Administrative Data | Agricultural Research Collection (AD|ARC)

Agricultural Research Collection – England (ARCE)

An Introductory Guide for Researchers

Version 2.0

12/02/2024

Authors

Dr Paul Caskie (Agri-Food and Biosciences Institute)

Sian Morrison-Rees (Swansea University/ADR Wales)

Nicholas Webster (Welsh Government/ADR Wales)

Laura Madden (Welsh Government/ADR Wales)

Sarah Lowe (Welsh Government/ADR Wales)

On behalf of the wider AD|ARC team and project partners: Erica Chisholm (Agri-Food and Biosciences Institute), Claire Jack (Agri-Food and Biosciences Institute), Maisie Duckham (DEFRA), Thomas Pearson (DEFRA), Jen Hampton (ONS ADR Data Linkage Team), Nathan O'Connor (ONS ADR Data Linkage Team), Roya Shahrokni (ONS ADR Data Acquisition Team), Sean Scully (Swansea University), Jenny Thyer (Swansea University), Katy Addison (Welsh Government/ADR Wales), Liam Crowley (Welsh Government/ADR Wales), Matthew Curds (Welsh Government/ADR Wales), Sorcha Egan (Welsh Government/ADR Wales), Matthew Kelly (Welsh Government/ADR Wales), Rosie Kirk (Welsh Government/ADR Wales), Rachael Loftus (Welsh Government/ADR Wales), Stuart Neil (Welsh Government), Richard McFerran (Welsh Government/ADR Wales), and Freya Pryce (Welsh Government/ADR Wales).

Funding

This work is supported by ADR UK (Administrative Data Research UK), an Economic and Social Research Council (part of UK Research and Innovation) programme and by the Welsh Government.

Acknowledgements

The AD|ARC project in England notes with thanks the many collaborators who have supported the development of the Research Ready Dataset in England, including the Welsh Government, Public Health Wales and the AD|ARC Scientific Advisory Board and England Stakeholder Reference Group, DEFRA, and the ADR Data Acquisition Team and ADR Data Linkage Team at ONS.

Similarly, the AD|ARC team notes with thanks the support of other data linking projects in the development this introductory guidance, most notably the Data First and ECHILD projects.

This report describes the AD|ARC Agricultural Research Collection – England (ARCE) which includes data from DEFRA, Rural Payments Agency and the ONS. These parties do not accept any responsibility for any inferences or conclusions derived by the authors.

Glossary of Terms and Acronyms

a08	The EU Farm Structure Survey records farm activities at a holding level identified by an a08 number.
Beta Testing	Beta testing is the process of giving access to a dataset to a small subset of potential users to identify any problems before allowing wider access.
Blocking	Blocking is when you split a dataset into 'blocks' or clusters of records using a particular attribute. This limits the number of potential matches according to the blocking attribute, improving efficiency while maintaining accuracy. For instance, blocking on place of birth and sex means that a linking algorithm looking for Margaret Rutherford from Chalfont St. Peter would only search within the set of candidate matches of women born in that county.
Business ID	The Unique Identifier for Farm Businesses
CBUA	Contiguous Built Up Area classification – this classification is used by AD|ARC as a proxy for rurality
Cleaning	Data cleaning is the process of addressing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data in a dataset. Cleaning methods vary depending on the variable or dataset but may be as simple as fixing a clear typo to deleting invalid data altogether.
Data Dictionary	A Data Dictionary is a set of information describing the contents, format and structure of a dataset or database
Deterministic Matching	Deterministic Matching uses the exact correspondence between instances of an identifier or set of identifiers to produce a match. Deterministic matching assumes that unique identifiers such as names or dates of birth can be used to definitively link across datasets. Also see Probabilistic Matching.
Jaro-Winkler string comparisons	The Jaro–Winkler distance gives more favourable ratings to strings that match from the beginning for a set prefix length.
LSOA	Output Areas (OAs) are the lowest level of geographical area for census statistics and were first created following the 2001 Census. Lower layer Super Output Areas (LSOAs) are made up of groups of OAs, usually four or five. They comprise between 400 and 1,200 households and have a usually resident population between 1,000 and 3,000 persons.
Matchkey	Matchkeys tend to be fixed-length, compressed strings built from a combination of the words and numbers in a name or address such that relevant variations have the same match key value.
Missingness	Missingness is the state of being missing and is usually applied to missing values, where no data value is stored within a variable for a particular case e.g. a farm business or a person living in a farming household.
Probabilistic Matching	The process of using statistical analysis to determine the overall likelihood that two records match. Probabilistic matching is the preferred method for matching large data sets when unique identifiers such as names or dates of birth are unavailable or are not of sufficient quality to be used to definitively link across datasets. Also see Deterministic Matching.
Research Ready Dataset (RRD)	The term 'research-ready data' is used to communicate the processing of administrative data to make it available for research. However, research-readiness has multiple dimensions. In the case of AD|ARC, this means a dataset that has been built by linking a number of agricultural and wider datasets, standardising variable names and

