Examples of automated zone design in practice

Hello I am David Martin and in this series of short videos I have been explaining the principles of automated zone design.

In this video I'm going to look at some real-world examples where automated zone design techniques have been applied in the creation of zones for which social science data are readily available.

Since the census in 2001 in England and Wales the automated zone design techniques that we've been talking about have been used to underpin the creation of quite a wide range of officials statistical areas and this commenced with 2001 census output areas and then the aggregation of those in 2004 into a set of units called lower layer super output areas, which I'll explain some more about in a minute. The whole system was revised for the 2011 Census to produce output areas and another set updated set of lower layer super output areas, and then a completely new geography for 2011 known as workplace zones and those are all available covering England and Wales. The same approach was used to create the Northern Ireland 2001 output areas and 2011 small areas and at April 2016, the time of recording, we're shortly expecting the publication of a set of UK wide workplace zones including Scotland and Northern Ireland which had been in production over the last few months.

For users of the data it's important to recognize that there are some common features to all of these geographical units and in particular the original output area creation has had quite a strong influence on all of the subsequent zones of geographies. The original idea was to separate the output areas from the geography of enumeration in 2001 and the underlying zone design is based on a set of postcode, or sometimes part-postcode polygons, which had been built up from Thiessen polygons around addresses and they include some ancillary information such as major rivers roads and railways but they're not highly detailed in the sense that they don't follow property boundaries, don't follow every minor road and so those features will not be reflected in the placement of the boundaries of the statistical units on the map. And the nature of the Thiessen polygons and the fact that these polygons have to cover the whole land surface can mean that the boundaries are quite irregular and quite spiky-looking on the map, particularly in areas where there's very little population, so in remote rural areas and these are features which are common to the entire data set.

The output areas provide the foundation for two layers, lower and middle layers, of super output areas and also the workplace zones so the output area creation came first each time around and therefore it influences the way in which all those small area census data has subsequently been aggregated and published and everything in this hierarchy nests within the local authority district boundaries which were in place at the time of their creation. So if we start briefly by thinking about those 2001 census output areas they were built from synthetic postcode polygons and they were constrained within ward and parish boundaries which were present at the time. They had a minimum population threshold of 100 people and 40 households and a target size of 125 households and the shape was controlled using an accessibility statistic which was intended to keep settlements together such as in rural areas we wouldn't split one settlement into a large area of low density land and another part of the same settlement going off into another adjacent area, but rather to keep zones tightly packed where possible around those small settlements. There is also a metric included which made it unattractive to combine urban with rural postcodes and again the goal there was to provide a clear demarcation of the urban areas

Homogeneity in this original design was controlled using an intra area correlation statistic and it was based on the tenure and the dwelling type classifications from 2001 Census and the whole zonation was implemented after the census had been collected and before the data were published and it meant that there were no zones in the output which had to be suppressed because the populations were too small. Effectively every zone was designed to meet census confidentiality criteria. What happened in 2011 with the boundaries which are currently available is that an updated set of the areas was created and that included a review process of the previous boundaries and some manual adjustments if the 2001 boundaries were found to be unsatisfactory in very specific circumstances, and it was also necessary to incorporate some imposed local authority boundary changes where the boundaries had themselves broken the original output area geography. But only 2.6 per cent of the output areas have actually changed between the two censuses. So they have a very strong degree of continuity. Where change has occurred it's mostly been due to the need to subdivide zones where the populations become too large and too large in this sense means over 625, mostly due to population growth in the decade between the censuses. And there would in some instances be aggregation of zones whether populations become too small and that's much rarer. The same automated zone design procedures were re-applied in the situations where subdivision was required and that was done using the AZTool Software which has been described in a previous video. One of the 2001 constraints to the ward and parish boundaries was dropped in 2011 in order to maintain a new statistical geography system and that results in a 171,372 of these output areas.

