Introduction to automated zone design

Hello I am David Martin and in this series of short videos I'm going to be providing an overview of some of the methods that we can use for automated zone design.

In this first video I'm going to review what exactly we mean by zones and the ways in which the design can be important in research and in fact in daily life. So what do I mean when I'm talking about zones? In this context zones are divisions of geographical space. Geographers would usually define them and think of them as polygons, but to many social scientists they may just be shaded areas on the map and usually we would consider that one zone is represented by a single polygon, although if there is an area with islands we might find that one zone is represented by more than one polygon. Lots and lots of examples that different people will be familiar with: we're used to seeing regions, counties, local authorities, wards, electoral districts, and in particular countries we might think of the states of the United States, communes in France and mesh blocks used for official statistics in Australia, and in the UK particularly things like postcode sectors and census output areas.

So to take one of those in a little more detail and try to understand how they've been constructed we will think for a moment about the 2011 census output areas in England and Wales. It's worth reviewing the basic characteristics because they tell us something about the way in which the zones have been created. The mean population size is three hundred and twenty-five people but importantly they must always have more than a hundred people and more than 40 households to preserve confidentiality and many of them are in fact based on the output areas from the previous census in 2001. There are lots of different constraints and considerations which are brought to bear on how those zones are placed on the ground and that would have included attempting to match as closely as possible to the unit postcodes, which is the smallest units in the postal geography, attempting to control for the shape of those zones so that they are not too irregular on the map and trying to observe social homogeneity so that were dealing with zones which try to keep neighbourhoods with similar social characteristics together, and it's important to realize the purpose of the zonation. In this case these are zones which were used for the publication of small area census statistics and so the placement of zone boundaries in this instance is a result of a set of a very explicit design procedures. If we take a look at a map such as this map of zones in East London these are the census output areas from 2011 and in this particular map I'm mapping some census data. We're using a census variable around dependent children in households, and essentially it can be thought of as a map where the red areas are the family areas, over in the east, and the city centre where there are very few children living are the green areas over on the left, in the west, of this map. And if we just reflect on the structure of that map it's apparent that the zones are very varying sizes that's because those populations have been kept very similar round about three hundred and twenty-five people but the underlying population distribution is very uneven and therefore the geographical sizes vary widely and that's a very important basic principle of zone design in most social science applications. If the geographical sizes were kept similar the populations themselves would vary very very widely and there may be very good reasons why we want to control that trade-off between size and population.

With this map which is from the brilliant Datashine website we can understand a little more by seeing the shading just on the built-up areas, but we've added some extra features to the map for example the main roads, railways and the River Thames through the centre of London which is now apparent. The pattern of the mapped zones reflects the underlying landscape features so we can see the river and the principal roads have influenced the design

of the zones and I'm highlighting here two particular features: there's a very characteristic U-shaped bend in the river Thames by London's Docklands and also down in south east of this map there's a straight line which is actually the route of the Roman Watling Street from Dover to London and so, intriguingly, in this very modern design from the 2011 census we see that physical geography features and historical cultural features have actually worked their way through into the algorithm which has placed those zones on the map.

So we're talking about zone design as a process. What exactly do I mean? In this case the placement of those boundaries is the result of a carefully worked-out design procedure and if the zones are going to be used for a purpose such as enclosing statistical units, they might be addresses, members of the population – which is what we're doing in the census situation, then effectively the placement of those zones determine which groups of units are going to be aggregated together. It's got particularly significant implications for statistical disclosure control because individuals with distinctive characteristics may be quite identifiable in small populations. So the zoning system which we have at the end may result, for different purposes, from a combination of historical, administrative or explicit statistical design criteria. And it might be the result of a very careful consideration, or in some cases we may find that were dealing with zones the placement of which is the result of a quite arbitrary process.

Why does it matter? Well, depending on the purpose to which the zonation is going to be put, size and position of boundaries matters in many different ways and geographers know of this as the Modifiable Areal Unit Problem. That's the phrase originally coined by a geographer called Stan Openshaw in 1984 and it's conventionally divided into scale and aggregation problems. Now although that term may be familiar primarily only to geographers the same phenomenon is quite widely recognized in the manipulation of electoral boundaries which we call gerrymandering, so let's take a look at that example.

