User Guide: Household emissions dataset

Contents

1 Introduction

This document accompanies the GB Household emissions dataset available to download from the Centre for Sustainable Energy website (www.cse.org.uk/gb-household-emissions-dataset).

The dataset was initially developed as part of a Joseph Rowntree Foundation-funded research project: '*Distribution of Carbon Emissions in the UK: Implications for Domestic Energy Policy'*. The full project report and four-side summary document are available from the JRF website[1].

The dataset and this guidance document have been compiled and made publicly and freely available by the Centre for Sustainable Energy as part of its Open Data project funded by the Esmée Fairbairn Foundation.

This document fulfils two key functions:

(i) To provide a brief overview explaining the methodology behind the dataset; and
(ii) To provide detailed guidance on how to use the dataset.

IMPORTANT! Please note: This document is not designed to provide a full and detailed methodology of how the dataset was developed. This is available in a separate document:

> 'Technical Report: Developing the datasets used in the modelling and analysis for the study: *Distribution of carbon emissions in the UK: Implications for domestic energy policy'.* Available at: www.cse.org.uk/downloads/file/jrf_social_impacts_technical_report.pdf

It is highly recommended that users of the GB Household emissions dataset read this paper, which includes detailed information on the different data sources and how emissions estimates were derived from survey data (including the carbon emissions factors applied).

[1] www.jrf.org.uk/publications/carbon-emissions

2 Background to the dataset

The data contained within the dataset of GB household emissions is derived from a number of different nationally representative surveys, namely the: Expenditure and Food Survey, National Travel Survey and Civil Aviation Authority Air Passenger Survey (Table 1).

Table 1. Surveys used to derive household emissions estimates

Survey	Source & Years	Input (raw survey data)	Output
Expenditure and Food Survey (EFS) ***['Recipient' survey]***	ONS 2004/05 -2007	Expenditure on all household fuels (electricity, gas and non-metered)	Annual consumption of all household fuels (kWh) and associated CO_2 emissions ($kgCO_2$).
National Travel Survey (NTS) ***['Donor' survey]***	DfT 2002 -2006	Private vehicle mileage Distance travelled by all modes of public transport	Annual CO_2 emissions from all personal (non-business) travel by private vehicle and public transport.
Air Passenger Survey (APS) ***['Donor' survey]***	CAA 1999 - 2008	Start airport, destination airport (international only) and flight class for all GB leisure passengers	Distance travelled and associated CO_2 emissions from (non-business) international aviation.

Each of these surveys is undertaken independently and therefore exists as a distinct dataset. However, they are all designed to be representative (through sampling and weighting design) and they each contain socio-demographic information (for example household income, dwelling type, tenure etc). Using variables common to two or more datasets, it is possible to develop imputation models to take (or rather 'impute') data from one survey into another.

A process of survey harmonisation and imputation (see Figure 1 and Box 1) was therefore applied to combine data from each of these surveys into a single dataset to represent carbon emissions of households in Great Britain. The socio-demographically representative sample of UK households surveyed in the ONS **Expenditure and Food Survey** (EFS)[2] provides the core or 'recipient' dataset. It is also the source of data used to derive estimates of household emissions from the consumption of energy in the home (electricity, gas and all non-metered fuels).

Data from four EFS years was combined for the purpose of the JRF-funded research (survey years 2004/5 to 2007), generating a sample size of over 22,500 cases.

