmental sustainability (specifically whether the data contained information on carbon footprints or water use)[4,5].

Using these criteria, a first list of datasets was created from authors combined knowledge of (publications about) data collection efforts describing diets in the UK. Next we consulted the UK data service (see https://www.ukdataservice.ac.uk/get-data/themes/food.aspx). To identify additional datasets, we contacted researchers through personal networks who undertake empirical analysis of food consumption. We also reached out to private companies that collect diet information (not necessarily in the UK) and to expert groups such as the Food and Climate Research Network (FCRN) Google group (https://groups.google.com/forum/#!topic/fcrn-l/TRMs4BnUWYc).[6]

To be able to reconstruct a complete diet, we defined that the data should cover at least one complete consumption day (e.g. through a 24h Dietary Recall (24h-DR), a Diet Diary (DD), an extensive Food Frequency Questionnaire (FFQ), or a purchase diary of at least a week). To focus on UK population dietary change, we excluded datasets that focus exclusively on children or the very old, as well as datasets that consist of secondary data collection efforts (i.e. merging data collection efforts done elsewhere).

For each dataset identified, we collected characteristics and metadata from the description that accompanied the dataset. In some cases, we referred to the original survey questionnaires, the raw data, or to publications that use the dataset to find this information. Where we required further information, institutions were contacted to verify entries and asked for missing information, though not all data holders returned answers (see table 2).

One of the main challenges we faced was inconsistency across the way diets and nutrition have been measured and reported. For example, some datasets record food consumption, others food purchases. These cannot easily be combined, but they do complement each other, as there is high correlation between what is purchased and what is consumed – with food waste data then used to further 'triangulate (Reynolds et al. 2019a)'. Some datasets maybe be combined using matching methods[7] (such as propensity score matching) which enable a construction of a comparison group that is similar to the group of interest. However, there are caveats to these methods, namely that the variables on which the matching is being made may have been collect in different ways and may not be capturing the same characteristics used for the matching process.

---

[4] We acknowledge that one of the limitations of this methods is that we have not referred datasets that are used to construct these panels. For instance, there are several food composition tables available in the UK that are the basis for the nutrition information provided in commercial panel data. Undoubtedly there is a need to curate those data sources, but that is beyond our goal on this paper.

[5] Please note, as very few datasets have the capacity to immediately calculate these aspects, that for aspects of (1) healthiness and (3) environmental sustainability, we assessed whether each dataset contained 'sufficient' purchase/consumption data to calculate these dimensions by combining this data with tables of nutrition profiles and GHGE impact of the diet. We acknowledge that there are multiple additional factors that can be used to measure sustainability, such as water use, land use, biodiversity loss etc. however as stated in the main text, GHGE has multiple linked datasets already in wide use.

[6] Final searches (and citation mining) of Google Scholar were carried out using search terms combinations of "Diet", "Food", "UK", "Recall", "Cohort", "Questionnaire", "Diary", as well as the dataset names to find any additional datasets.

[7] Matching methods are statistical and econometric methods were developed to combine datasets collecting similar information on different units observation (see Rosenbaum and Rubin (1983){
P.R. Rosenbaum, D.B. Rubin **The central role of the propensity score in observational studies for causal effects** Biometrika, 70 (1983), pp. 41-55}) to enable causal analysis by constructing a counterfactual.

4. Results

This section presents the datasets identified and provides further details on those that met our main and secondary criteria. In table 1, we list all datasets that were identified in our search, highlighting in bold those meeting our inclusion criteria. In table 2, we describe, in detail, nine datasets that provided a complete overview of at least one full day of consumption or purchase data that can be accessed for research purposes (noting which data is proprietary).

Table 1: List of UK household panel datasets gathering data on food expenditure and consumption. Public datasets are those collected by governmental agencies or funded by public research funds. Private sources are those collected by commercial companies, generally through home or retail scanners, surveys or through apps. Public datasets are divided into *open* or *restricted*, with restricted meaning that further access permissions where institutional associations need to be verified and sometimes special permission requests need to be provided. Private datasets are divided into those that are available for a fee and those that are generally not shared outside the company (restricted private datasets).