	providing metadata in order to provide a resource researchers can readily access via the Trusted Third Parties of the UK.
Secure Research Service (SRS)	The ONS Secure Research Service (SRS) is a DEA-accredited Trusted Research Environment. The ONS SRS gives accredited/approved researchers access to de-identified, unpublished data to work on research projects for the public good.
Single Payment Scheme 2010	Single Payment Scheme (SPS or Single Farm Payment) replaced the Common Agricultural Policy (CAP) in 2003, introducing a new method of direct subsidy payments to landowners.
UPRN	The Unique Property Reference Number or UPRN is the unique identifier for every addressable location across the UK.

Contents

1 What is the aim of this guide?

This user guide is intended to support researchers in seeking access to the Agricultural Research Collection – England (ARCE). It provides an overview of the Dataset, including information on data content, quality, and security.

The guide also provides guidance on the processes required to access the Collection in England. This includes the process of becoming an accredited researcher, guidance on how to access the Secure Research Service (SRS) and the research governance processes to seek approval for access to AD|ARC for specific research projects. Links to the AD|ARC Data Dictionaries are also included.

2 What is the Agricultural Research Collection – England (ARCE)?

The Research Ready Datasets (RRDs) have been created as part of the AD|ARC project. The aim of the project is to integrate the human dimension of farming with data on farming activities, to better understand the demographic, health, education, and economic characteristics of farm households associated with different types and sizes of farm businesses. It is hoped this will provide the insight decision makers need to formulate better policy for the Sector, thereby enhancing the wellbeing of farmers and their households across the UK.

AD|ARC links de-identified electronic records that are already collected by departments across the governments of the UK. The AD|ARC datasets include variables from the following sources:

EU Farm Structure Survey 2010

Rural Payments Agency (RPA) 2010

ONS Census of Population for England and Wales 2011

Inter-Departmental Business Register (IDBR) 2006-2018

It should be noted that the AD|ARC Project Team have not carried out any imputation on the datasets. However, imputation of various kinds may have been carried out by the original data owners.

The AD|ARC England cohort is defined as: **all members of households in England that have been successfully linked to farm businesses claiming under the Single Payment Scheme in 2010.**

The cohort includes households where no individual had reported their main occupation as ‘farming’ in the Census 2011. However, we have included these households in the cohort because they lived at an address for which a farm business received a subsidy payment under the Single Payment Scheme in 2010.

This means that the cohort contains both:

- households where at least one individual reported an agricultural main occupation in the Census 2011; and
- households where no-one reported an agricultural main occupation in the Census 2011.

Because both types of households received subsidy payments, both are considered 'farming households' for analysis purposes.

The cohort includes household members of all ages, including children.

However, it should be noted that farm businesses who chose not to claim Single Payment Scheme subsidies for which they were eligible will not appear in the RRD. Farm businesses that did not meet the Single Application rules for the Basic Payment Scheme will also not appear in the RRD. Furthermore, where there was a second household residing at a different address, such as an adult offspring who had moved but was still part of the business, they will not be included unless their address was also provided in the RPA application.