Additional data on household emissions from personal travel was derived from:

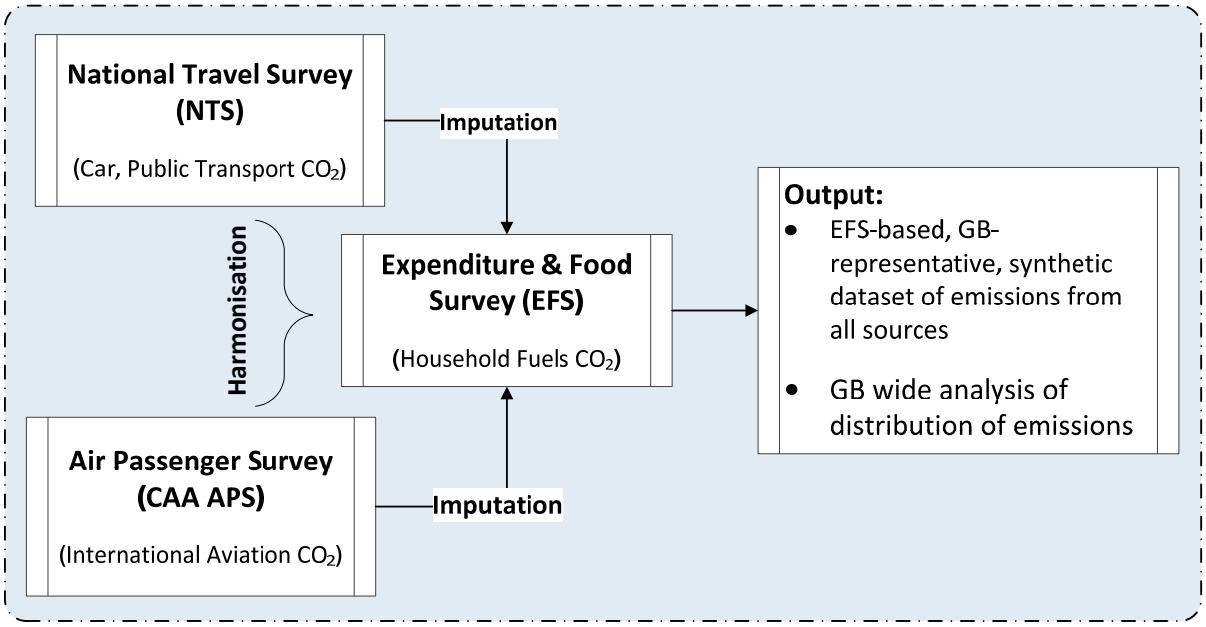
- **National Travel Survey** (NTS, 2002-2006): used to derive estimates of household emissions from all personal (defined as leisure and commuting) travel by private road vehicle and public transport.

[2] In April 2001 the *Family Expenditure Survey* (FES) and *National Food Survey* (NFS) were combined to form the *Expenditure and Food Survey* (EFS), which completely replaced both series. From January 2008, the EFS became known as the *Living Costs and Food* (LCF) survey module of the *Integrated Household Survey* (IHS).

- **Civil Aviation Authority Air Passenger Survey** (CAA APS, 1999 to 2008): used to derive estimates of household emissions from international air travel.

The emissions estimates derived within the NTS and CAA APS datasets were then imputed to the EFS to create a single 'synthetic' dataset representing annual household carbon emissions from the consumption of energy in the home and all personal travel, for all households in Great Britain.

Figure 1. Overview of methodology



Box 1. Overview of key technical stages in the creation of the GB household emissions dataset

1. **Derive carbon emissions estimates from survey data**: taking raw survey data (for example, actual household expenditure on heating fuels (in the EFS), or annual distance travelled by private car for leisure purposes (in the NTS)) methods were developed to apply relevant carbon emissions factors and derive estimates of carbon emissions. Different methods were required for each survey, reflecting the different nature and types of information collected and provided by each survey (Table 1). Full details of the methodology applied to each survey are provided in the separate technical report[3].

2. **Survey harmonisation**: In creating a single dataset representative of household carbon emissions in Great Britain, data has to be imputed from each of the individual survey datasets into a single dataset (based on the EFS). Before the imputation can be undertaken, however, the survey datasets need to be 'harmonised'. This essentially means ensuring that key concepts used in each of the surveys are defined and measured in the same way (for example, income can be defined in a number of ways – disposable, gross etc; for income to be used in an imputation model to impute data from one survey dataset to another, it must be defined in the same way). The technical report includes detail on the survey harmonisation process, including a list of the harmonised variables.