Dataset or survey name	Public		Private	
	Open	Restricted	Fee	Restricted
EPIC Norfolk (Day et al., 1999)		√		
EPIC Oxford (Davey et al., 2003)		√		
Family Food module of Living Cost and Food Survey (LCFS) (Department For Environment and Office For National Statistics, 2017) (Office For National Statistics, 2019)		√		
Fenland study ("Fenland Technical Summary - MRC Epidemiology Unit," n.d.)		√		
Kantar consumption panel			√	
Kantar purchase panel			√	
National Diet and Nutrition Survey (NDNS) (Laboratory and Research, 2019)		√		
UK Women Cohort Survey (UKWCS) (Cade et al., 2015)		√		
UKBiobank (Sudlow et al., 2015)	√			
Health Survey for England		√		
1000 family study		√		
85+ study		√		
ASH30		√		
ALSPAC		√		
FAO statistics	√	√ [2]		
Food and Drink in Scotland		√		
Gateshead Millennium Cohort		√		
GfK (company)			?[1]	
Global Dietary Database (GGD)		√		
Loyalty card data collections (e.g. Dunnhumby, Tesco, Sainsbury, Waitrose)			√	√[3]
Million Women Study		√		
MyFitnessPal (company)				√
Nielsen (company)			?[1]	
Scottish Health Survey		√		
Slimming world (company)				√
Weightwatchers (company)				√

[1] Data for the UK for these companies may not be available, but this was not conclusively verified (the companies did not respond to an information request). [2] Greater detail available via application for restricted data for some areas.[3] Some Loyalty card data available through UKDS and the CDRC.

Table 1 can be considered a tentative index, where we categorise the datasets identified according to their ownership and accessibility.

Table 2 presents and characterizes the nine datasets that met our main criteria. Next, we briefly explain the characteristics of these data in three dimensions: sampling and recruitment, data collection methods and economic information therein.

<<<<Table 2 here>>>

Study design, recruitment and sample characteristics

Three types of designs can be recognized in the overview. First, two of the nine studies concerned non-cohort studies (National Diet and Nutrition Survey, and the Family Food Module of Living Cost and Food Survey, FFM-LCFS). Both of these cross-sectional studies targeted UK households using a multistage stratified sampling strategy in which households were identified from Postcode Address Files (PAF) and recognized as small users, and clustered in Primary Sampling Units (PSUs). Households were then drawn from a number of PSUs. Samples sizes ranged from about 1000 participants annually in the NDNS to 6000 households annually.

Second, six datasets concerned cohort studies (EPIC Norfolk, EPIC Oxford, the Fenland Study, the UKBiobank and the UK Women's Cohort Survey, UKWCS, Million Women Study). Targeted populations varied considerably. Some studies targeted specific diets (non-red-meat-eating, vegetarian), some geographical regions (Norfolk, Cambridgeshire) and two study targeted women only. All studies targeted a middle-age range with participant ages ranging 20-79. NHS registers and membership lists were used to recruit people. Cohort sizes of ranged from roughly 12,500 (Fenland Study) up to approximately 211,000 (UKBiobank) or 688,000 (Million Women), although sample sizes at the level of individual recordings range 1,600-100,000.

The datasets collected by Kantar are the only commercial datasets and the only ones that monitor participants' diets over an unrestricted time frame (4x per year with 10,000 people in the consumption panel and 30,000 people in the purchase panel). Advertisements on social media were used to recruit people, although more targeted methods were also used to obtain a representative sample size.

Dietary assessment methods, administration method and method of portion size estimation

A variety of methods to assess dietary consumption or purchases can be found between and within the databases. These include Food Frequency Questionnaires (FFQ), 24-Hour Dietary Recalls (24h-DR), Diet Diaries (DDs), and purchase diaries. There are well known completion biases with all food intake questionnaires when assessing the dietary intake of a free living population; with a linear association between participant burden and accuracy. None of the datasets assessed used the "gold standard" duplicate diaries to assess food intake and most used standard portions to assess food quantities. It must also be noted that datasets which convert food intake to nutrient and energy intake use food composition tables which are limited by the small number of foods they include and the age of the data within the dataset. Therefore, only a "best fit" approach was used to crudely estimate intake.

