Cumulative Curves for Exploration of Demographic Data: a Case Study of Northwest England

Natalia Andrienko and Gennady Andrienko

Fraunhofer AIS - Autonomous Intelligent Systems Institute,
SPADE – Spatial Decision Support Team,
Schloss Birlinghoven, Sankt Augustin, D-53754, Germany

Summary

The paper introduces the idea of generalising a cumulative frequency curve to show arbitrary cumulative counts. For example, in demographic studies generalised cumulative curves can represent the distribution of population or area. Generalised cumulative curves can be a valuable instrument for exploratory data analysis. The use of cumulative curves in an investigation of population statistics in Northwest England allowed us to discover interesting facts about relationships between the distribution of national minorities and the degree of deprivation. We detected that, while high concentration of national minorities occurs, in general, in underprivileged districts, there are some differences related to the origin of the minorities. The paper sets the applicability conditions for generalised cumulative curves and compares them with other graphical tools for exploratory data analysis.

Keywords: Geographic visualisation, Exploratory data analysis, Data visualisation, Interactive graphics.

1 Introduction

Exploratory data analysis (Tukey 1977) is the process of examining a data set with an objective to detect significant, previously unknown patterns and regularities. This process usually involves visualisation of the data on various graphical displays. For this reason the exploratory data analysis is sometimes also called "visual data mining". Very often an analyst builds multiple displays for viewing the data from various perspectives. Geographically referenced data are visualised on thematic maps – this gives an opportunity to explore their spatial distribution and detect relationships between the spatial variation of characteristics and various geographical phenomena and features, both natural (relief, land cover, climate, etc.) and related to people's activities (roads, land use, etc.). Combination of maps with other types of graphs provides conditions for more comprehensive analysis and gaining additional insights into inherent data characteristics.

In this paper we describe a case study of exploring geographically referenced demographic data concerning Northwest England using an interactive map and a cumulative curve display, an original data visualisation tool we have recently developed. First of all, we briefly describe the data set. Then we introduce the idea of the cumulative curve display. Besides having interesting properties that make it useful for data exploration, the display has excellent scalability characteristics and can be applied even to very large data sets. Next, we show how we arrived at interesting, previously unknown facts concerning relationships between some population characteristics and indices representing deprivation. Finally, we compare the cumulative plot display with other statistical graphics and interactive visualisation tools for exploring relationships between attributes.

2 Northwest England Dataset

The dataset with census data of Northwest England, on the ward level of administrative division, was provided to us by MIMAS (Manchester Information & Associated Services, University of Manchester, http://www.mimas.ac.uk/) within the EU-funded project SPIN! – Spatial Mining of Data of Public Interest (IST Program, project No. IST-1999-10536). The dataset contains values of various demographic attributes referring to 1011 spatial objects (wards): population number in total and in different groups by gender, nationality, education, employment, etc., number of households in total and by particular classes (e.g. households with no car), and so on. Besides these counts, there are values of four so-called deprivation indices: DoE (Department of the Environment's Index of Local Conditions), Jarman Underprivileged Area Score,

Carstairs Deprivation Score, and Townsend Material Deprivation Score. Each index is calculated in its own way from a different set of original census attributes. An overview of the commonly used deprivation indices can be found at http://www.swpho.org.uk/pat18discuss.htm (Bunting 2000).

The data provider expected that, after exploring the data set using interactive visualisation tools we had developed, in particular, within the SPIN! project, we would be able to characterise wards with high deprivation in terms of the

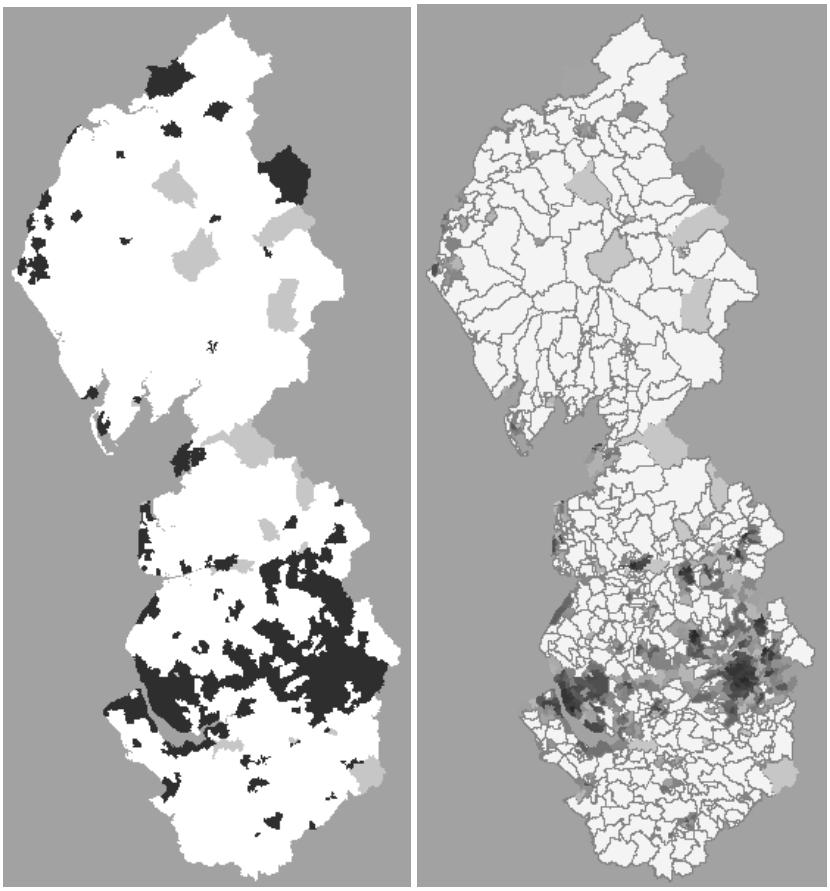

Figure 1. The map on the left shows wards with positive values of the Doe deprivation index by dark colouring. On the right, these wards are painted in different shades according to the values of the Doe index. Darker shades correspond to higher values.

demographic attributes available. Of course, this did not relate to the attributes used in the computation of the deprivation scores.

