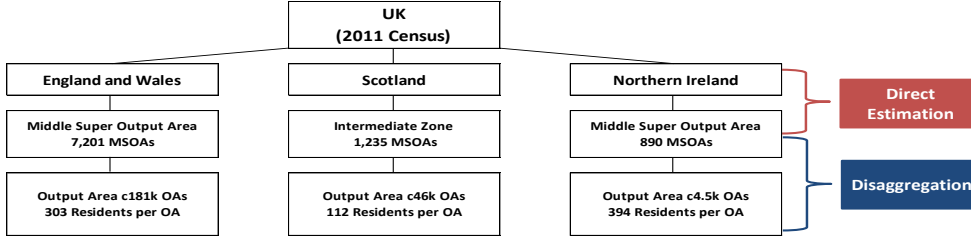


Guide	Predictor Data
<p>Summary</p>	<p>Our Predictor data sets are the low-level geographic data sets that More Metrics use to model high-level geographic data to Output Area. For example, they are used to model UK Government petition data published by parliamentary constituency to “propensity to support” models tagged at Output Area.</p> <p>Predictors are produced from Census data, the latest set being for the 2011 Census. Where appropriate they are age-standardised and geographically smoothed to impute values for nil results and to remove noise from the raw data.</p> <p>Predictors are also available for purchase by clients for use as modelling variables, or for direct application.</p>
<p>Output Areas</p>	<p>2011 Output Areas (OA11) were created for 2011 Census data, and the OA is the lowest geographical level at which census estimates are provided.</p> <p>OAs are built by ONS from clusters of adjacent unit postcodes and have similar population sizes and are as socially homogenous as possible based on tenure of household and dwelling type (homogeneity was not used as a factor in Scotland). Urban/rural mixes are avoided where possible, and they have approximately regular shapes and tend to be constrained by obvious boundaries such as major roads.</p> <p>OAs were required to have a specified minimum size to ensure the confidentiality of data.</p> <p>The total number of 2011 OAs is 171,372 for England and 10,036 for Wales and have an average population in 2011 of 309. In Scotland there are 46,351 OAs and Northern Ireland uses 4,500 small areas.</p>
<p>Predictor derivation for census variables with no age split</p>	<p>Predictors are produced by More Metrics from ONS published Census datasets. Each data set is based on relevant census counts, for example the number of cars and vans available per household.</p> <p>The refined Predictors are standardised for the age distribution in each Output Area, using the ONS population counts by Age at the time of the Census. The Predictor values are found by calculating OA Actual / OA Expected (A over E) values with the expected values calculated from the OA Age distribution. Adjustments are made for Output Areas with low population counts when calculating Predictor values. This is done to avoid problems that would otherwise arise because of nil values (and other small counts) for actuals. Adjustments routinely include nearest neighbour smoothing and adjustment of the raw A over E distributions by OA population size bands to remove heteroscedasticity.</p> <p>Predictors are tagged by Output Area which can be matched to postcode using a free file available on the More Metrics website, and in some cases by sex where ONS publish data split by sex.</p> <p>Expected values are based on an England and Wales baseline because the cross-tabbed census data by Age Band for Scotland and Ireland is often structured differently to England and Wales. The choice of baseline has no material effect on the relative Predictor values calculated across the UK</p>
<p>Predictor derivation for census variables split by age</p>	<p>Some data published at OA level by ONS have an actual age split in their structure already. Examples include the self-reported health and disability data. In these cases the A over E calculation is not required and the standardisation values are the calculated proportion of the total OA population within each ONS category. However, it is important to note that the same approach to adjustments are made for OAs with small population counts as for Predictors to deal with heteroscedasticity, so some variation will be seen to the raw proportion values that result from this.</p>

<p>How More Metrics use Predictors in disaggregation</p>	<p>Predictors and other standardised variables were developed by More Metrics to produce granular models from high-level open source data. An example is mortality modelling. The ONS publishes population and death data at Middle Super Output Area level (Intermediate Zone in Scotland), approximately 7,600 inhabitants. From this and population-level mortality tables More Metrics directly estimates relative mortality at MSA level based on an A over E calculation.</p>  <p>We then use the Predictors, which contains details on the population at Output Area standardised by age. We use this data to model relative mortality at Output Area through our own disaggregation modelling, based on standard small area estimation modelling techniques.</p>
<p>Data Sheets</p>	<p>The Predictor and other standardised variables related to a target variable each have their own data sheets that cover the following information:</p> <ul style="list-style-type: none"> • Reference data • Description • Geography • Uses • GDPR status • Source and predictor data • Keys fields • Database formats available • Example data • Example geographic spread map • Example data value spread • Notes • Copyright
<p>Client use of Predictors in data modelling</p>	<p>Clients can use Predictors as additional variables in any modelling they are doing. They have proved to add sufficient valuable to be included in a range of client models for marketing, client value, underwriting and other purposes.</p> <p>Attaching the Predictors to client data is straight-forward. The use of postcode to lookup OA values is covered in MM Guide 10001, and there is over a 99.97% match rate using postcode.</p> <p>Clients may wish to select Predictors by sex or age range depending on their data. They may also wish to exclude certain Predictors in line with legal or internal restrictions, for example ethnicity for insurance pricing models.</p> <p>In practice clients may choose not to use Predictors directly in their models, but bin OAs in, say, 10 buckets for use in modelling. Actual Predictors are provided to give full flexibility enabling clients to set their own bin ranges as required.</p> <p>Databases tagged only by OA, with other keys in the column names, are available for this use.</p>
<p>Client direct use of Predictors</p>	<p>Some Predictors may be used directly in applications. For example, number of cars or vans available to a household, or Income.</p> <p>As Predictors are modelled value these are most likely to be used directly as selectors in a marketing selection.</p> <p>Databases formatted tagged by all keys are available for this use.</p>

Predictor files			
Target variable	MM code	Additional Keys to OA11 & target	Target variable values
Age	30001		15 states 0-15 to 85+
Cars and vans	30002		3 states 0, 1 and 2+
Country of Birth	30003		12 states UK to Oceania
Communal	30004		9 states Household residents to Communal Total
Self-Reported Disability	30005	Sex, AgeBand1	3 states Lot, Little, Not Limited
Economic Activity	30006	Sex	9 states P-T Employee to Other Inactive
-	30007		
Ethnicity	30008		10 states White to Other Ethnic
Self-Reported Health	30009	Sex, AgeBand1	5 states Very Good to Very Bad
Household Ref. Person	30010		14 states Single <35 to Other communal
-	30011		
-	30012		
Industry	30013	Sex	8 states UK SIC 2007 A to F, to No Industry
Lone Parent	30014		4 states lone parent working F-T to Other
Marital Status	30015		5 states Single to Widowed
Socio-economic (NSSeC)	30016	Sex	11 states Managers large Er.s to Students
OA Class (OAC)	30017		8 states Rural Residents to Hard-Pressed Living
Std. Occ. Class (OAC)	30018	Sex	9 states Managers to Elementary Occ.s
-	30019		
Qualification	30020		5 states No qualifications to Degree-level
-	30021		
-	30022		
Persons per room	30023		3 states <0.5 to >1.5
Social-Economic Group	30024		7 states I, II, IIIN, IIIM, IV, V, U
Tenure	30025		5 states Owned outright to Living rent free
Consolidated SPR file		232,296 OA11's	x 398 variables