Food Frequency Questionnaires (FFQs) were used in four studies. These questionnaires asked about habitual consumption frequency in the past 12 months on a range of food items (28 to 217 food items). Participants were requested to rate their consumption frequency from never to six per day on nine frequency choices. Some exceptions to this are that one study (UKWCS) used a 10-point frequency scale and two smaller FFQs in EPIC Oxford used a 6-frequency scale. Portion sizes were generally estimated by framing the question such that it asked for the consumption of standard portion sizes. The standard portion size was then described with the item or category, for example one sausage or one portion of carrots. Some questionnaires omitted portion size and only asked for a frequency – this was to determine 'general' diet over a period of time rather than what was eaten on a specific day. We note that some of the smaller FFQs do not include a full range of foods, only categories of interest to the study. However, other assessments in the same study do. The mini FFQ's were included for completeness.

The 24h-DR was used in three studies. These asked about the consumption of the previous day. Methods used varied from pen and paper recordings, accompanied with suggestions on standard portion sizes, to online forms that required to rate their portion sizes in standard measures. The 24h-DRs were all self-administered, either at the test centre or at home.

Diet Diaries (DDs) were used in five studies. These asked the participants to track their consumption for several days (ranging between studies from 4 to 7 days). In both EPIC studies and the NDNS paper, DDs were used in combination with suggestions for standard portion sizes, supported by pictures of various portion sizes that participants could refer to. In the UKWCS, participants were asked to list weight or volume of consumed products which had to be measured or read from packaging (standard measures were allowed on some occasions). The DD in Kantar was performed on a computer. Participants selected per meal the products that they had used, but did not specify amounts consumed.

Purchase diaries where used in two studies. The FFM-LCFS used pen and paper entries or allowed participant to attach their receipts. In the Kantar purchase panel, participants were asked to scan each purchase receipt using a digital clicker. Both purchase diaries are self-administered and completed at home. One of the limitations of purchased data is that they are only proxies for consumption, as they don't factor in wastage (though it can be examined with further inquiries or complementary studies) or the delayed consumption. However, this type of data has information of food prices and collects data on disposable income and permits estimation of expenditure by category.

Economic information

Income is recorded for five out of the nine studies we describe (the NDNS, the FFM-LCFS, the Fenland Study and both Kantar datasets), while prices and/or expenditure are also recorded in the purchase panels (FFM-LCFS and Kantar datasets). One of the problems with recording economic and income information in the datasets we identified is that it is not consistent. For example, the Kantar data enables a verifiable estimation of weekly expenditure as it is based on actual shopping receipts, but this does not necessarily provide accurate information on what is actually consumed, nor does it distinguish who in the household consumes what. Another issue is that some datasets collect information on individuals, while others do so across households which prevents a combination of different datasets. Still, insofar as these datasets capture information on disposable income and purchases, they enable an assessment of affordability. Moreover, it may enable comparisons across segments of the population and identify opportunities to improve the sustainability of diets within the budget limits of household. When geographical information on location of households is available it may be possible to understand how the food retail and service environment may determine the food choices.

A full economic assessment of transitions to sustainable diets would need to include other variables that are not currently collected, for example time spent planning, shopping and preparing meals. There are datasets that provide information on time use (for example the Gershuny and Sullivan (2017) survey on how much time different groups of the population spend their time), which can be used to estimate more accurately the costs of sustainable diets.

Environmental information

There was no environmental impact information found within the datasets surveyed. However, GHGE emission datasets have been linked to multiple datasets presented in table 1. This includes the NDNS (Bates et al. 2019) and LCFS (Wriden et al 2019). We also found that the USDA's FoodAPS (Boehm et al. 2018) and the European Food Standard Agencies FoodEx2 (Reynolds et al. 2019c) have also been matched to GHGE databases.

2. Discussion

In this study, we have identified, classified and described nine datasets on diet, food consumption, or expenditure that are available in the UK to the research community. Individually, each dataset has limited effectiveness to monitor transitions to sustainable diets and for direct comparisons between datasets. This is because they were not designed for either of these purposes. The datasets use different units of observation, sampling sizes[8], sampling rates, and study durations. In addition, the datasets recorded either food purchased or consumed. In this regard, our findings are consistent with the data limitations identified by Perignon et al. (2017), who found that there is a lack of relevant and good-quality datasets for assessing the environmental, health and socio-economics impact of current diets.