We started our analysis with comparing values of the four indices and found that they are highly correlated: the pairwise computed correlation coefficients ranged between 0.899 (DoE vs. Carstairs) and 0.960 (Townsend vs. Carstairs). Hence, it was reasonable to focus on a single deprivation index rather than to try to analyse all of them. We selected the DoE index since we were informed that it was the most often used in demographic studies. The index comprises seven census-based variables: unemployment, children in low earning households, overcrowding, housing lacking basic amenities, lack of car ownership, children in 'unsuitable' accommodation, educational participation at age 17. Scores greater then zero indicate greater levels of material deprivation.

The map in Figure 1, left, shows the spatial distribution of values of the DoE index. The wards with positive values of the index are painted in dark and those with negative values are light. The map on the right represents only positive values by degrees of darkness: the higher the value, the darker the colour.

By visual examination of the map, one can detect two spatial clusters of high deprivation scores on the south of the studied area. They correspond to big cities Liverpool (on the west) and Manchester (on the east). In general, one can observe that high values are mostly associated with smaller wards that are located in urban areas and, hence, can be expected to have quite high population density.

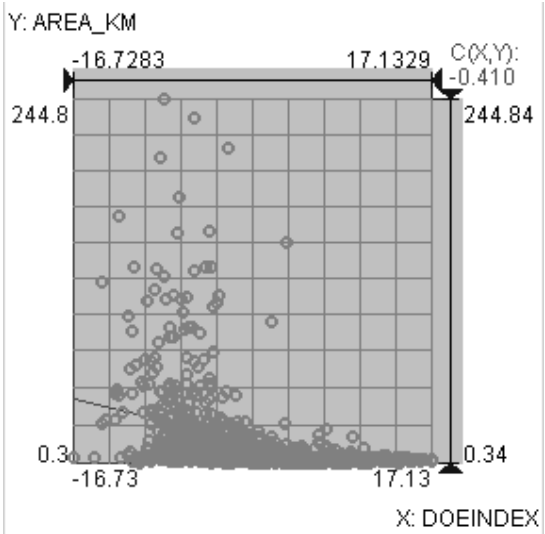

Figure 2. Wards with big areas have low values of the Doe index.

In order to check this statistically, we visualised the values of the DoE index and areas of the wards on a scatterplot display that shows also the correlation coefficient for this pair of attributes (Figure 2). It may be seen from the plot that, indeed, big areas mostly correspond to negative values of the DoE index while the highest deprivation scores correspond to small areas. At the same time, the correlation coefficient (-0.410) does not indicate a strong dependency.

This situation shows us that one should not rely in data analysis solely on computed figures but try to gain additional information from graphical representations of the data. A scatterplot is, possibly, the simplest but, nevertheless, in many cases a very effective representation supporting investigation of relationships between attributes. However, a scatterplot is seldom suitable for attributes representing absolute counts. For analysing such attributes, we have devised the cumulative curve display.

3 Cumulative Curve Display

Cumulative frequency curve, or ogive, is one of well-known methods for graphical representation of statistical distribution of attribute values. In such a

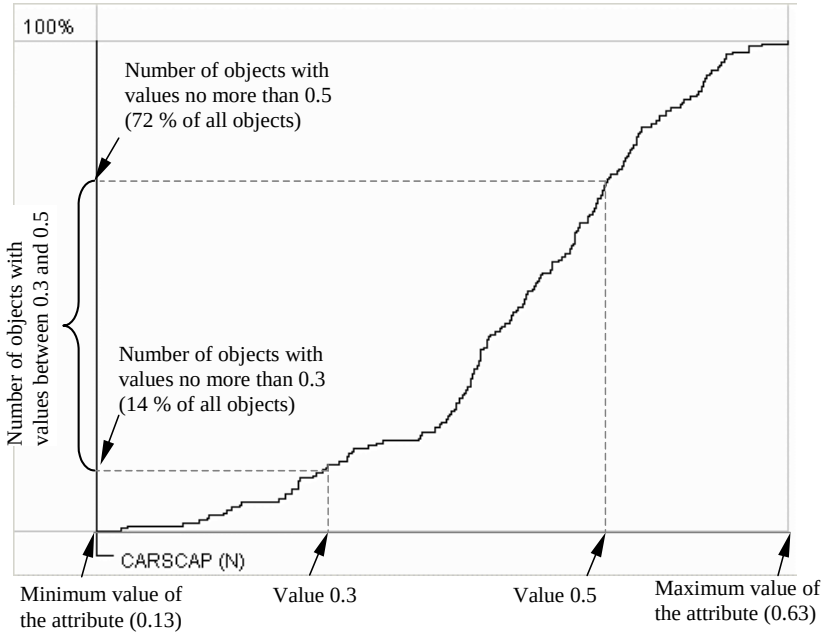

Figure 3. How a cumulative frequency curve is constructed.

graph, the horizontal axis represents the value range of an attribute. The vertical position of each point of the curve corresponds to the number of objects with values of the attribute being less than or equal to the value represented by the horizontal position of this point. The method of construction of an ogive is demonstrated in Figure 3. Here, the ogive represents the distribution of values of the attribute CARSCAP (number of cars per capita) related to districts of some territory.

