# Notes on the Large Data Set OCR Mathematics A (H230/H240)

These notes outline the requirements for OCR’s large data set, include some notes on interesting features of the large data set and some links to some useful resources.

Please note that the approach outlined here and within the activities to be found on the website are suggestions only; you are free to deliver this in any way that suits you, your students and your setting.

If you have any comments on the contents of this guide, including suggestions for other activities, please get in touch with the Mathematics Subject Advisors at [Maths@ocr.org.uk](mailto:Maths@ocr.org.uk) .

In the “Overview” section we take a look at the *requirements* in terms of teaching from the Department for Education and in terms of assessment from Ofqual, and introduce the large data set chosen for OCR’s Maths A specification.

In the “Key Features” section we will take a look at some important aspects of the structure and set up of the data. We will also explore some aspects of data cleaning in this section, including a list of things to look out for in the way that the data were collected and presented by the ONS, along with some ways to explore these issues and to set up your students to stumble across them.

In “Online resources” we highlight some useful places to go for more contextual information, maps, more data, tools and so on.

On the website, you will find two packs of “Starter Activities”, which provide a selection of short activities, i.e. 5-10 min and usually a response to a particular presentation of data. These include some prompt questions to use and some teacher notes on details to look out for. There are also two longer activities. These might be a substantial part of a lesson, a whole lesson or even a series of linked lessons. They can all be used and adapted in many ways.

***DISCLAIMER***

This resource was designed using the most up to date information from the specification at the time it was published. Specifications are updated over time, which means there may be contradictions between the resource and the specification, therefore please use the information on the latest specification at all times.If you do notice a discrepancy please contact us on the following email address: [resources.feedback@ocr.org.uk](mailto:resources.feedback@ocr.org.uk)

### Contents

[Notes on the Large Data Set OCR Mathematics A (H230/H240) 1](#_Toc512347403)

[Contents 2](#_Toc512347404)

[1 Overview 3](#_Toc512347405)

[1.1 Introduction 3](#_Toc512347406)

[1.2 OCR’s Large Data Set 3](#_Toc512347407)

[2 Key Features 4](#_Toc512347408)

[2.1 Geography Codes 4](#_Toc512347409)

[2.2 Data Cleaning and manipulation 5](#_Toc512347410)

[3 Online resources 7](#_Toc512347411)

[3.1 Useful maps 7](#_Toc512347412)

[3.2 Other useful websites 8](#_Toc512347413)

# 1 Overview

### 1.1 Introduction

All AS and A Level Mathematics for first teaching from September 2017 include the requirement to work with a given large data set (LDS). The criteria from the Department from Education says that

AS and A Level Mathematics specification must require students to:

* become familiar with one or more specific large data set(s) in advance of the final assessment (the data must be real and sufficiently rich to enable the concepts and skills of data presentation and interpretation in the specification to be explored)
* use technology such as spreadsheets or specialist statitstical packages to explore the data set(s)
* interpret real data presented in summary or graphical form
* use data to investigate questions arising in real contexts

Specifications should require students to explore the data set(s) and associated contexts, during their course of study to enable them to perform tasks that assume familiarity with the contexts, the main features of the data and the ways in which technology can help explore the data. Specifications should also require students to deomnstrate the ability to analyse a subset of features of the data using a calculator with standard statistical functions.

The prerelease LDS is primarily a resource for the classroom, to encourage the use of real, LDSs when learning statistics. *Some* questions in the assessment, on the statistics content, will be set in the context of the LDS, in such a way as to provide an advantage to students who have spent time exploring the data. However, this is only a small part of the exam and students will not have access to the LDS, or to a computer, in the assessment. The focus of this resource is very much on teaching and learning.

Note that the first bullet point above includes the phrase “to enable the concepts and skills of data presentation and interpretation in the specification to be explored”. This refers to that particular section of the content document, ie section 2.02 of the OCR Mathematics A specification. You are welcome to explore the full range of concepts and skills, but the focus of the requirement, and therefore of the assessment, is on data presentation and interpretation.

### 1.2 OCR’s Large Data Set

OCR’s Large Data Set consists of data taken from the 2001 and 2011 Census data for England and Wales on Method of Travel to Work and Age Structure. The data is organised by Local Authority area, giving a fine grain of detail without making the data set too large and unwieldy. The data set includes a metadata sheet which describes the various terms used and which explains some aspects of the data collection process. Rather than repeat this information here, please refer to that sheet. The terminology and information on the metadata sheet of the LDS is part of the data set, so students are assumed to be at least familiar with that terminology and information. That’s part of the point of putting it there, i.e. that you then know that these are the words students should understand. *However*, the assessment is not a test of memorisation of the details of the data, only that students have worked with the data enough to have some familiarity with the key features. We will take a more in-depth look at *some* features in section 2.