However, we propose that collectively these datasets have the potential to assess transitions and changes towards sustainable diets in the UK. This is because they are complementary and can become elements of multi-layered analysis combining food consumption or purchase with other information affecting the households or individuals on which data is collected. For these purposes, the identified datasets have to be matched with other existing databases containing geographical information, further socio-economic information and environmental impact information of the foods consumed or purchased. As we pointed out in the results, some of these datasets already existed, though necessarily easily accessible. As already mentioned, matching methods (Rosenbaum and Rubin (1983); Stuart (2010)) are increasingly used to combine datasets and construct counterfactuals that enable causal analysis when, as is the case, it is challenging to design suitable experiments. While this matching may not always be feasible and could be labour-intensive to varying degrees, due to the different levels of food classification and dimensions for data-aggregation in each database, there are already ways to automate the mapping and linking of dietary and environmental impact databases (Eftimov et al., 2017). Even if they are not linked directly to environmental impacts, these databases can still be used to collectively conduct analyses at the social-economic strata level to investigate general trends and to compare different cohorts regarding the changes in dietary patterns.

In the best case, a collaborative and coordinated data collection effort - that takes account of possible linkages, and upcoming data needs - must be part of any new food strategy for the UK. This strategy could extend beyond the datasets identified in this paper to include linkages to food composition tables, lifecycle and environmental impact studies from different food categories, along with data from alternative production systems, and archived consumer survey instruments. It is beyond the purpose of this study to provide those sources of information, but we acknowledge the need for such a strategy and repository of complementary datasets that could be easily searched and used. For example, a preliminary search for Composition tables in the UK identified a Governmental source of data (the CoFID- Composition of Foods Integrated dataset) and the Carter et al (2016) new branded UK composition database. Along with a list of databases, it may be informative to provide potential

[8] For some datasets it is uncertain whether they present a representative sample of the British population.

users with a quality assessment of the data in repositories commenting on the methods used to collect the data and its limitations.

The public datasets we identified are generally accessible, have a snapshot nature, and are suitable to evaluate how different groups have changed diets and facilitate cross sectional analysis. The value of the household food purchases panel data (such as Kantar) is that it enables researchers to observe transitions with a much finer granularity. However, this analysis has the caveat that it does not capture individual consumption, but rather expenditures. Still, it enables comparison on how different households are changing consumption of a given food category and whether they are shifting to healthier, more sustainable food categories, as well as across household types, and time periods (52 weeks over a year in the case of Kantar, or weekly once a year e.g. for LCFS). In isolation, these datasets do not necessarily gather information on the health status of the household they recruit. In addition, there is a lack of detail in current panel data on the traceability and origin of food; this additional information is needed to truly understand sustainability of different foodstuffs.

It should be highlighted that there is a certain degree of self-selection bias on the households that are included in both public and private panels that were reviewed. Moreover, these datasets have not inquired about households' attitudes to - or perceptions of - sustainable dimensions of food consumption or purchase (this would be required to understand reasons why people make changes in what they eat). Moreover, there is limited information about the home and neighbourhood context as well as on the food preparation and consumption practices with which to explore more deeply what may motivate or hinder transitions at the households or individual level. Indeed, the food availability landscape is not necessarily captured in the datasets we have identified. However, those factors are important determinants of consumption and purchase. Consequently, as the existing datasets do not carry data on 1) attitudes and 2) the food environment there must be caution when interpreting this data to assess and draw conclusions as to what may have changed dietary behaviour and consumption/purchase patterns over time.

Still, the complementarity between the more frequent and rich information on products gathered in panel data and the broad coverage of large cohort studies presents a clear opportunity for assessing general transitions to sustainable diets. The household panel data could be employed to identify trends and micro-responses to interventions, in turn the cohort studies can be used to confirm how they are impacting broader aggregate measures. Another opportunity lies with matching both private and public datasets to geographical information (which is recorded in differing detail in each dataset) to further our understanding of how changes in regional or urban food policies may be affecting consumption patterns, as well as environmental and health outcomes.