Peculiarities of value distribution can be perceived from the shape of the ogive. Steep segments correspond to clusters of close values. The height of such a segment shows the number of the close values. Horizontal segments correspond to "natural breaks" in the sequence of values.

We found it possible to generalise the idea of cumulative frequency curve and to build similar graphs summarising values of arbitrary quantitative attributes (counts). Examples of such attributes are area, population number, gross domestic product, number of households, etc. A generalised cumulative curve (GCC) is built as is demonstrated in Figure 4 by the example of a cumulative population curve. The same objects (districts) and the same attribute (CARSCAP) are used as in Figure 3. The horizontal axis of the graph corresponds, as before, to the attribute CARSCAP (we shall use the term "base attribute" to denote the attribute represented by the horizontal axis). For each position *x* on the horizontal axis, the corresponding vertical position is obtained by counting the total population in all districts having no more than *x* cars per capita. The resulting sequence of points is the cumulative population curve with respect to the attribute CARSCAP. Figure 4 shows the cumulative population curve along with the cumulative frequency curve

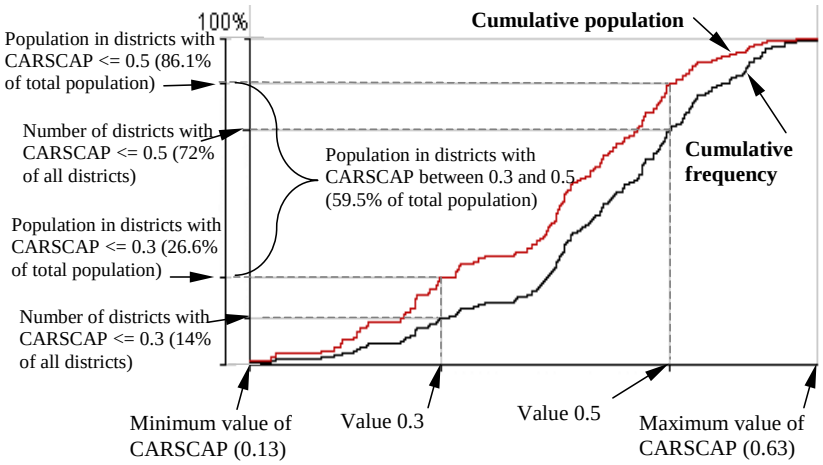

Figure 4. Construction of the cumulative population curve for districts ordered by the number of cars per capita (attribute CARSCAP).

for the attribute CARSCAP. It may be noticed that both curves have similar shapes, but the population curve rises a bit steeper for low values of the attribute CARSCAP while at the right end it is more gradual than the frequency curve. This indicates that the population is not distributed evenly among the districts: the districts with fewer cars per capita are more populated than the districts where the number of cars per capita is high.

Another popular way to represent a distribution is frequency histogram, which can also be generalised in the same way as we did for cumulative frequency curve. However, unlike a cumulative frequency curve, a histogram requires a prior division of the value range of the base attribute into intervals. Such a division leads to substantial loss of information because it hides the distribution of values within each interval. At the same time, a cumulative curve graph *can* represent division into intervals by means of additional graphical elements. Thus, the horizontal axis of the graph may be suited to show interval breaks. In our implementation of the cumulative curve display (Figure 5), we use for this purpose segmented bars with segments representing the intervals. The positions of the breaks are projected onto the curve, and the corresponding points of the curve are, in their turn, projected onto the vertical axis. The division of the vertical axis is also shown with the use of coloured segmented bars. By construction, the lengths of the segments are proportional to the numbers of objects with attribute values fitting in the respective intervals. Hence, it becomes easy to compare the sizes of the object subsets (classes) corresponding to the intervals. For example, the interval breaks shown in Figure 5 divide the whole set of objects into 3 classes of approximately equal size that is demonstrated by the equal lengths of the bar segments on the vertical axis.

The cumulative curve display as we designed it allows the user to add a GCC for any quantitative attribute to the cumulative frequency curve. The curves are overlaid, i.e. drawn in the same panel (see Figure 6). This is possible since they share the same base attribute. To be easier distinguished, the curves differ in colour. The horizontal axis is common for all of them. The vertical axes are shown beside each other on the left of the graph. Each of the vertical axes is divided into the same number of segments, but the positions of the breaks are, in general, different. Thus, Figure 5 shows us that 33.5% of all districts fit in the class corresponding to the first interval (from 0.134 to 0.405). They occupy only 9.6% of the total area but contain 48.6% of the total population living on the territory.

As may be seen from this example, GCCs may be used for exploring relationships between several attributes. In particular, they allow one to see the distribution of population with respect to various indices characterising districts of territory division. This makes them especially suitable for demographic studies, such as the exploration of the deprivation in Northwest England.

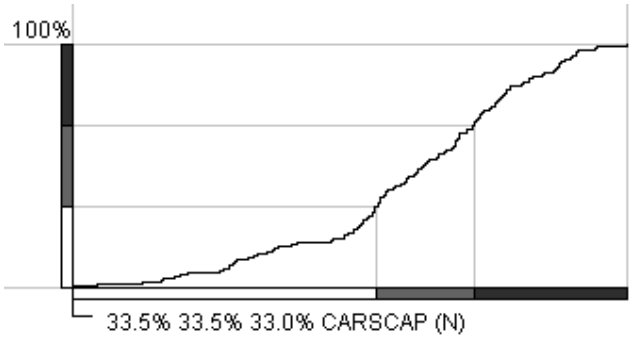

Figure 5. Values of the attribute CARSCAP are divided by 2 breaks (0.405 and 0.492) into 3 intervals. As a result, the objects have been divided into 3 classes of approximately equal sizes (33.5%, 33.5%, and 33.0% of the whole set). The segmentation of the horizontal axis shows the positions of the interval breaks, and that of the vertical axis shows the corresponding division of the objects into classes.