The Agricultural Research Collection – England (ARCE) consists of an Individual-Level Table and a Household-Level Table, each of which is briefly described below. The methodology to produce these Tables is described in Section 3 of this User Guide.

2.1 The Individual-Level Tables

There are four individual-level tables for the AD|ARC England RRD – the table relating to the cohort of individuals living in 'farming households' ([Table Name]_INDIVIDUAL) and a table each of three control groups.

The ([Table Name]_INDIVIDUAL) includes individual-level records for all household members included in the cohort, allowing analysis to be conducted with the individual as the primary unit of analysis. The Table includes:

- the 2011 Census variables for each individual, e.g. age, gender, approximated social grade, National Statistics Socio-economic Classification, highest level of qualification, self-reported general health, whether the individual has a long-term health problem or disability etc.;
- selected agricultural variables, e.g. numbers of livestock, utilised agricultural area etc.; and
- variables derived from the Census or agricultural sources, e.g. whether the individual reported their occupation as 'farming' on the Census, farm type, farm size etc.

The AD|ARC England Data Dictionaries include more detailed information about every variable in each dataset, including the variable type, the source of the variable (e.g. Census), the format in which the data is held and the magnitude of any missingness. As noted above, the methodology to produce these Tables is described in Section 3 of this User Guide.

In addition to the records for members of households linked to farm businesses ([Table Name]_INDIVIDUAL), three further tables contain the following control groups:

- Up to 3 individuals per cohort member (i.e. per farming household member) matched on age (+/- 3 years), gender and Contiguous Built Up Area classification (CBUA) as a proxy for rurality ([Table Name]_AGEGENCBUA)
- Up to 3 individuals per cohort member (i.e. per farming household member) matched on age (+/- 3 years), gender and employment status (e.g. self-employed) in the Census 2011 ([Table Name]_AGEGENEMP)
- Up to 3 individuals per cohort member reporting their main occupation as farming (i.e. per self-reported 'farmer') matched on age (+/- 3 years), gender and a set of similar occupation codes in the Census 2011 ([Table Name]_AGEGENOCC)

Please note that individuals were selected for the control groups based purely on the individual characteristics listed above and not on either household size or household characteristics. So, for example, a female child living in an area with fewer than 2,000 residents may live with a lone parent in the control group but with two adults and three siblings in the cohort or vice versa. For example, this allows researchers to examine the relative household sizes of 'farming' households when compared with non-farming, rural controls.

For each matched control individual, the Table includes the same 2011 Census variables as for the cohort individual, however the fields for the agricultural variables are either not included or contain a null value.

The AD|ARC England Data Dictionary includes more detailed information about every variable in the dataset, including the source of the variables e.g. Census, the format in which the data is held, the values, and the magnitude of any missingness. The methodology to produce this Table is described in Section 3 of this User Guide.

2.2 The Household-Level Table

There are two household-level tables for the AD|ARC England RRD – the table relating to the cohort of farming households ([Table Name]_HOUSEHOLD) and a table for a control group ([Table Name]_CONTROL_HH).

The ([Table Name]_HOUSEHOLD) includes household-level records for each farm business included in the cohort, allowing analysis to be conducted with the household as the primary unit of analysis. The Table includes:

- the 2011 Census variables that relate to the household and property e.g. the tenure of the dwelling, household size, car/van availability, central heating etc.;
- variables relating to the farm business and activity data e.g. numbers of livestock etc.;
- variables derived from agricultural sources, e.g. Farm Type farm size etc.; and
- variables from the IDBR, e.g. turnover.

There is a household level control group:

- Up to 3 households (using Household_ID) from the Census 2011 per farming household matched via the LSOA for a similar geographical area ([Table Name]_HH_Controls).

For each matched control household, the Table includes the same 2011 Census variables as for the cohort household, however the fields for the agricultural variables are either not included or contain a null value.

Please see the Data Dictionary for more information on the variables referred to above.