[3] 'Technical Report: Developing the datasets used in the modelling and analysis for the study: *Distribution of carbon emissions in the UK: Implications for domestic energy policy.'* Available at: www.cse.org.uk/downloads/file/jrf_social_impacts_technical_report.pdf

3. **Impute data**: Multiple imputation was used to impute carbon emissions data from one survey to another. This process involves developing predictive models where the 'predictor' variables are the survey-harmonised socio-demographic variables. Several different imputation models had to be developed for the purpose of this study – one for each of the variables imputed to the EFS. Finally, adjustments were made post-imputation to ensure imputed data summed to match original totals in the donor surveys.

3 How to use the dataset

This 'How-to' guide outlines some key steps to enable users to understand and analyse the data provided in CSE's GB household emissions dataset. These principally relate to joining additional variables from the Expenditure and Food survey to the GB household emissions dataset to enable analysis of the emissions data by different socio-demographic descriptors.

Please note: The GB emissions dataset is designed to be analysed using SPSS and the instructions provided in this document reflect this. We also assume users have a good level of understanding of structure of the Expenditure and Food Survey datasets. For more information about the EFS and the various datasets available please refer to ONS documentation[4].

IMPORTANT! The GB household emissions dataset is designed for use and analysis by users familiar with both SPSS and the Expenditure and Food Survey datasets.

3.1 Step 1: Request GB household emissions dataset from CSE

To request the data please email **opendata@cse.org.uk** stating your name, your organisation (if applicable), the data format you require (SPSS or Excel) and the reason you are interested in this data. We are very interested to find out what people are doing with this data, to help us understand its utility and how it could be improved.

3.1.1 Content of the dataset

As outlined above the GB household emissions dataset is based on the Expenditure and Food Survey household data file, to which additional information on household travel emissions has been imputed. Several years of EFS data were combined to increase the sample size. The GB household emissions dataset therefore contains 22,591 cases (individual rows) each of which is representative of 1 or more households in Great Britain (see below for information on weightings).

A total of ten different variables are provided in the GB household emissions dataset, as shown below (Table 2). All emissions estimates represent annual estimates of kilograms of carbon dioxide.

[4] For more information about the EFS (now the 'Living Costs and Food Survey') please refer to survey documentation: www.esds.ac.uk/findingData/efsTitles.asp

Table 2. Variables provided in the GB household emissions dataset

Variable Name	Variable label
Case_ID	CSE Case ID
Annual_weight	Adjusted annual weight [a]
Private_vehicle_CO2	Private vehicle emissions (kgCO2/yr)
Public_transport_CO2	Public transport travel emissions (kgCO2/yr)
International_aviation_CO2	International air travel emissions (kgCO2/yr)
Total_transport_CO2	Total transport emissions (kgCO2/yr)
Household_heating_CO2	Household heating emissions (kgCO2/yr)
Household_power_CO2	Household power (kgCO2/yr)
Total_household_fuels_CO2	Total household fuels emissions (kgCO2/yr)
Total_household_CO2	Total household emissions (personal travel and household fuels) (kgCO2/yr)

[a] The weighting is explained later in section 3.4.

3.2 Step 2: Accessing Expenditure and Food Survey data

The GB household emissions dataset provided by CSE contains household level emissions estimates only. Licensing restrictions prevent inclusion of any original EFS survey data. However, the user may access EFS data themselves, which can then be joined to the emissions dataset (see Step 3). To do this, there are two key steps:

Step 2a: Register with UK Data Archive

The user must have a valid registration with the UK Data Archive in order to access EFS survey datasets. Registration can be done here: www.data-archive.ac.uk/sign-up/credentials-application

Step 2b. Download EFS Survey data for years corresponding with CSE's GB household emissions dataset

Having registered with the UK data archive, use your log-in to access and download EFS datasets to correspond with the survey years used in creating CSE's GB household emissions dataset, namely EFS survey years: 2004/05; 2005/06; 2006; and 2007.