Please note that further commentary on specific aspects of the LDS can be found within the resources on the OCR website for this qualification.

# 2 Key Features

### 2.1 Geography Codes

Each Local Authority is allocated to a Region and is given a Geography code that gives some information about that Local Authority. Students do not need to know any technical details about how these authorities are set up and so forth. Essentially they need to know enough to distinguish between them so that they can build a little knowledge about their general characters.

As a simple summary:

E06: Unitary authority in England

Larger than a normal Local Authority: see the section at the bottom of the metadata sheet about how unitary authorities were created between 2001 and 2011. They don't need to know the details; just that these were created by combining Local Authorities and that some of the information from 2001 has been lost by this amalgamation.

E07: Non-metropolitan district in England  
A ‘normal’ Local Authority: not heavily urbanised and (in general) with less infrastructure for public transport.

E08: Metropolitan borough in England

A more heavily urbanised Local Authority with good provision of public transport infrastructure.

E09: London borough.

A Local Authority in Greater London. these tend to share features with Metropolitan Boroughs.

W06: Unitary authority in Wales

All Local Authorities in Wales fall in this category, regardless of their different characters and sizes.

### 2.2 Data Cleaning and manipulation

This list is a few details covering data cleaning issues, and any data manipulation which was done to create comparable data sets between 2001 and 2011, along with a couple of possible stumbling blocks to watch out for.

1. Unitary authorities: these were created in 2009, so the original 2001 data set has more LAs than the 2011 set. It also means that we needed to do some work on the 2001 data, amalgamating LAs into these UAs, otherwise it would be necessary for you (or your students) to do a significant amount of work before having workable data. There is a many-to-one relationship between UAs and LAs (as can be seen in the metadata sheet) so this did not require anything other than amalgamation of the data. It means that we’ve got sets we can compare, but we have lost a bit of granularity from the 2001 data set.
2. This amalgamation is why we can’t retain mean and median age from KS02, as mentioned in the metadata point 5. It is worth discussing whether or not these two measures are reconstructable from KS02, particularly because neither is possible to do exactly for the UAs, and the estimate of mean would be more accurate than that of the median. Of course the LAs which remained unchanged could take their values straight from KS02.
3. Look at Isles of Scilly for Underground, Metro, Light Rail and Tram: in 2011 it has a zero; in 2001 it has a dash. This is how the data came from the ONS and it shows how easy it is to create difficulties.  
   The issue with it is that if you find the mean of that column in 2011 then Excel will divide by 348, but in 2001 it will divide by 347 because it doesn’t recognise the dash as a value.  
   It is a small point, but the hyphen is the only piece of “bad” data as such and it makes the population mean 2046 instead of 2040 so it is not entirely insignificant.   
   You can fix this in Excel by using AVERAGEA() instead of AVERAGE(), but it will of course affect other calculations involving that cell, for example if you find the proportion of each LA travelling by Underground, Metro, Light Rail and Tram then you get #VALUE in that cell in Excel, and even using AVERAGEA() will not help you….
4. The columns in Method of Travel are in a different order in 2011 and 2001. Again this is how it came from the ONS sources. It can be fixed easily enough by moving the Taxi column in one of the data sets.  
   It is easy to set up students to fall over this one by giving them calculations to perform of graphs to produce from one of the sets and then copy-pasting across to the other set.
5. The “All Categories of people in Employment” column in Method of Travel 2011 has been created by summing the other columns because the original data set only had the values for each category along with the category of people not in employment.  
   This latter category has been kept in column P for interest, though it cannot be replicated easily in 2001 without some more research. The gap before column P has been left in so that students using the ctrl+ features of Excel to select blocks of data don’t include this column in error.
6. An interesting point of discrepancy between the Age Structures data sets is that in 2001 they treated reported ages of over 110 years as invalid, but in 2011 they used 115 years as the cut off. This is only really relevant if trying to estimate some measure from the (grouped) data because the “Age 90 and over” class is not the same width in the two sets.
7. The Method of Travel for a given person is the method used for the longest part, by distance, of the usual journey to work. So for example one shouldn’t really say that it is only those people in the category “Driving a car or van” who use a car to get to work, this language is a bit loose. However, this is the language used in the sample assessment material and throughout the resources, if only because the precise language would be too cumbersome. You should simply remind your students of what the categories actually mean.

If you want a “clean” version of the data set to work with or to give to students, then carry out the following steps

1. move the Taxi column in 2001 Method of Travel to column I.
2. For Isles of Scilly in Method of Travel to Work 2001change “-“ to “0”.
3. Delete Column P of Methods of Travel 2011.