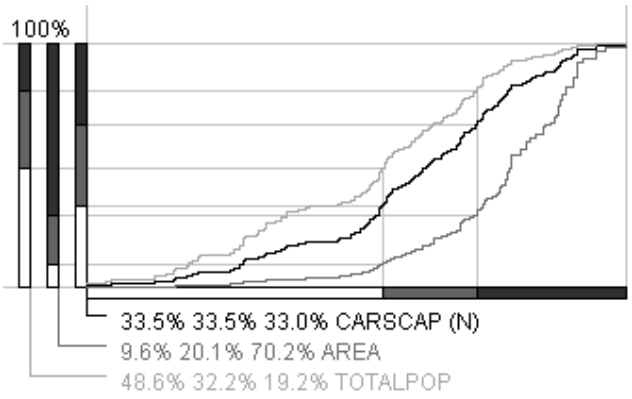

Figure 6. Generalised cumulative curves are built for the attributes AREA and TOTALPOP (total population) using CARSCAP as the base attribute.

By construction, GCC is similar to Lorenz curve (Schmid and Schmid 1979; first described in Lorenz 1905), which is used mainly in economic studies. Like GCC, Lorenz curve is also applied to quantitative attributes (counts), most often to attributes representing incomes. The method of the construction of a Lorenz curve is demonstrated in Figure 7.

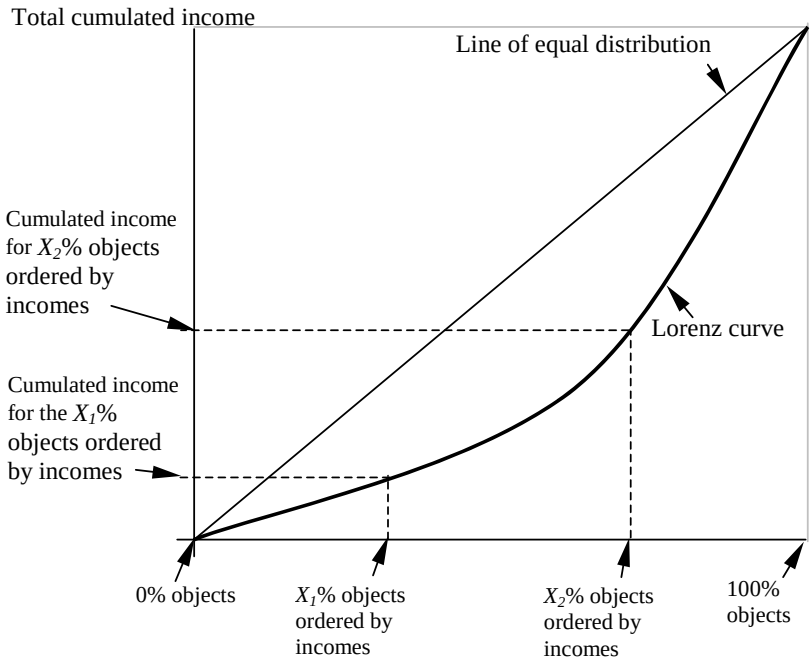

Figure 7. Construction of a Lorenz curve.

For building a Lorenz curve, the objects under investigation are ordered by the values of the studied attribute (in our example income). Then, for each position *x* on the horizontal axis the corresponding vertical position is the cumulative income of the first *x*% objects in this row. Hence, like in the case of GCC, Lorenz curve involves the accumulation of values of some attribute. However, the difference is that the latter requires the objects being ordered by values of the same attribute that is used for accumulation while the former can be based on any numeric attribute. Furthermore, a Lorenz curve is a mapping (in the mathematical sense) between cumulative amounts and the corresponding *numbers of objects* whereas a GCC is a mapping between cumulative amounts and the *values of the base attribute*. This property makes a GCC suitable for exploring relationships between the cumulated attribute and the base attribute, while a Lorenz curve is helpful for examining the distribution of a single quantitative attribute, more specifically, for estimating how far this distribution is from the equal distribution (see the straight line in Figure 7).

For both GCC and Lorenz curve, it is possible to put several curves on the same display. Again, these two kinds of display support different exploratory tasks. Two

or more Lorenz curves are used for comparing distributions, for example, the distribution of income at different time moments (see Schmid and Schmid 1979, p.138) or in different countries. Two or more GCCs allow one to see whether attributes B, C, … are related to the base attribute A similarly or differently. Let us demonstrate the use of this property in exploratory data analysis by the example of the investigation of the deprivation in Northwest England.

4 Exploration of Deprivation in Northwest England

The cumulative curve display in Figure 8 (in the panel right of the map) shows the statistical distribution of the values of the DoE index over the set of wards of Northwest England. The range of values of the DoE index is divided by breaks –3 and 3 into three subintervals: low deprivation (below –3), medium (from –3 to 3), and high deprivation (over 3). Accordingly, the wards are grouped into three classes. The spatial distribution of the classes can be seen on the map. The wards with high deprivation are clustered at the locations of Liverpool (south-west) and Manchester (south-east). These clusters are mostly surrounded by districts with medium deprivation, although there is a "belt" of wards with lower deprivation on the north of the western cluster. However, in the northern part of the territory, which contains mostly wards with low deprivation, there are some spatial outliers, i.e. wards with medium and high deprivation scores.

From the segmentation of the vertical axis of the cumulative curve display and from the figures below the graph one can see the distribution of the wards over the classes: 32.4% of the wards fit in the class with low deprivation, 28.5% in the class with medium deprivation, and 36.9% in that with high deprivation (for the remaining 2.2% wards the values of the DoE index are missing).