To overcome the aforementioned limitations of current datasets and to develop new datasets, we suggest harnessing technological developments to better assess dietary transitions and changes towards sustainable diets. We therefore briefly highlight the potential of digital wearable devices to collect data on food choices, as well as the use of data science methods to provide new methods of data harmonization and mapping.

In principle, data science methods (including frequentist statistics, probabilistic methods, data matching as well as different techniques from machine learning and artificial intelligence) can be used for two main purposes with respect to the existing datasets:

1) improving the data-quality and reducing sparsity (filling gaps, e.g. data imputation (Jerez et al., 2010)),

2) linking datasets (e.g. through auto-correlation) ("Automated census record linking: a machine learning approach," n.d.),

3) clustering datasets or supersets, creating new sectioning or subsets (e.g. using autoencoders (Baldi, 2012)),

4) optimizing future / ongoing data collection (Sra et al., 2011) and

5) prediction.

At the same time, with the growing capabilities and affordability of sensors and increased computational capacity easily available in the cloud, digital technology, including devices and software applications opens interesting opportunities for improving data collection and research efforts. Digital data streams can be very complex and have a high sampling rate – which can at times even emulate real-time "natural fidelity" recording, compared to what is feasible with more traditional data collection efforts. This area can be split into four main elements:

1) quantified self and community applications with a) self-reporting tools, such as consumption / intake trackers (Bradley et al., 2016), or b) habit tracking / forming apps (Stawarz et al., 2015),

2) general dietary information tools (Boulos et al., 2015),

3) professional practice support (Simons et al., 2012) and

4) indirect information sources (such as product sales data, raw materials uptake / tracking, supply-chain monitoring, distributed ledgers, as well as production and transport cost /energy expenditure monitoring).

The ethical implications (and possibilities for additional bias) due to the use of such technologies, sensors, wearables, and the internet of things are of considerable extent and beyond the scope for this paper. Possible future research questions include: when and how should researchers be allowed to gain access to data from wearables? How can researchers ensure that an individual's data is used with care? How can researchers ensure that we are not neglecting harder-to-access members of society such as the poor? Who pays for these wearables? And how do we overcome the "big brother" nature of these devices?

5. Conclusions and future work

We identified and classified existing data sources with the potential to be used in research on monitoring transitions towards more sustainable diets in the UK. We present a catalogue of datasets classified in key sustainability dimensions and discuss potential of these datasets for such analysis. We conclude that neither of the datasets fulfils the requirements for reliable monitoring or prediction. Most of the datasets are also limited to traditional data sources, such as survey responses. This clearly suggests two pathways for future work: improving the quality and enable matching of the existing data sets, as well as a broader effort to collect coherent data on transitions towards more sustainable diets that combines - in a single data collection instrument - individual-level data, including motivations, and objective behaviour and food consumption over regular time periods. This instrument needs to be carefully designed tacking into account existing datasets with complementary information, such as food composition tables, environmental impact assessment and economic information. If designed and implemented ethically, digital technologies can play a key role and enable novel approaches and insights. These technologies include software with supportive algorithms and user interfaces, which can, for example, gauge shopping behaviour, shopping, and the engagement with – and social communication about – diet information sources, as well as (sensing) hardware devices that allow for objective measurements e.g. of eating behaviour.

We also acknowledge that we have not documented and critically examined other data sources that complement the datasets we covered – this includes a) food composition tables; b) datasets of environmental outcomes for different foods; c) food price datasets; d) survey data on attitudes surrounding food purchasing behaviour. These were outside the scope of this particular effort. We recognize that such information is valid and believe that future work should fill that gap. Also, we hope our limitations inspire researchers interested in measuring sustainable diets to create and curate a library of datasets facilitating further work in this area.