3 How has the data been processed to produce the Agricultural Research Collection – England (ARCE)?

The Agricultural Research Collection – England (ARCE) brings together datasets capturing information about farm households at four different levels:

1. Individual level – the individuals who make up farm households as identified by the Person_ID in the ONS Population Census 2011. There may be multiple individuals in each farm household.
2. Holding level – the EU Farm Structure Survey records farm activities at a holding level identified by the a08 number.
3. Business level – the Rural Payments Agency data records payments made to farm businesses at a farm Business Level, identified by a Customer Number. Some farm businesses in England have multiple holdings.
4. Household level – the households identified by the Household_ID in the Census 2011.
   N.B. There are farm households in England that correctly link to multiple farm businesses.

Rural Payments Agency subsidies are paid at the farm business level. The EU Farm Structure Survey is collected at the holding level. A farming household can have several farm business IDs. Equally, more than one farming household can be associated with a farm business. A farm business can have several holdings. However, each holding will have a distinct farm business ID (i.e. holdings are not split across farm businesses or farming households).

The steps below outline how this data has been processed to create the individual and household tables mentioned above.

3.1 Preparation of the agricultural data

The preparation of agricultural datasets was done within DEFRA before the combined agricultural dataset was sent to the ADR Data Linkage Team at ONS. The file of contact details used to collect the EU Farm Structure Survey in 2010 was linked to the Rural Payments Agency data via the address details. The subsidy payments data for 2010 was appended to the EU Farm Structure Survey 2010.

The Ordnance Survey Address Base Unique Property Reference Number (UPRN) was then added to the dataset.

A record for every farm business included in the EU Farm Structure Survey for England 2010 was then uploaded to the ADR Data Linkage Team at ONS via their standard procedure.

3.2 Data Linkage between the Agricultural data and Census 2011

As noted above, the AD|ARC England cohort is defined as: all members of households in England that have been successfully linked to farm businesses claiming under the Single Payment Scheme in 2010.

The ADR Data Linkage Team at ONS conducted the linkage and provided a report, summarised below:

a) The contact data was cleaned and standardised e.g. standardisation of case and abbreviations, removal of white spaces and concatenation.
b) Linkage was conducted using hierarchical match keys from the name and address on both datasets.
c) Quality checks were undertaken on iterations of match keys.

d) Jaro-Winkler similarity scores were generated for address variables of candidate pairs.
e) Record pairs were assigned a match category based on the respective strength of this score, with record pairs with higher scores (i.e., closer to 1) retained in favour of conflicting pairs with lower scores.
f) A manual clerical review was undertaken from a sample.

