Technical note: How Most Similar Groups are formed

What are Most Similar Groups?

Most Similar Groups (MSGs) are groups of local areas that have been found to be the most similar to each other using statistical methods, based on demographic, economic and social characteristics which relate to crime.

Areas which have similar demographic, social and economic characteristics will generally have reasonably comparable levels of crime.

MSGs are designed to help make fairer and more meaningful comparisons between areas. Police forces operate in very different environments and face different challenges. It can be more meaningful to compare an area with other areas which share similar social and economic characteristics.

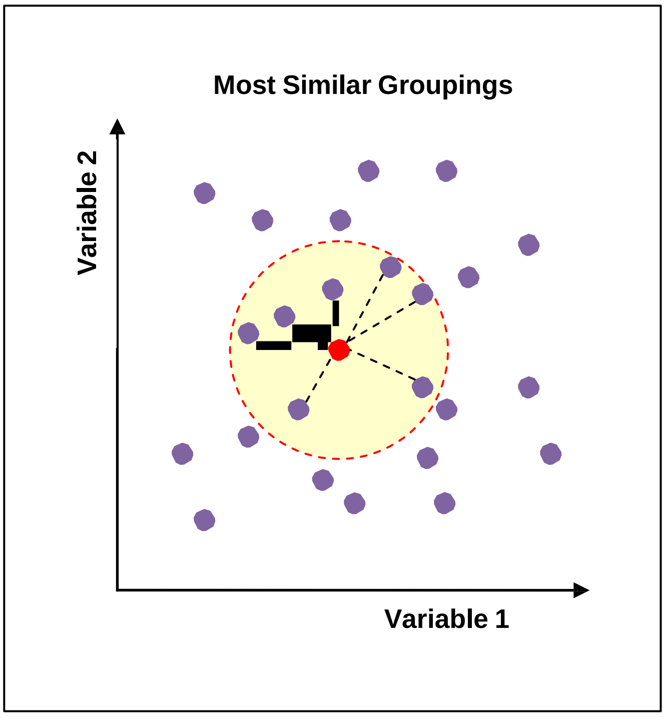
The development of the MSG approach involved stakeholders from the Home Office, Association of Chief Police Officers, HMIC and others. The current approach was chosen following advice from independent academics.

Setting Most Similar Groups

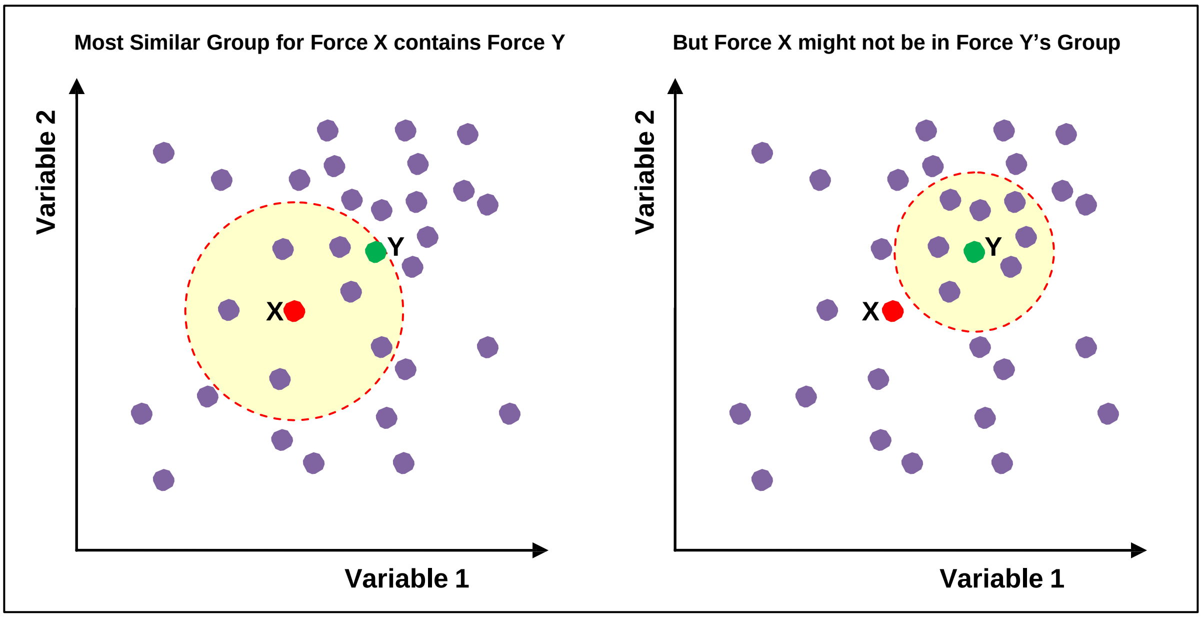
Seven variables were selected to generate MSGs (see Appendix A). They were identified by considering the levels of correlation with one or more of crime, fear of crime, or incidents. These variables are combined using a technique called Principal Component Analysis to determine two new, uncorrelated variables that best describe the variation between areas. The MSGs are determined by identifying the areas which are most similar on the basis of these new variables.