6. Bibliography

Aleksandrowicz, L., Green, R., Joy, E.J.M., Smith, P., Haines, A., 2016. The impacts of dietary change on greenhouse gas emissions, land use, water use, and health: A systematic review. PLoS One 11, e0165797. doi:10.1371/journal.pone.0165797

Aston, L.M., Smith, J.N., Powles, J.W., 2012. Impact of a reduced red and processed meat dietary pattern on disease risks and greenhouse gas emissions in the UK: a modelling study. BMJ Open 2. doi:10.1136/bmjopen-2012-001072

Automated census record linking: a machine learning approach [WWW Document], n.d. URL https://open.bu.edu/handle/2144/27526 (accessed 6.30.19).

Bates, R.L., Chambers, N.G. and Craig, L.C.A., 2019. Greenhouse gas emissions of UK diets. Proceedings of the Nutrition Society, 78(OCE2). **doi:10.1017/S0029665119000910**

Baldi, P., 2012. Autoencoders, unsupervised learning, and deep architectures. Proceedings of ICML workshop on unsupervised and transfer learning 37.

Bandy, L., Adhikari, V., Jebb, S., Rayner, M., 2019. The use of commercial food purchase data for public health nutrition research: A systematic review. PLoS One 14, e0210192. doi:10.1371/journal.pone.0210192

Blanquer, M., García-Alvarez, A., Ribas-Barba, L., Wijnhoven, T.M.A., Tabacchi, G., Gurinovic, M., Serra-Majem, L., 2009. How to find information on national food and nutrient consumption surveys across Europe: systematic literature review and questionnaires to selected country experts are both good strategies. Br. J. Nutr. 101 Suppl 2, S37–50. doi:10.1017/S0007114509990572

Blundell, J.E., Cooling, J., 2000. Routes to obesity: phenotypes, food choices and activity. Br. J. Nutr. 83 Suppl 1, S33–8.

Boehm, Rebecca, Parke E. Wilde, Michele Ver Ploeg, Christine Costello, and Sean B. Cash. "A comprehensive life cycle assessment of greenhouse gas emissions from US household food choices." Food policy 79 (2018): 67-76.

Boulos, M., Yassine, A., Shirmohammadi, S., Namahoot, C., Brückner, M., 2015. Towards an “internet of food”: food ontologies for the internet of things. Future Internet 7, 372–392. doi:10.3390/fi7040372

Bradley, J., Simpson, E., Poliakov, I., Matthews, J.N.S., Olivier, P., Adamson, A.J., Foster, E., 2016. Comparison of INTAKE24 (an Online 24-h Dietary Recall Tool) with Interviewer-Led 24-h Recall in 11-24 Year-Old. Nutrients 8. doi:10.3390/nu8060358

Burlingame, B., Dernini, S., 2012. SUSTAINABLE DIETS AND BIODIVERSITY DIRECTIONS AND SOLUTIONS FOR POLICY, RESEARCH AND ACTION.

Cade, J.E., Burley, V.J., Alwan, N.A., Hutchinson, J., Hancock, N., Morris, M.A., Threapleton, D.E., Greenwood, D.C., 2015. Cohort profile: the UK women's cohort study (UKWCS). Int. J. Epidemiol. doi:10.1093/ije/dyv173

Carter MC, Hancock N, Albar SA, et al. Development of a New Branded UK Food Composition Database for an Online Dietary Assessment Tool. Nutrients. 2016;8(8):480. Published 2016 Aug 5. doi:10.3390/nu8080480

Davey, G.K., Spencer, E.A., Appleby, P.N., Allen, N.E., Knox, K.H., Key, T.J., 2003. EPIC-Oxford: lifestyle characteristics and nutrient intakes in a cohort of 33 883 meat-eaters and 31 546 non meat-eaters in the UK. Public Health Nutr. 6, 259–269. doi:10.1079/PHN2002430

Day, N., Oakes, S., Luben, R., Khaw, K.T., Bingham, S., Welch, A., Wareham, N., 1999. EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. Br. J. Cancer 80 Suppl 1, 95–103.