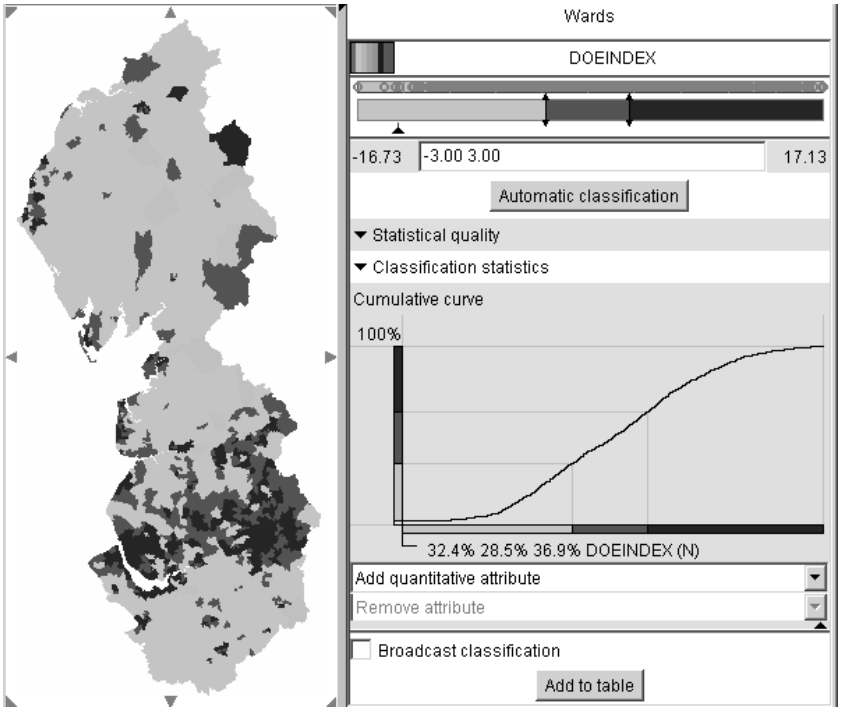

Figure 8. The cumulative frequency curve (on the right) shows the statistical distribution of values of the DoE index. On the map (right), the wards are classified according to the values of the DoE index into classes with low deprivation (below –3), medium (from –3 to 3), and high (over 3). The classification intervals are represented on the cumulative curve display by segmentation of the horizontal axis. The segmentation of the vertical axis shows the distribution of the total number of wards among the classes.

Let us now add cumulative curves for areas and population to the cumulative curve display. The result is presented in Figure 9. One can see that the shape of the cumulative population curve is very similar to that of the cumulative frequency curve while the cumulative area curve looks rather differently: it goes up much faster than the other two curves for low values of the DoE index, especially on the interval approximately between –10 and -3. The distribution of the total area among the classes of wards is the following: 68.7% of the total area of Northwest England is occupied by wards with low deprivation, 16.3% by those with medium deprivation, and 9.8% by wards with high deprivation (the remaining 5.2% of the total area belong to the districts with missing values of the DoE index). The distribution of the population is 21.3%, 30.2%, and 47.7%, respectively (and 0.8%

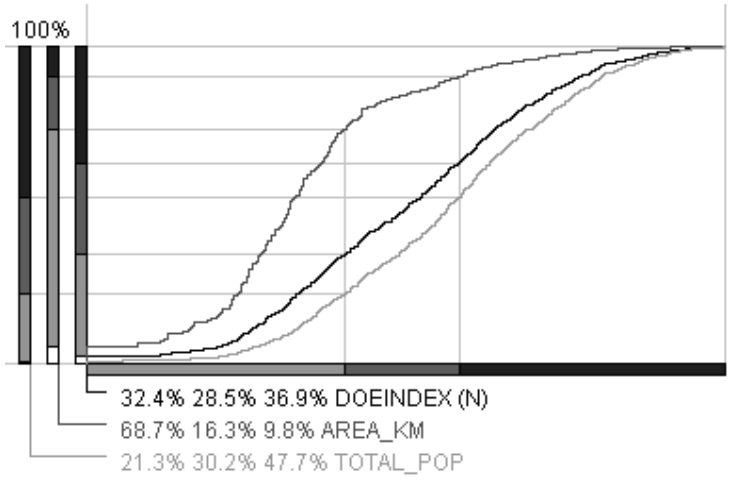

Figure 9. The distribution of the total area and population of North-West England in relation to the values of the DoE deprivation index.

in the districts with missing DoE values). Hence, almost a half of the total population lives in districts with high deprivation, which constitute only 9.8% of the whole area. Thus, the cumulative curve display has allowed us to confirm the impression we got when we looked at the map in Figure 1, that is, that high deprivation scores mostly occur in smaller wards having higher population density.

Let us see what is the relationship between the deprivation and the distribution of national minorities over the territory. For this purpose we select the attribute TOTAL_minorities (total number of people originating from foreign countries) for the representation on the cumulative curve display and compare the shape of the curve for this attribute with that of the curve for the total population (Figure 10). We see that the curve for the national minorities grows much more gradually than the curve for the total population on the intervals with low and medium values of DoE index and is very steep at the end, i.e. where the deprivation scores are the highest. Below the graph we see how the total number of the national minorities is distributed among the deprivation classes: while only 6.3% and 13.0% of people with foreign roots live in the wards with low and medium deprivation, respectively, 80.5% fit in the class with high deprivation. These figures become especially striking in comparison with the distribution of the total population (21.3%, 30.2%, and 47.7%, respectively).