3.3 Step 3: Joining EFS data to the GB household emissions dataset

There are several key steps needed in order to join any additional EFS variables to the CSE GB emissions dataset for the purpose of analysis:

Step 3a: Identify EFS variables of interest. A huge range of variables is provided, grouped into a number of different datasets as part of the Expenditure and Food survey. These include 'raw' and 'derived' datasets and person-level and household-level variables. The user should be familiar with the nature and content of the different EFS datasets in order to identify variables of interest for joining to the CSE emissions dataset. Please refer to the ONS EFS user guides for further information about the different EFS datasets and data contained therein (full supporting documentation available to download from UKDA with the EFS datasets).

Step 3b: Aggregate EFS data to household level if needed. CSE's GB emissions dataset is provided at the household level, therefore all data to be joined to this dataset needs to also be at the household

level. It is likely that most EFS variables of interest and relevance for analysing the GB emissions data will already be available as derived household level data fields (in the EFS dataset: "[YEAR]_dvhh_ukanon"). However, some data may be at person level. This can (and needs to) be aggregated to household level using the case number (EFS variable 'case') and 'aggregate' function in SPSS before joining to the emissions dataset.

Step 3c: Create new 'Case_ID' in the downloaded EFS household datasets. As described above, four years of EFS data were combined to create CSE's GB household emissions dataset. In doing so, a new unique ID for every case (row) was created. This new case number is provided in the GB household emissions dataset ('Case_ID'). To join (or 'merge' in SPSS terms) variables from the EFS datasets onto the CSE GB household emissions dataset, the user must create the CSE 'Case_ID' in the UKDA EFS datasets downloaded in Step 2. The CSE 'Case_ID' is derived from the original EFS case number (EFS variable name 'case') supplied in the UKDA EFS datasets. SPSS syntax for computing CSE 'Case_ID' in each EFS survey year respectively is provided in Box 2 below.

Box 2. SPSS syntax for computing CSE variable 'Case_ID'

For EFS survey year 2004-05:
compute Case_ID = 10000 + case .
exe .

For EFS survey year 2005-06:
compute Case_ID = 20000 + case .
exe .

For EFS survey year 2006:
compute Case_ID = 30000 + case .
exe .

For EFS survey year 2007:
compute Case_ID = 40000 + case .
exe .

Step 3d: Merge variables from the downloaded EFS datasets to the GB emissions dataset using new 'Case_ID'. Having created the CSE Case_ID in the UKDA EFS 2004/05 to 2007 (inclusive) household-level datasets, the user can then use this variable to match cases in the GB emissions dataset to add variables of interest from the EFS. There are two key points to note when merging EFS variables to the household emissions dataset:

1 **UK to GB:** The emissions dataset represents GB households, whilst the EFS covers the whole of the UK. Cases in the EFS representing households in Northern Ireland should therefore be excluded in the join (see Step 3d (iii) below).

2 **Financial years to calendar years:** From January 2006 the Expenditure and Food Survey has been conducted on a calendar year basis, rather than the previous financial year basis. As a result, the last three months of financial survey year 2005/06 and the first three months of

the 2006 calendar year survey overlap. To avoid duplication of survey data in the multi-year combined GB household emissions dataset, the first three months of 2006 (January to March) were dropped, as shown by the smaller sample size for 2006 in Table 3.

There are several different approaches that can be taken for joining EFS variables to the GB household emissions dataset. The simplest and recommended approach is to:

(i) First combine the EFS household datasets for each survey year into a single SPSS file. This can be done using the 'Data -> Merge files -> Add cases' function in SPSS.
(ii) Join variables of interest from the combined EFS household dataset (now containing all four survey years) on to the GB household emissions dataset (using the 'Data -> Merge files -> Add variables' function in SPSS, matching on key variable 'Case_ID')
(iii) Once the join is complete, delete all cases in the newly merged file where the variable 'Annual_weight' is missing (using the 'Select if' function in SPSS[5]). This will delete all records from the dataset that are either in Northern Ireland or from the 3 month overlapping period of survey calendar year 2005/06 to financial year 2006.