De Keyzer, W., Bracke, T., McNaughton, S.A., Parnell, W., Moshfegh, A.J., Pereira, R.A., Lee, H.-S., van't Veer, P., De Henauw, S., Huybrechts, I., 2015. Cross-continental comparison of national food consumption survey methods--a narrative review. Nutrients 7, 3587–3620. doi:10.3390/nu7053587

Department For Environment, F., Office For National Statistics, 2017. Living Costs and Food Survey, 2015-2016. UK Data Service. doi:10.5255/ukda-sn-8210-4

Eftimov, T., Korošec, P., Koroušić Seljak, B., 2017. StandFood: Standardization of Foods Using a Semi-Automatic System for Classifying and Describing Foods According to FoodEx2. Nutrients 9. doi:10.3390/nu9060542

Fenland Technical Summary - MRC Epidemiology Unit [WWW Document], n.d. URL http://doi.org/10.22025/2017.10.101.00001 (accessed 6.28.19).

Gershuny, J., Sullivan, O. (2017). United Kingdom Time Use Survey, 2014-2015. [data collection]. UK Data Service. SN: 8128, http://doi.org/10.5255/UKDA-SN-8128-1

Green, R., Milner, J., Dangour, A.D., Haines, A., Chalabi, Z., Markandya, A., Spadaro, J., Wilkinson, P., 2015. The potential to reduce greenhouse gas emissions in the UK through healthy and realistic dietary change. Clim. Change 129, 253–265. doi:10.1007/s10584-015-1329-y

Hobbs, D.A., Lovegrove, J.A., Givens, D.I., 2015. The role of dairy products in sustainable diets: modelling nutritional adequacy, financial and environmental impacts. Proc. Nutr. Soc. 74. doi:10.1017/S0029665115003572

Horgan, G.W., Perrin, A., Whybrow, S., Macdiarmid, J.I., 2016. Achieving dietary recommendations and reducing greenhouse gas emissions: modelling diets to minimise the change from current intakes. Int. J. Behav. Nutr. Phys. Act. 13, 46. doi:10.1186/s12966-016-0370-1

Horton, P., 2017. We need radical change in how we produce and consume food. Food Sec. 9, 1323–1327. doi:10.1007/s12571-017-0740-9

Ingram, J., 2017. Perspective: Look beyond production. Nature 544, S17. doi:10.1038/544S17a

Institute of Medicine, 2014. Sustainable Diets: Food for Healthy People and a Healthy Planet: Workshop Summary, The National Academies Collection: Reports funded by National Institutes of Health. National Academies Press (US), Washington (DC). doi:10.17226/18578

Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M., Franco, L., 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med 50, 105–115. doi:10.1016/j.artmed.2010.05.002

Kearney, J., 2010. Food consumption trends and drivers. Philos. Trans. R. Soc. Lond. B, Biol. Sci. 365, 2793–2807. doi:10.1098/rstb.2010.0149

Laboratory, M.E.W., Research, N.S., 2019. National Diet and Nutrition Survey Years 1-9, 2008/09-2016/17. UK Data Service. doi:10.5255/ukda-sn-6533-13

Monsivais, P., Perrigue, M.M., Adams, S.L., Drewnowski, A., 2013. Measuring diet cost at the individual level: a comparison of three methods. Eur. J. Clin. Nutr. 67, 1220–1225. doi:10.1038/ejcn.2013.176

National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Food and Nutrition Board, Food Forum, 2019. Sustainable diets, food, and nutrition: proceedings of a workshop, The National Academies Collection: Reports funded by National Institutes of Health. National Academies Press (US), Washington (DC). doi:10.17226/25192

Oddy, D.J., 2003. From plain fare to fusion food: British diet from the 1890s to the 1990s. Boydell Press, Woodbridge, Suffolk.

Office For National Statistics, 2019. Living Costs and Food Survey, 2017-2018. UK Data Service. doi:10.5255/ukda-sn-8459-1

Orr, J.B., 1937. Food health and income: Report on a survey of adequacy of diet in relation to income. Macmillan and Company Limited,, London.

Perignon, M., Vieux, F., Soler, L.-G., Masset, G., Darmon, N., 2017. Improving diet sustainability through evolution of food choices: review of epidemiological studies on the environmental impact of diets. Nutr. Rev. 75, 2–17. doi:10.1093/nutrit/nuw043

Poore, J., Nemecek, T., 2018. Reducing food's environmental impacts through producers and consumers. Science 360, 987–992. doi:10.1126/science.aaq0216

Pye, S., Li, F.G.N., Price, J., Fais, B., 2017. Achieving net-zero emissions through the reframing of UK national targets in the post-Paris Agreement era. Nat. Energy 2, 17024. doi:10.1038/nenergy.2017.24

Reynolds, C., Hodgson, H., Bajzelj, B., 2019b. An improved picture of the UK diet: Linking production, consumption and waste data to provide a better dietary picture. WRAP.