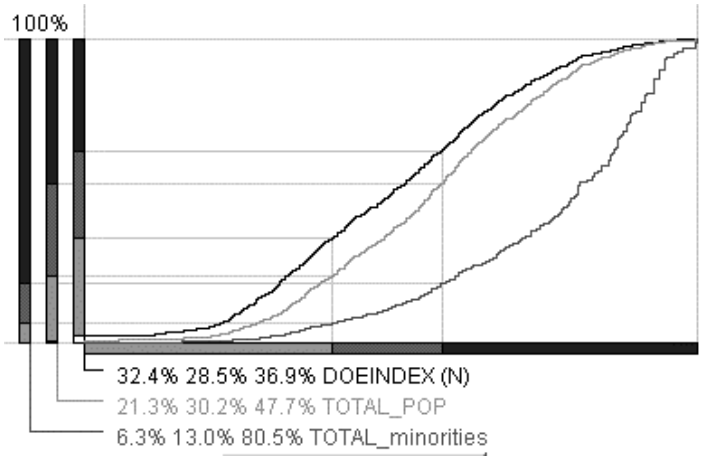

Figure 10. The distribution of national minorities in comparison to the distribution of the whole population.

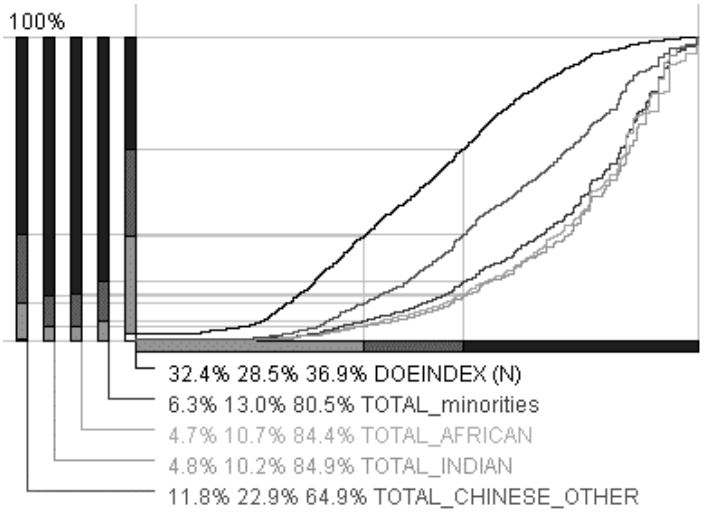

Figure 11. The distribution of people originating from Africa and India is close to that for all national minorities, while Chinese and other nationalities are distributed differently in relation to the deprivation.

It can be anticipated that the relationship to the deprivation is not the same for minorities with different origins. In order to detect the differences, we add the attributes TOTAL_AFRICAN (people originating from Africa),

TOTAL_INDIAN (people originating from India), and TOTAL_CHINESE_OTHER (people of Chinese and other origin) to the cumulative curve display and compare the curves built to the curve for the total minorities (Figure 11). For simplification, we removed the curve for the total population.

We see that the curves for African and Indian population almost completely repeat the shape of the curve for all minorities. The figures of distribution among the classes are also close for these three attributes. For the population with the Chinese and other origin, however, the situation is quite different. The curve grows slowly on the interval with very low values of the DoE index but then the growth is almost even on the remaining part of the value range. There is no such steep increase in the area of the highest values of the DoE index as for the other national minorities. The distribution among the classes is also notably different from that for the other minorities: 11.8% in the class with low deprivation, 22.9% in the class with medium deprivation, and 64.9% in the class with high deprivation.

Hence, using the cumulative curve display, we have uncovered interesting facts about the relationships between the level of deprivation and the distribution of national minorities across the territory of Northwest England. In particular, the cumulative curve display exposed the dissimilarity in how the minorities with different origins (African and Indian, on the one hand, and Chinese, on the other hand) are related to the deprivation.

5 Discussion

In our study, we used GCCs for exploring relationships between attributes. It is appropriate to compare this technique with other visualisation methods typically used for the same purposes. One of such methods is scatterplot. In our study, scatterplots occurred to be not very useful due to the peculiar feature of the distribution of the national minorities over the territory: the proportion of the national minorities is relatively low almost everywhere except for a few districts with much higher values, which are in this case statistical outliers. Thus, in Figure 12 the counts for the different minorities (from left to right: all, Indian, African, and Chinese) are plotted against the corresponding values of the DoE index. In principle, it is possible to detect certain differences between the scatterplots, but they are less clearly visible than on the cumulative curve display in Figure 11. The logarithmic transformation of the scatterplots (Figure 13) does not help much due to the overplotting. This is a serious problem that makes scatterplots inappropriate for large datasets. On the opposite, generalised cumulative curve has a very important property of being scalable. This method of data visualisation is independent of the number of objects it is applied to; it is equally applicable to

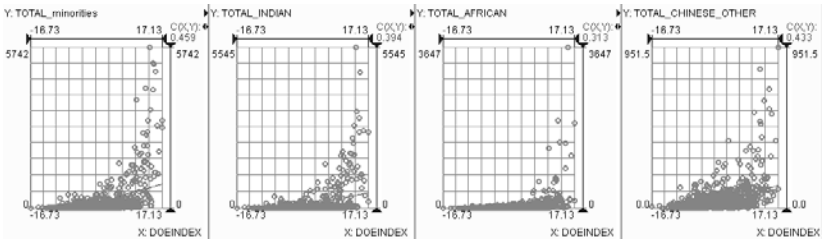

Figure 12. Due to a particular distribution of the national minorities over the territory, scatter plots occur to be inconvenient for exploring the relationship to the deprivation.