Step 3e: Verify sample count in the final merged data file. We recommend that users ensure that the (unweighted) sample count of cases in the final merged dataset matches those shown in Table 3 below for each survey year in the CSE GB household emissions dataset.

Table 3. Sample sizes and case numbers in the original EFS household datasets and the CSE GB household emissions dataset

EFS Year	Original EFS household dataset		CSE GB household emissions dataset	
	Sample count	Case number range	Sample count	Case number range
2004-05	6,798	1 to 6,798	6,265	10,001 to 16,265
2005-06	6,785	1 to 6,785	6,258	20,001 to 26,258
2006	6,645	1 to 6,645	4,528	30,001 to 36,059
2007	6,136	1 to 6,141	5,540	40,001 to 45,545

3.4 Weighting

The methodology developed and applied in this study was designed with the specific aim of creating a dataset representative of carbon emissions in Great Britain at the household level. As explained above, the dataset is based on the Expenditure and Food Survey household data file, to which additional information on household travel emissions has been imputed. Every row (case) in the EFS household level dataset is representative of a proportion of households in the UK, as indicated by case weight provided (the variable 'annual weight'; variable name 'weighta'). When applied, the count of households in the survey dataset sums to represent all households in the UK in the survey year. As several years of ONS EFS data were combined for the purpose of the research from which the GB emissions dataset stems, this annual survey weight has been adjusted to reflect this. This new adjusted weight is provided in the GB household emissions dataset and should be applied in all analysis of the data.

[5] SPSS syntax for the Select If function: Filter off. Use all. Select if(not missing(Annual_weight)). execute.

All analysis of the GB household emissions dataset should be undertaken with the adjusted annual weight applied.

Some key statistics describing the GB household emissions dataset are shown below. This includes information on: the sample size (the number of cases in the dataset – the 'unweighted' count); number of households represented (the 'weighted count'); and the sum total and average (mean) estimates of carbon emissions from each source.

We recommend that users ensure they can replicate the figures shown in the table below as a check that the data and weighting are being used correctly.

Table 4. Annual household emissions estimates derived from the GB Household Emissions Dataset

	Mean ($kgCO_2$)	Sum ($MtCO_2$)
Private vehicle emissions	2,644	64.0
Public transport travel emissions	302	7.3
International air travel emissions	1,182	28.6
Total transport emissions	4,128	99.9
Household heating emissions	3,725	90.2
Household power	1,951	47.2
Total household fuels emissions	5,675	137.4
Total household emissions (personal travel & household fuels)	9,836	238.1
Unweighted sample count	22,591	
Weighted count (thousands)	24,207	

4 Potential uses of the data

The GB household emissions dataset was analysed as part of the JRF-funded research to explore the distribution of emissions across GB households by a range of socio-demographic variables. As far as we are aware, this is the first integrated analysis of emissions from both the consumption of energy in the home and personal travel based entirely and directly on nationally representative survey data. The analysis provides new evidence and insight into who is responsible for emitting how much carbon dioxide, and identifies the relative contributions of different aspects (i.e. energy consumption in the home, private road travel and aviation) of household carbon emissions. The results of this analysis are summarised in section 4 of the main project report and in more detail in a supplementary project paper.

For results of the analysis of the GB household emissions dataset undertaken as part of the JRF-funded project, please refer to:

- The main project report: www.jrf.org.uk/publications/carbon-emissions
- Supplementary project paper: www.cse.org.uk/downloads/file/project_paper_1_household-emissions-distribution.pdf

The analysis presented in the JRF study gives an indication of the relationship between household carbon emissions and socio-demographic descriptors. However, this is limited to bivariate analysis of only a selection of socio-demographic variables. The scope and depth of this analysis could be extended to include additional variables and a multivariate analysis approach could be applied, to explore multiple relations between multiple variables simultaneously.