Technical detail on how MSGs are generated is provided in Appendix B. In summary, areas are compared in pairs to find the „distance‟ between them for each variable. The overall distance between the pairs of areas is then calculated by summing the squared distances for all the variables. Each area is then grouped with up to seven other areas to which it is „closest‟, based on these distances.

The picture to the right shows an example identifying the seven most similar areas to a given area based on two variables. This illustrates the method used to generate the MSGs.



Note that it is possible for area Y to be in area X's most similar group without area X appearing in area Y's group. In the graphs below the seven areas that are most similar to X are circled (left hand side), one of which is area Y. However X is not one of the seven most similar areas to Y.


Most Similar Group for Force X contains Force Y
But Force X might not be in Force Y's Group
Variable 2
Variable 1
X
Y
Variable 2
Variable 1
X
Y

How are Most Similar Group averages calculated?

The average crime rate for an MSG is the sum of the crime rates in a group divided by the total number of forces in the group. All of the forces (including the force being compared) are included in calculating the average rate for the MSG. However, the MSG comparison charts also include an upper and a lower bound line. Given the spread of crime rates for all forces in the group, the chosen force‟s crime rate would be expected to lie between these lines. More information on how this range is calculated is set out in Appendix C.

Appendix A: Variables used to calculate Most Similar Groups

Socio-economic variables were chosen based on their correlation with crime levels. The full list of the seven variables used to determine the most similar groups is given below. They were chosen based on the levels of correlation with one or more of crime, fear of crime, or incidents.

1. **Percentage of ACORN 5 households:** ACORN category 5 ("Hard Pressed" neighbourhoods). (ACORN stands for "A Classification Of Residential Neighbourhoods", and is a system for categorising areas into various types based upon census data and other information such as lifestyle surveys). It draws in part on the 2011 census data.

2. **Percentage of terraced households.** The number of terraced households divided by the total number of households (both from 2011 Census) multiplied by 100.

3. **Output Area (OA) density.** A population-weighted average of the density (population/area) of each OA. It aims to give a better indication of population density as it will highlight small pockets of densely populated housing.

4. **Percentage of overcrowded households.** From the 2011 Census. Households are classified as being overcrowded if they have an occupancy of more than 1 + number of bedrooms. This figure aims to represent the level of 'undesirable sharing' of rooms within a property.

5. **Percentage of single parent households.** From the 2011 Census, the percentage of households which contain one parent and dependent children (15 and under, or 16-18 if in full-time education).

6. **Population sparsity.** This variable gives an indication of the proportion of the population that lives in sparsely populated areas, 2011 census data is used to calculate this.

7. **Long-term unemployed per worker.** From the Office of National Statistics labour market statistics, the number of people (average of June 2010 to May 2013) claiming job seekers allowance for more than 6 months, as a percentage of the population of working age.

Appendix B: Generating Most Similar Groups – more detail

The MSG generation comprises four stages:

- Data preparation.
- Application of Principal Component Analysis to reduce the number of variables.
- Generation of the initial groupings.
- Pruning of the initial groupings to produce the final groups.

The data preparation stage involves the following steps:

1. Calculate values of input variables for each force area.

2. Transform the input variables by taking the natural logarithm.

3. Standardise the transformed variables by removing the mean and dividing by the standard deviation.

The Principal Component Analysis stage results in the number of variables being reduced from seven to two.

Generation of the initial groupings involves the following steps:

1. Calculate the „distances" between all the forces, based on the variables generated by the Principal Component Analysis. (The distances used are Euclidean distances in 2-dimensional space.)

2. For each force, work out the seven forces that are closest, based on the distances calculated in step 1.

Pruning of the initial groupings involves the following steps:

1. Calculate the standard deviation of the distance between each area and its median group member.

2. Calculate the distance between each area and the „centroid" of its MSG, for each group member, if the group members were added sequentially.

3. Remove all group members for which the distance calculated in step 2 is larger than twice the quantity calculated in step 1.

Appendix C: Calculating the red lines

The calculation of the red lines (upper and lower bounds) involves six stages:

1. Calculate the „normalised crime rate‟ for each area (force), defined as the crime rate divided by the MSG average.

2. Transform the normalised crime rates by taking the natural logarithm.

3. Calculate the mean and standard deviation of the transformed rates.

4. Use a standard result to convert this mean and standard deviation into the upper and lower quartile of the transformed rates.

5. Transform the upper and lower quartile back to the upper and lower quartile of the normalised rates by using the exponential function.

6. Multiply the resulting upper and lower quartile by the MSG average to obtain the upper and lower bounds.