In my Toytown example, a vote's to be held in 35 neighbourhoods – in this case they're square neighbourhoods of equal sizes – and the result of the vote was that 15 neighbourhoods vote green and 20 vote blue. If we were to organize those neighbourhoods into constituencies perhaps to return candidates to a senate or parliament, in this particular configuration of five constituencies, blue wins all five because blue is in the majority in all five of those constituencies. But if we were to redraw the boundaries of five constituencies, in this configuration we've almost reversed the result because green wins four and blue only wins one and that's the result of having concentrated blue's core area into a single zone such that in the other four zones on the left in this illustration green is in the majority.

We can also change the number of zones in this case three where Green wins two and blue wins one, or we could even come up with a configuration in this case with seven constituencies which exactly reflects the proportion of the vote at the neighbourhood level here green wins three blue wins four. So the decision as to whether there should be three, five, seven or any other number of constituencies is known as the scale problem and that refers to the size of the zones and how many of them were going to have. But at a given scale there are multiple possible configurations and this is known as the aggregation problem: it's the configuration of the boundaries at a particular scale. So the familiar gerrymandering scenario is where those boundaries are manipulated in order to produce a particular statistical result in this case a bias in favour of one party or another. And certainly in a UK context in the press the concept of a 'postcode lottery' is frequently cited which refers to people having entitlement to use of local services just because of the way in which a set of arbitrary boundaries, by reference their postcodes, have been placed on the map. People living in one

side of the street might be able to access services from one authority which are denied to those somewhere else. So we see that the zonation can have very important real-world consequences. Whether the design of zones is actually a problem depends on the intended purpose.

So one final area that I'd like to highlight is that the zonation can also have an impact on the statistical relationships which we see in data and this is perhaps the biggest implication for social science research. The way in which counts are grouped into zones will affect any kind of ecological associations and any kind of ecological analysis which we want to perform on those data, because different relationships are likely to hold different geographical scales and we may also describe them differently when we re-aggregate at the same scale. So in this illustration which is from some work that I did myself with my colleague Samantha Cockings, nearly a decade ago now, we see the relationship between four census variables and a deprivation index and long-term limiting illness as measured by the census. On the x-axis we've got population size and we re-aggregated the same census data (it was for the then County of Avon) into many different zonations with different average zone sizes. On the y axis we've got the correlation with long-term limiting illness so if we were to look for example at the very lower left corner of this diagram where the overcrowding line comes down to the lowest point represented by the line with the square boxes, we see that when the zone size on average is only about four hundred people the correlation with long-term limiting illness is less than 0.5 but as we redraw the boundaries to create zones with successively larger populations that correlation increases by the time zone size is perhaps around about two and a half thousand in the middle of the diagram if you trace the line then the correlation has risen to about 0.8 and then it plateaus it doesn't increase greatly with increasing zone size.

A very similar patterns can be seen for the other individual variables in this case unemployment, no car ownership, not living in owner-occupied accommodation and overcrowding, and also for the overall deprivation score which is a Townsend score – it's a multivariate index built from those variables. So the conclusion we can draw from that would be the researcher who only came along and looked at one particular aggregation scale, one set of zones, would actually understand the relationship only at one point on that curve and the importance is to recognize that there is a dependence between what we see in aggregate data and the way in which the zones are being constructed. This second diagram from the same study shows one further point of interest. So here we've run multiple aggregations at the same scale so around about 750 people we've redrawn that population re-aggregated them into ten different zonations, and even at the same scale here we see that there is a spread in those correlations – not a very large one but there's a distinct range of values. And so when we go to any particular aggregation scale the value that we get is not a central value necessarily for that scale but is one drawn from a distribution, and as we can see here with successively increasing population sizes that distribution spread out a little way.

So in summary, the zones which we encounter in social science are all around us for many different purposes. They may be used for statistical analysis, they might have policy implications and the specific design and placement of those boundaries could have very big impacts both on research findings and interpretation and on everyday life. So really all researchers who want to use zone-based data need to understand the methods by which the zones have been created and, where appropriate, researchers should consider perhaps designing their own zones which are designed and optimized for the purpose of the research. And we should also note that we've seen that the zone design has particular significance in

ensuring the confidentiality of data that are being aggregated over geographical areas. And so in the next video we will take a look at some of the principles we can apply to the zone design process.