5 Data health warnings

5.1 Sample sizes

The unweighted count is a count of the actual number of cases in the sample. We recommend that all analysis maintains a minimum sample size of 200 cases.

5.2 Data sources

The data provided in the GB household emissions dataset is sourced from three different surveys of households in the UK/GB. We have not sought to reconcile the resulting carbon emissions totals derived from the survey data with published figures at the national level. This is because the two use very different approaches and are designed for very different purposes. Our approach, utilising representative household survey data is very much 'bottom-up', deriving emissions relating directly and explicitly to activity in the home and for personal travel, whilst national figures adopt a 'top-down' approach. For example, DECC figures for emissions by 'final user' category includes direct emissions from domestic premises, (e.g. from burning gas, coal or oil for space heating) but also includes emissions from power stations generating the electricity used by domestic consumers and emissions from refineries, coal mines from the extraction, storage and distribution of mains gas for the domestic sector.

However, the totals for household fuel consumption derived from the EFS dataset (used for estimating emissions from energy consumption in the home) have been compared with DUKES overall totals and with mean values from DECC's National Energy Efficiency Data frameworks (NEED) database. The EFS results are within +/- 5% of NEED with a maximum range of -9% and +6% by income band.

All resulting CO_2 sum totals derived from survey data for this research have also been compared at the aggregate (dataset) level with DECC GHG emissions reporting, as shown below. In addition to the above explanations, discrepancies between the figures will also relate to:

- Our values represent Great Britain only, whilst DECC figures cover the whole of the UK (including Channel Islands and Isle of Man);
- All transport-related emissions in the GB household emissions dataset represent leisure and commuting only (i.e. no business travel).

Table 5. Comparison of CO_2 totals in the GB household emissions dataset with national figures

	Source Survey	GB Survey estimate $MtCO_2$	*National data* $MtCO_2$	*Source of National Data and Notes*[6]
Household fuels	**EFS**	137.4	*154.4*	*DECC (a): Residential combustion, by final user*
Private car	**NTS Vehicle data**	64.0	*75.6*	*DECC (a): Passenger cars & motorcycles, by source*
Public transport	**NTS Journey data**	7.3	*6.4*	*DECC (a): Buses and rail, by source*
International aviation	**CAA APS**	26.7	31.8	*DECC (b): CO2e from UK international aviation bunkers*

DECC (a) Table 4: Estimated emissions of carbon dioxide (CO_2) by National Communication source category, type of fuel and end-user category, 1970-2010. Values shown represent average for the corresponding survey years.
www.decc.gov.uk/en/content/cms/statistics/climate_stats/gg_emissions/uk_emissions/uk_emissions.aspx
DECC (b) Table 8: Greenhouse gas emissions arising from use of fuels from UK international aviation bunkers. Values shown represent average for the corresponding survey years (1999-2008).

6 Data protection and intellectual property

The methodology and derived variables used to produce the dataset of GB household emissions are the Intellectual Property Rights of CSE. CSE has made this dataset publicly and freely available under an Open Data license. We welcome feedback on how others are utilising the data and are happy to answer any questions about the methodology, if these are not answered in the full methodology report.

Please send any feedback or queries to: opendata@cse.org.uk

If the user wishes to join additional variables from the EFS to CSE's GB household emissions dataset, ONS should be referenced accordingly as the source of the former[7].

[6] DECC figures are reported 'by source' or 'end user'. This difference in reporting mainly affects emissions related to electricity generation from power stations. By source, these emissions are allocated to the energy supply sector since the power stations are responsible for producing the electricity. Reporting by end-user reallocates all these emissions to the final users of the electricity, such as to homes and businesses. Hence figures quoted above are by 'end user' for residential emissions, but 'by source' for transport.
www.decc.gov.uk/assets/decc/Statistics/climate_change/407-uk-emissions-stats-faq.pdf

[7] Source: Expenditure and Food Survey, National Statistics. © Crown Copyright material is reproduced with the permission of the Controller of Her Majesty's Stationery Office (HMSO)