Reynolds, C.J., Horgan, G.W., Whybrow, S., Macdiarmid, J.I., 2019a. Healthy and sustainable diets that meet greenhouse gas emission reduction targets and are affordable for different income groups in the UK. Public Health Nutr. 22, 1503–1517. doi:10.1017/S1368980018003774

Reynolds, C.J. Schmidt Rivera, X. Frankowska, A. Kluczkovski, A. da Silva J. T., Bridle S. L. Levy, R. Rauber, F. Quadros, V. P. Balcerzak, A. Sousa , R. F. Ferrari, M. Leclercq , C. Koroušić Seljak, B. Eftimov,T. (2019c) A Pilot Method Linking Greenhouse Gas Emission Databases To The Foodex2 Classification, Livestock, Environment and People (LEAP) Conference 2019. Saïd Business School, Oxford 10th December 2019

Reynolds, C.J., Macdiarmid, J.I., Whybrow, S., Horgan, G., Kyle, J., 2015. Greenhouse gas emissions associated with sustainable diets in relation to climate change and health. Proc. Nutr. Soc. 74. doi:10.1017/S0029665115003985

Simons, L.P., Hampe, J.F., Guldemond, N.A., 2012. Designing Healthy Consumption Support: Mobile application use added to (e) Coach Solution. Bled eConference 34.

Sra, S., Nowozin, S., Wright, S.J. (Eds.), 2011. Optimization for machine learning. The MIT Press. doi:10.7551/mitpress/8996.001.0001

Stawarz, K., Cox, A.L., Blandford, A., 2015. Beyond Self-Tracking and Reminders: Designing Smartphone Apps That Support Habit Formation, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI' ' ’15. Presented at the the 33rd Annual ACM Conference, ACM Press, New York, New York, USA, pp. 2653–2662. doi:10.1145/2702123.2702230

Stuart E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical science : a review journal of the Institute of Mathematical Statistics, 25(1), 1–21. doi:10.1214/09-STS313

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12, e1001779. doi:10.1371/journal.pmed.1001779

Tilman, D., Clark, M., 2014. Global diets link environmental sustainability and human health. Nature 515, 518–522. doi:10.1038/nature13959

Walker, P., Mason, R., Carrington, D., 2019. Theresa May commits to net zero UK carbon emissions by 2050 | Environment | The Guardian. The Guardian.

Willett, W., Rockström, J., Loken, B., Springmann, M., Lang, T., Vermeulen, S., Garnett, T., Tilman, D., DeClerck, F., Wood, A., Jonell, M., Clark, M., Gordon, L.J., Fanzo, J., Hawkes, C., Zurayk, R., Rivera, J.A., De Vries, W., Majele Sibanda, L., Afshin, A., Chaudhary, A., Herrero, M., Agustina, R., Branca, F., Lartey, A., Fan, S., Crona, B., Fox, E., Bignet, V., Troell, M., Lindahl, T., Singh, S., Cornell, S.E., Srinath Reddy, K., Narain, S., Nishtar, S., Murray, C.J.L., 2019. Food in the Anthropocene: the EAT-Lancet Commission on healthy diets from sustainable food systems. Lancet 393, 447–492. doi:10.1016/S0140-6736(18)31788-4

Wrieden, W.L., Leinonen, I., Barton, K.L., Halligan, J., Goffe, L., 2017. Is the UK diet sustainable? Assessing the environmental impact, cost and nutritional quality of household food purchases. Proc. Nutr. Soc. 76. doi:10.1017/S0029665117001811

Wrieden, W., Halligan, J., Goffe, L., Barton, K., & Leinonen, I. 2019. Sustainable Diets in the UK—Developing a Systematic Framework to Assess the Environmental Impact, Cost and Nutritional Quality of Household Food Purchases. Sustainability, 11(18), 4974. doi:10.3390/su11184974