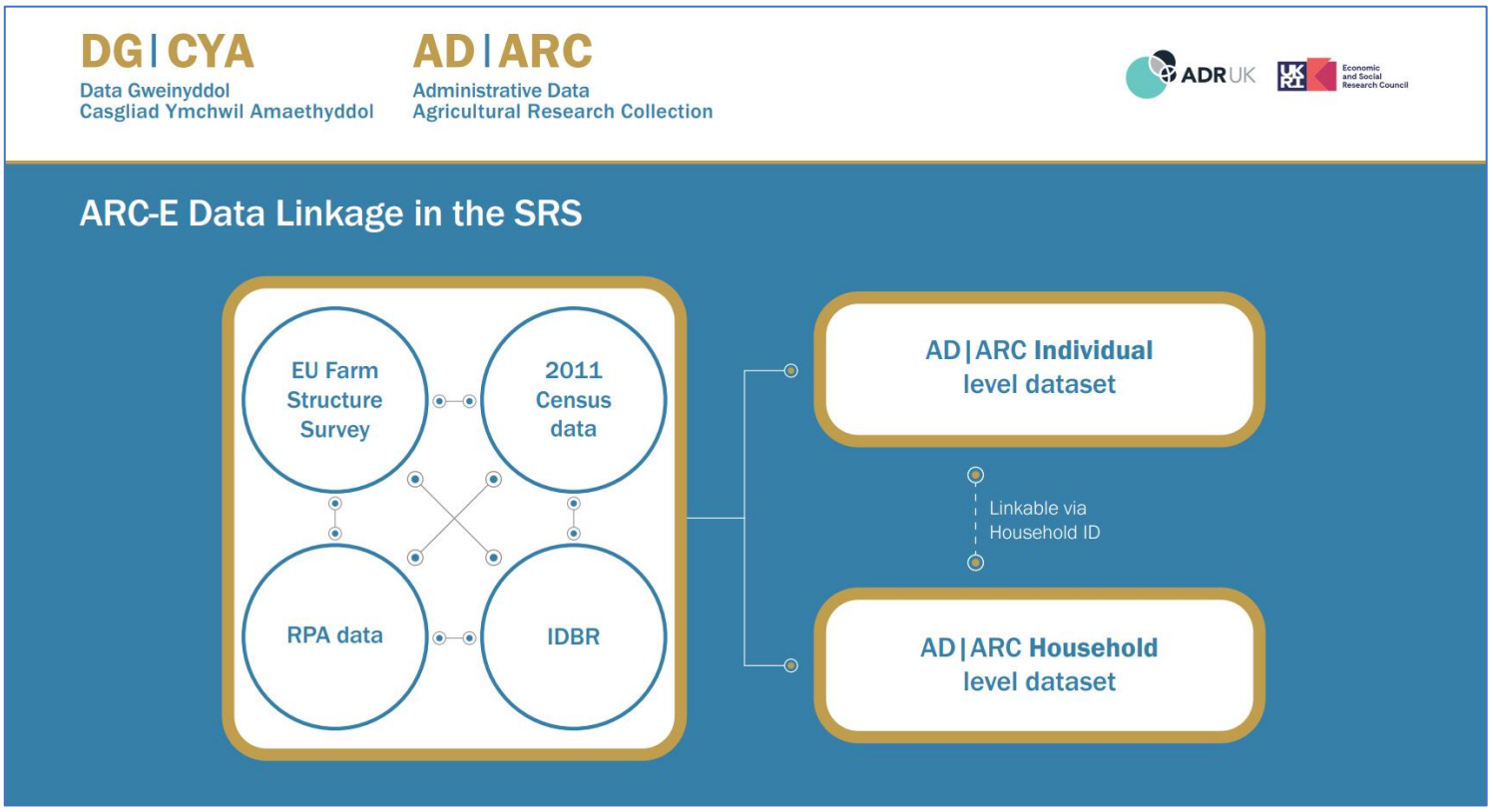
The same process was conducted to incorporate the IDBR data.

3.3 Construction of the Agricultural Research Collection – England (ARCE) Research Ready Dataset

The data was de-identified and provisioned into the SRS for the project. The attribution data was provided in separate tables alongside the de-identified linkage tables. The project then used the linkage tables to construct the Agricultural Research Collection – England (ARCE) from the attribute tables to create an individual table and a household table that include the derived variables and variables aggregated to a household level (please see Figure 1, below, for a diagram of the linking process).

As noted in Section 1, 'farmers' who chose not to claim subsidies for which they were eligible, did not meet the eligibility criteria for subsidies or lived at a second address not listed in the RPA application will not be included in the cohort.

Figure 1: Diagram of how the Agricultural Research Collection – England (ARCE) datasets were linked in the SRS



3.4 Aggregation of agricultural data to household level

As noted above, multiple Holdings (a08s) and farm businesses can be linked to the same farm household. The EU Farm Structure Survey and Rural Payments Agency variables have therefore been aggregated to the farm household level. Derived variables for a count of the number of holdings (a08_NUM) and farm businesses (business_NUM) for each household are included in the dataset (see Data Dictionary for more information).

Farm Type has been recalculated at a household level using the same Robust type 10 calculations and methodology used to calculate Farm Type in the EU Farm Structure Survey.

Because, as noted above, some farm businesses are linked to more than one household, caution is required when aggregating farm business variables.

4 Key Considerations for Research using the Agricultural Research Collection – England (ARCE)

All data has been linked and stored in either the Individual-level or the Household-level Table ready for analysis (see accompanying Data Dictionary). As noted above, additional linking may be required for more complex research questions but potential researchers should contact the core AD|ARC team who can advise on whether additional links are necessary and, if so, how to perform them.