One further point of explanation is that the lower layer super output areas which were originally created for the government's indices of deprivation in 2004 are the result of taking the output areas and putting them back into the beginning of the zone design process and then running the whole process again so the output areas become the building blocks in this second zone design problem. And here the thresholds are larger they are 1000 persons or 400 households because the data which should be made available are potentially much more detailed and sensitive so the confidentiality threshold is higher. The 2001 output areas were the building blocks for the 2001 super output areas, which actually were published in 2004 and then in 2011 the 2011 output areas the building blocks for the new 2011 super output areas. In each case there are constrained within local authority boundaries and typically the original populations had a mean of around 1,500 so they would often comprise five output areas combined. There were 34,753 in England and Wales in 2001 and then in the new zonation that's gone down slightly to 32,844 in 2011. Users of the statistics may note that there is also a zone called a middle layer super output area that's an aggregation of these lower layers but it's not been produced as a result of the direct automated process. It was much more process of consultation which led to the creation of those zones.

So in the map that we have here were seeing for an area of East London the super output areas from 2011. The statistic which is being mapped is actually a shaded set of deciles from the indices of deprivation, the 2015 version. We can now superimpose on there the output area boundaries which allows you to see the difference in scale between the very detailed output areas and the less detailed super output areas lower layer.

The same approach has been used in quite different way to create workplace zones so following the 2011 Census the Office for National Statistics ONS recognized the demand for detailed data on place of work and the workforce which not really well suited by output areas based on residential geographies and so, using the information which comes from people's

answers to the census questions about where they work, the same data have been completely reconfigured to create a set of workplace zones and these have a minimum of 200 workers and three workplace postcodes as their basic population thresholds. In some cases they'll be the same as the 2011 output areas if there are neighbourhoods where similar numbers of people live and work, but it's very often the case that in residential areas the workplace zones are aggregations of output areas and in city centres and business districts where many people are working they are subdivisions of the output areas and so we're adjusting the trade-off between number of statistical units that are present, in this case the workers, and the size of the zones. In each case if subdivision is needed to be performed it's based on the same automated zone design process using synthetic postcode polygons and these are the postcodes which were described as workplaces in the census.

The workplace zone geography is understandably a slightly larger set of zones than the output areas and there's 53,000 or so of them in England and Wales. In this map from the Datashine website http://datashine.org.uk we see the same area that we've used in the previous illustrations but this time for the workplace zones what's notable here is that unlike the output areas the smallest zones are those over on the left hand side in the west of the map and City of London because very large numbers of people are working in very high density small areas, so we have a fine subdivision of the geographical space, and across to the right in the east which are mostly residential areas and some low density industry, and zones are much larger because there are smaller numbers of people working in those areas. This particular map shows a classification of those workplace zones into different industry and worker types.

So in summary we see that the automated design processes which we've been talking around in these videos have underpinned all of the census zones used in England and Wales in 2001 and 2011 and also appear in some of the census outputs in Northern Ireland and Scotland particularly the workplace geographies for 2011. So all research users of the small area census data need to understand the implications of those zone design processes on the data which they're using. In particular we've seen that the design has a strong bearing on the relationship to other geographical units. These zones will all match perfectly to local government geography. They will match quite closely to the small area postal geography, but there may be other units of interest for research purposes, for example electoral geographies, to which this will not have a very close association, at least not in detail. The exact placement of those boundaries and the shapes which we see within them are going to be an artefact of the road centrelines, railways, the rivers, the features which have been included in the building blocks geographies and that explains some of the patterns which we see in small areas census mapping

Its the relationship between the zones and the underlying population geography which determines the size and shapes of the zones. They would be very much larger in rural areas because they're aimed to target the same population size so we've got to go much further to find the same number of people. In really understanding what the implications would be for any statistical analysis of the census aggregate data, and for investigation of ecological relationships in those data, the researcher needs to be aware of the thresholds that have been used to preserve confidentiality, the size ranges which were allowable and the way in which social homogeneity, based on tenure and dwelling type, has been used to influence the zone sizes and shapes because that may well have a bearing on any analyses which are going to use those same variables.

Finally we might note that researchers who are wanting to use these data could consider

running alternative zone designs at the same scales to understand to what extent the relationships which they see in their data would be preserved under different configurations, effectively different aggregations at the same scale.