# 3 Online resources

### 3.1 Useful maps

Please note, this maps are offered as useful tools to get a sense of where places are and what sort of places they are.

**Regions of England and Wales**

This wikipedia page includes a useful map of the regions of England, which is great for getting a high-level sense of where places are.

<https://en.wikipedia.org/wiki/Regions_of_England>

**Local Authority Districts**

This map from the Office of National Statistics shows all Local Authority Districts and Unitary Authorities. The text is searchable, so it is an easy way to find out where a given LA is.

<https://data.gov.uk/dataset/local-authority-districts-counties-and-unitary-authorities-december-2012-map-in-the-united-king2>

**A Landcover Atlas of the United Kingdom**

This Atlas, including a full set of Local Authority level maps, is a good way to get a quick view of a Local Authority and how urban or rural it is.

<https://figshare.com/articles/A_Land_Cover_Atlas_of_the_United_Kingdom_Document_/5266495>

<https://figshare.com/articles/A_Land_Cover_Atlas_of_the_United_Kingdom_Maps_/5219956>

**OCR GCSE Geography (9-1) resource**

This resource created for OCR GCSE Geography (9-1) includes a very useful map of depravation, which might be an interesting tool to look at a given local area of interest. Scroll down the left hand side until you get to Choropleth maps. This is initially centred on Portsmouth, but it is zoomable and searchable.

<http://ocr.maps.arcgis.com/apps/MapJournal/index.html?appid=b56d896f39194809b47f233ba89f9271>

### 3.2 Other useful websites

As mentioned earlier, a classic spreadsheet is not the only way to interact with the LDS. These links are a starting point for exploring other tools.

**Tinkerplots**

TinkerPlots is a simple, but powerful, data visualization and modelling tool developed for use by schools.

<https://www.tinkerplots.com/>

**Gnumeric**

Gnumeric is an open-source spreadsheet which is particularly useful for statistics, including easy ways to create data visualisations, take repeated samples etc. It is primarily developed for Linux, but can be run in Windows. This is a technical option which requires some knowledge of IT to install, but is very easy to use.

<http://www.gnumeric.org/>

**Office for National Statistics**

The original source for our data and a huge wealth of possible resource. It includes the Nomis service, which has an easy to use function for searching for tables from the 2001 and 2011 census.

<https://www.ons.gov.uk/>

<https://www.nomisweb.co.uk/>

**Geogebra**

Geogebra is a free dynamic mathematics tool, including graphing, 3D graphing, geometry, CAS and (most importantly) a spreadsheet. the website also hosts a vast collection of materials.

<https://www.geogebra.org/>

**R**

Finally, R is a free software environment for statistical computing and graphics. It is readily available on a wide variety of operating systems. The interface and language may take some getting used to, but the flexibility and power rewards the effort.

<https://www.r-project.org/>

We’d like to know your view on the resources we produce. By clicking on ‘[Like](mailto:resources.feedback@ocr.org.uk?subject=I%20liked%20the%20AS%20and%20A%20Level%20Maths%20A%20Teacher%20Guide%20about%20the%20large%20data%20set)’ or ‘[Dislike](mailto:resources.feedback@ocr.org.uk?subject=I%20disliked%20the%20AS%20and%20A%20Level%20Maths%20A%20Teacher%20Guide%20about%20the%20large%20data%20set)’ you can help us to ensure that our resources work for you. When the email template pops up please add additional comments if you wish and then just click ‘Send’. Thank you.

Whether you already offer OCR qualifications, are new to OCR, or are considering switching from your current provider/awarding organisation, you can request more information by completing the Expression of Interest form which can be found here: [www.ocr.org.uk/expression-of-interest](http://www.ocr.org.uk/expression-of-interest)

Looking for a resource? There is now a quick and easy search tool to help find free resources for your qualification:   
[www.ocr.org.uk/i-want-to/find-resources/](http://www.ocr.org.uk/i-want-to/find-resources/)

**OCR Resources**: *the small print*OCR’s resources are provided to support the delivery of OCR qualifications, but in no way constitute an endorsed teaching method that is required by the Board, and the decision to use them lies with the individual teacher. Whilst every effort is made to ensure the accuracy of the content, OCR cannot be held responsible for any errors or omissions within these resources.   
© OCR 2018 - This resource may be freely copied and distributed, as long as the OCR logo and this message remain intact and OCR is acknowledged as the originator of this work.

OCR acknowledges the use of the following content: Adapted from materials written for St Paul’s School, London SW13 9JT.

Please get in touch if you want to discuss the accessibility of resources we offer to support delivery of our qualifications: [resources.feedback@ocr.org.uk](mailto:resources.feedback@ocr.org.uk)