Linking rate plus any examination of bias in linking rate to follow.

Administrative data refers to information collected for non-statistical reasons, either to enable the delivery of a public programme or service or to maintain records. Administrative data sources can be rich sources of information that can be used in quantitative analysis and evaluation without imposing an additional burden on data subjects or additional costs to data controllers. However, because they are not designed for research purposes, they often have limitations.

Although the AD|ARC project team has undertaken the steps documented above to create a Research Ready Dataset, the following limitations should be kept in mind when both undertaking analysis and reporting findings:

- The dataset is ‘research ready’ in that we have linked together several different datasets to enable research. The data is, however, not fully ‘analysis ready’. Researchers will still be required to assess the quality of the data and consider any necessary data preparation steps in accordance with their own research needs.
- The cohort should not be seen as a census of all ‘farming’ households. As noted in Sections 1 and 2.1, the cohort includes all members of households in England that have been successfully linked to farm businesses claiming under the Single Payment Scheme in 2010. ‘Farmers’ who chose not to claim subsidies for which they were eligible, did not meet the eligibility criteria for subsidies, or lived at a second address not listed in the RPA application will not be included in the cohort. This means that some individuals or households selected for the various AD|ARC control groups who have not stated a main occupation of farming in the 2011 Census may be involved in farming activities. This should be kept in mind when interpreting findings.
- Some farm businesses are linked to more than one household. Caution is therefore required when aggregating farm business variables.

- Farm businesses may make their RPA application through an agent. If this is the case, they may have been included in the data. However, agents are unlikely to have been operating from a residential address and therefore to be successfully linked to the 2011 Census.

The challenges and limitations documented above should be taken into consideration by researchers when preparing their application to use the Agricultural Research Collection – England (ARCE) database. The AD|ARC team welcome queries and/or potential collaborations prior to an application for use of the data.

When researchers are writing reports based on the analysis of AD|ARC project data and need to add data quality indicators or commentary, they should feel free to quote the text from this User Guide.

Because the AD|ARC project datasets are complex and only just beginning to be used, early research projects could be considered beta-testers and should expect to find issues with the data, which the Project Team would be grateful if they would inform the team about so that, where possible, we can address them and, where not, we can provide appropriate documentation. Equally, this User Guide should be considered a 'starter for ten' and will be updated to include additional information or explanation where users ask the Project Team questions, so please do get in touch!

5 How is the Data Protected?

All the data held in the Agricultural Research Collection – England (ARCE) is de-identified. The associated data processing is administered under the framework outlined in the Digital Economy Act (2017) (DEA), which enables government to prepare administrative data for the purposes of research, and to provide de-identified versions of those data to researchers and projects accredited by the UK Statistics Authority (UKSA). The Secure Research Service is an accredited Trusted Research Environment under the DEA.

The Secure Research Service uses the Five Safes Framework to ensure the safety and security of its stored data. This is a set of principles adopted by a range of secure research environments to provide complete assurance for data owners.

The Five Safes are:

Safe People	– trained and accredited researchers are trusted to use data appropriately
Safe Projects	– data are only used for valuable, ethical research that delivers clear public benefits
Safe Settings	– access to data is only possible using secure technology systems
Safe Outputs	– all research outputs are checked to ensure they cannot identify data subjects
Safe Data	– researchers can only use data that have been de-identified

The Agricultural Research Collection – England (ARCE) is stored in the SRS. For more information on how data is de-identified and protected in the Agricultural Research Collection – England (ARCE) please follow the links below to SRS resources.

Secure Research Service

6 How do I access the data?

Any researcher interested in accessing the Agricultural Research Collection – England (ARCE) is advised to contact the AD|ARC Project Team to discuss their proposal.

To apply to access the data, please see the guidance on the Secure Research Service website.

Prior to an application to the SRS, if not already DEA accredited, researchers will be required to complete the ONS accredited researcher training.