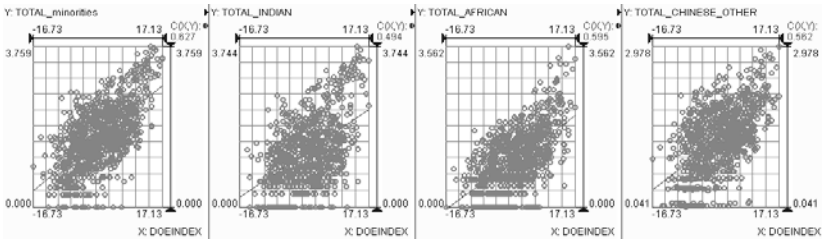

Figure 13. The result of applying a logarithmic transformation to the scatterplots from Figure 12.

twenty or to a million objects. A large number of objects is not even a serious obstacle to the display interactivity. Certainly, the initial sorting takes relatively long time, but after that, when the user changes the class breaks, the sizes of the segments on the axes can be recomputed and redrawn very efficiently.

Let us compare GCC to other types of statistical graphics that can be applied to a pair of attributes without suffering from overplotting, for example, Q-Q (quantile) plot and P-P (percent) plot (Wilk and Gnanadesikan 1968, see also Cleveland 1993). The idea of a Q-Q plot is to plot quantiles of one attribute against the corresponding quantiles of another attribute rather than the original attribute values. For building a P-P plot, the two attributes need to have the common value range (or, at least, their value ranges must significantly overlap). Then, for each attribute value, the percentage of the objects having this or lower value of one of the attributes is plotted against the percentage of the objects having this or lower value of the other attribute.

The use of P-P plots for our data seems inappropriate because of the difference of the value range of the DoE index from those of the counts of the national minorities. In principle, we could normalise the attributes, but the presence of the outliers makes it difficult to find a suitable normalisation method. Since Q-Q plots

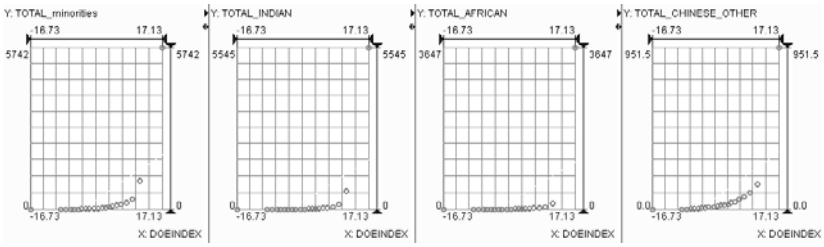

Figure 14. Q-Q plots for exploring the relationships between the distribution of various national minorities and deprivation. Quantiles of the minority counts (from left to right: all minorities, Indian, African, and Chinese) are plotted against the corresponding quantiles of the DoE index.

appear to be applicable to arbitrary attributes, we tried to use them for the visualisation of our data. The result can be seen in Figure 14.

The Q-Q plots show us that the distributions of all minority counts differ very much from the distribution of the deprivation scores. It can be also seen that the distribution of the Chinese minority is somewhat different from those of the other minorities. However, the Q-Q plots do not tell us anything about the relationships between the minority counts, on the one hand, and the deprivation scores, on the other hand.

From interactive data visualisation techniques suggested for analysis of relationships between attributes, the best known is the so-called Influence Explorer (Tweedie et al. 1999), which is based on the use of frequency histograms. A user may have on the screen several histograms for different attributes. By clicking on one or more bars in one of the histograms, the user selects the subset of objects having the attribute values within the respective intervals. In response, all other histograms show in which intervals the values of the other attributes of the selected objects fit, by highlighting segments of the corresponding bars. The heights of the segments are proportional to the number of objects with the values fitting in the intervals. The technique is demonstrated in Figure 15 (for producing this figure, we used our own implementation of histogram display). In this way one may explore how distribution of one of the attributes is related to those of the other attributes, e.g. whether high values of one attribute mostly co-occur with high or with low values of another attribute.

A disadvantage of a histogram is that it is based on a division of the attribute's value range into intervals, which leads to an inevitable information loss since it masks differences between values within an interval. A gap in a value series will be unnoticeable if it is covered by one of the intervals. On a histogram, it is also impossible to see concentrations of close values. When such concentrations are

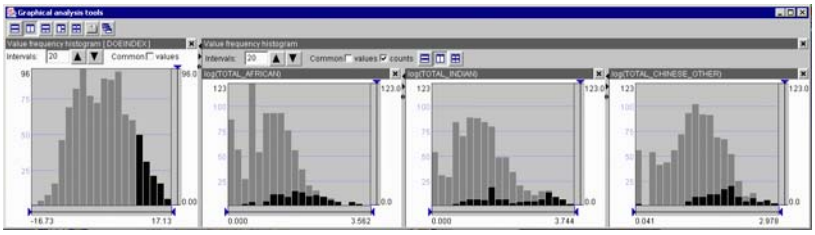

Figure 15. Four histograms are dynamically linked: selection of bars in one of them (left) results in highlighting the corresponding portions of bars in the other histograms. This idea is used in the Influence Explorer.

present in data, the shape of a histogram greatly depends on the choice of the intervals, i.e. whether a break occurs to be inside an interval of concentration or outside of it. As we already mentioned, a cumulative frequency curve is free of these disadvantages: it does not require any intervals and therefore minimises the information loss. Horizontal segments of such a curve reflect gaps between values and steep segments correspond to value concentrations.

Sometimes (as in our case) data may be so particularly distributed that histograms become practically useless. Thus, the histogram for any minority count has one very high bar for the values close to 0, than a few much smaller bars for the next intervals, and the bars for the remaining intervals have zero heights and can only be noticed when "zooming" is applied, which cuts the high bars and stretches the low ones. For producing the screenshot shown in Figure 15, we had to convert the original values of the attributes TOTAL_AFRICAN, TOTAL_INDIAN, and TOTAL_CHINESE_OTHER to logarithms. However, such a transformation makes it more difficult for a user to understand which value interval each bar corresponds to. At the same time, it may be seen from our examples that GCCs are quite tolerant to outliers in data, which is a clear benefit in our case.

Another advantage of a cumulative curve display is that the distributions of two or more attributes (in relation to the same base attribute) can be compared immediately, without prior selection of an object subset, while in Influence Explorer one needs to select one or more bars in one of histograms. For a comprehensive investigation, this procedure needs to be repeated many times. On the other hand, in Influence Explorer all attributes are represented and treated in the same way while in a cumulative curve display there is a difference between the base attribute and the cumulated attribute. It is easy to compare the curves for several cumulated attributes having the same base attribute, but changing the base attribute actually means that a new display is built, which may have quite different properties. This may cause a certain inconvenience. Another limitation of the cumulative curve display is that a GCC can only be built for an attribute representing some absolute amounts, or, in other words, when it makes sense to

sum up attribute values associated with different objects. Thus, it is appropriate to build a cumulative curve for the absolute numbers of representatives of the national minorities in each ward but not for their proportions in the ward's population.

Like with cumulative frequency curve, it is possible to generalise the idea of frequency histogram for representing not only frequencies but also summary counts of areas, population, etc. In this case the heights of the bars may be proportional to the sums of these values. However, several cumulative curves can be much easier combined within a single display and compared than several histograms. On the other hand, histograms are more usual and better understandable for users than cumulative curves.

As a non-spatial representation, a cumulative curve graph alone is unsuitable for exploration of spatially referenced data. Therefore a link between the graph and a map representing the spatial locations of the objects is important. We link the cumulative curve display with a map by means of classification, i.e. division of the geographical objects into groups according to values of the base attribute. The classes are assigned particular colours, and these colours are used for painting the objects on the map. Unfortunately, unlike a cumulative curve display, a map is not a scalable visualisation method. However, when there are clear patterns in the spatial distribution of attribute values, even a map with thousands of objects can be quite well perceived and useful for data analysis.

The representation of classes on a cumulative curve display makes it also a useful tool for classification of geographical objects. It should be noted that geographers and cartographers recognise the usefulness of supporting the classification procedure with graphs showing the statistical distribution of values of an attribute used for defining classes. Thus, Yamahira, Kasahara, and Tsurutani (1985) suggested using a frequency histogram and Slocum (1999) uses so called dispersion graphs. Such representations allow a map designer to balance between geographical and statistical criteria, that is, to define classes of geographical objects so that, on the one hand, the territory is divided into the smallest possible number of coherent regions, on the other hand, variation of data within each class is low while differences between the classes are maximal.

With the tool based on generalised cumulative curves, it is easy to account in classification not only for the statistical distribution of attribute values but also for such criteria as even distribution of population among the classes, or approximately equal total areas occupied by the classes, or other specific criteria that may emerge in this or that application domain. An analyst needs only to move the interval breaks and to observe the resulting segmentation of the corresponding vertical axis. In our implementation (Andrienko and Andrienko 1999), we provide a direct manipulation interface (Shneiderman 1983) for introducing and moving class breaks.

6 Conclusion

In general, there is no “ideal” data representation that would be suitable for any purpose and for any user. In exploratory data analysis, it makes sense to consider the same data from diverse perspectives by visualising the data in various ways. We have demonstrated that generalised cumulative curves have a potential of prompting a particular sort of insights into data characteristics, that is, it allows an analyst to see whether attributes A and B relate to attribute C in the same way or differently. GCCs are, perhaps, especially useful in demographic studies, since they are very convenient for visualising and comparing the distribution of the total population and various population subgroups in relation to certain characteristics of the districts where the people live.

A cumulative curve display combined with a tool for interactive classification of objects according to values of a numeric attribute can be used for defining object classes with desired properties, for example, classes with equal areas or population. A link to a map display allows a user to see immediately the spatial patterns formed by the resulting classes. This makes the cumulative curve display suitable for exploration of spatially referenced data.

We have implemented the cumulative curve display within our system CommonGIS, which is available in the web at the URL http://www.CommonGIS.de/.

Acknowledgement

We thank the reviewers of the paper and the editor for helpful suggestions concerning its improvement.

References

Andrienko, G., and Andrienko, N., 1999. Interactive maps for visual data exploration. *International Journal Geographical Information Science*, **13**, 355-374

Bunting, J., 2000. Measuring deprivation: a review of indices in common use, http://www.swpho.org.uk/pat18discuss.htm

Cleveland, W.S., 1993. *Visualizing Data*, Hobart Press, Summit, New Jersey

Lorenz, M.O., 1905. Methods of Measuring the Concentration of Wealth, *Journal of the American Statistical Association*, New Series, **70**, 209-219

Slocum, T.A., 1999. *Thematic Cartography and Visualization.* Prentice-Hall, New Jersey

Schmid, C.F, and Schmid, S.E, 1979. *Handbook of graphic presentation*. Second Edition, John Wiley & Sons, Inc., New York

Shneiderman, B., 1983. Direct Manipulation: A Step Beyond Programming Languages, *Computer*, August 1983, 57-69

Tukey, J.W., 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading

Tweedie, L., Spence, R., Dawkes, H., and Su, H., 1999. Externalising Abstract Mathematical Models. In Card, S.K., Mackinlay, J.D., and Shneiderman, B. (Eds.) *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann Publishers, Inc., San Francisco, California, pp. 276-284

Wilk, M.B., and Gnanadesikan, R, 1968. Probability plotting methods for the analysis of data. *Biometrica*, **55** (1), 1-17.

Yamahira, T., Kasahara, Y., and Tsurutani, T., 1985. How map designers can represent their ideas in thematic maps. *The Visual Computer*, **1**, 174-184