

Appendix 8: Statistical analyses of social exclusion indicators

Social indicators, of which social exclusion measures are an example, are of surprisingly recent origin. They were defined 40 years ago as “statistics, statistical series, and all other forms of evidence that enable us to assess where we stand and are going with respect to our values and goals” (Bauer, 1966). Social indicator systems are still heavily influenced by the 1975 UN System of Social and Demographic Statistics and the OECD Programme of Work on Social Indicators (OECD, 1982). The UN System was conceptualised by UK economist Richard Stone, who also jointly developed the system of national accounts, for which he won the Nobel Prize for Economics. In Europe the 1971 Delors report, *Les indicateurs sociaux*, has also influenced later work on social planning in all EU member states.

The nature of social indicators has evolved such that they can now be considered to be statistical time series “used to monitor the social system, helping to identify changes and to guide intervention to alter the course of social change” (Ferriss, 1988). However, the problem remains that the term ‘social indicators’ encompasses at least four different concepts:

- *Social statistics*: statistical time series that provide information on social goals, such as indicators, measures and indices.
- *Social accounts*, which assume the existence of a holistic interrelated social system in which the relationship between social goals is understood.
- *Sub-system variables* in which social indicators are output and input measures that can be used to judge ‘progress’ towards social goals (for example, progress in health, education, inclusion, etc). A causal model is implicitly assumed to exist.
- *Quality of life*: objective and subjective indicators of an individual’s (and sometimes society’s) sense of well-being.

The best methods of statistical analyses to use with social exclusion measures to provide evidence for policy makers crucially depends on which of these concepts is applicable, but also on other factors. All statistical analyses depend on both the

theoretical framework and the objective of the analysis. Without these, the question of the ‘best’ method is meaningless. The purpose of this report is to identify potential data on social exclusion within a sub-system variable framework. A causal model is implicitly assumed to exist such that exclusionary processes are assumed to result in individuals and groups of people becoming excluded. It is therefore necessary to consider what kinds of data and statistical analyses are needed to measure causality.

Measuring causality with statistical data

Social scientists agree that measuring causality is very difficult, but it is not impossible. Most attempts by social scientists in the UK to establish causal relationships with statistical data draw either implicitly or explicitly on epidemiological practice. The statistical research of Richard Doll and Austin Bradford-Hill, which established an unexpected causal relationship between smoking and lung cancer, has been highly influential.

Bradford Hill (1965) outlined nine viewpoint/guidelines that could be used to establish causality (rather than just association) with statistical data:

1. *Strength*: the statistical association between the cause and effect should be strong (highly significant).
2. *Consistency*: the same or similar results have been repeatedly found by different people, in different places, circumstances and times, in a range of studies.
3. *Specificity*: ideally one cause should lead to one effect.
4. *Temporality*: the cause should precede the effect.
5. *Biological gradient*: there should be a dose–response relationship between the cause and the effect, that is, the ‘stronger’ the cause the ‘stronger’ the effect.
6. *Plausibility*: it is helpful if the causation appears to be plausible given ‘current’ knowledge. (It should be noted that Bradford Hill had mixed feelings about this criteria since ‘current’ knowledge may be incorrect.)
7. *Coherence*: the cause-and-effect interpretation of the data should not seriously conflict with the generally known facts.
8. *Experiment*: It is helpful if there is some experimental or semi-experimental evidence for the cause and effect either from ‘deliberate’ or ‘natural’

experiments, which for example show that preventative interventions do indeed prevent the effect.

9. *Analogy*: that there is evidence that similar causes have similar effects.

There has been extensive debate on the relative merits of Bradford Hill's nine viewpoints over the past 40 years (Cox, 1992; Goldthorpe, 2001; Thygesen et al, 2005). Although it is possible to reject all these nine viewpoints as 'just' inductive criteria, they have proved useful in many pragmatic studies into causality. All of the nine viewpoints can be criticised, and examples/'special cases' found where they do not work; however, some remain more controversial than others. In particular, specificity has been criticised since it is rare in social science for a single cause to have a single effect: with social phenomena effects may be due to multiple causes and causes may have multiple effects.

Similarly, there is nothing in the philosophy of logic or science that requires a cause to occur before an effect – cause and effect can occur simultaneously and in modern physics an effect may even occur before the cause². Nevertheless, in social science temporality is often seen as a useful criterion and longitudinal panel data can be used in measuring this phenomenon. However, it should also be noted that the only major advantage longitudinal panel data have over repeated cross-sectional survey data is their ability to provide evidence on temporality (that cause occurs before effect); the other eight of Bradford Hill's criteria can often be better measured using other types of statistical data. It is not, therefore, the case that longitudinal data sets are preferable, especially if the range and quality of data in relation to the questions asked is limited.

The key point from the Bradford Hill viewpoints for this review is that in order to establish causal relationships in social exclusion studies, many data sets/surveys will be needed – causality is very unlikely to be established from the analysis of just one 'perfect' longitudinal panel survey or randomised control trial. Although, studies of social exclusion should be commissioned using the 'best' data sources, it is also important to check the results from these analyses using the 'second best' data sources. This is particularly important if the analysis of the 'best' survey data suffer from model selection bias.

The problem of model selection bias

Statistical theory assumes that models are known and have been clearly specified before the analysis. However, in reality models are almost never known in advance and the same data set is often used to both formulate the models (select the most ‘important’ variables) and then to undertake the analyses (fit the models). Almost all studies in social science undertake some analyses on the data prior to fitting a statistical model (for example, running a regression analysis) in order to select the best sub-set of variables to use in the modelling. In fact, many studies make use of step-wise procedures, where a lot of variables are entered into the model and the ‘computer’ removes the variables that are not statistically significant. Thus the explanatory variables in a model have usually been selected from a much larger initial set of variables. This is known as ‘data mining’ and suites of analysis software have been developed solely for this purpose.

Unfortunately, it is not possible to make inferences as if a model is known to be true when it has, in fact, been selected from the *same* data to be used for estimation purposes (Zhang, 1992). If this is done then the p values and other model parameters will be incorrect.

Chatfield (1995) argues that there are typically three main sources of uncertainty in any problem:

- a) uncertainty about the structure of the model;
- b) uncertainty about estimates of the model parameters, assuming that we know the structure of the model;
- c) unexplained random variation in observed variables even when we know the structure of the model and the values of the model parameters.

Uncertainty about model structure can arise in different ways such as:

- i) model misspecification (for example, omitting/including a variable by mistake);

- ii) specifying a general class of models of which the true model is a special, but unknown, case; or
- iii) choosing between two or more models of quite different structures.

Statistical theory has much to say about (b) and (c) and about the mechanics of the choice in (ii) (for example, F-tests in analysis of variance [ANOVA]), but it has little to say about (iii) and even less about (i), and largely ignores the effects of (a) in ensuing inferences. This is very strange given that errors arising from (a) are likely to be far worse than those arising from other sources (Chatfield, 1995, p 421).

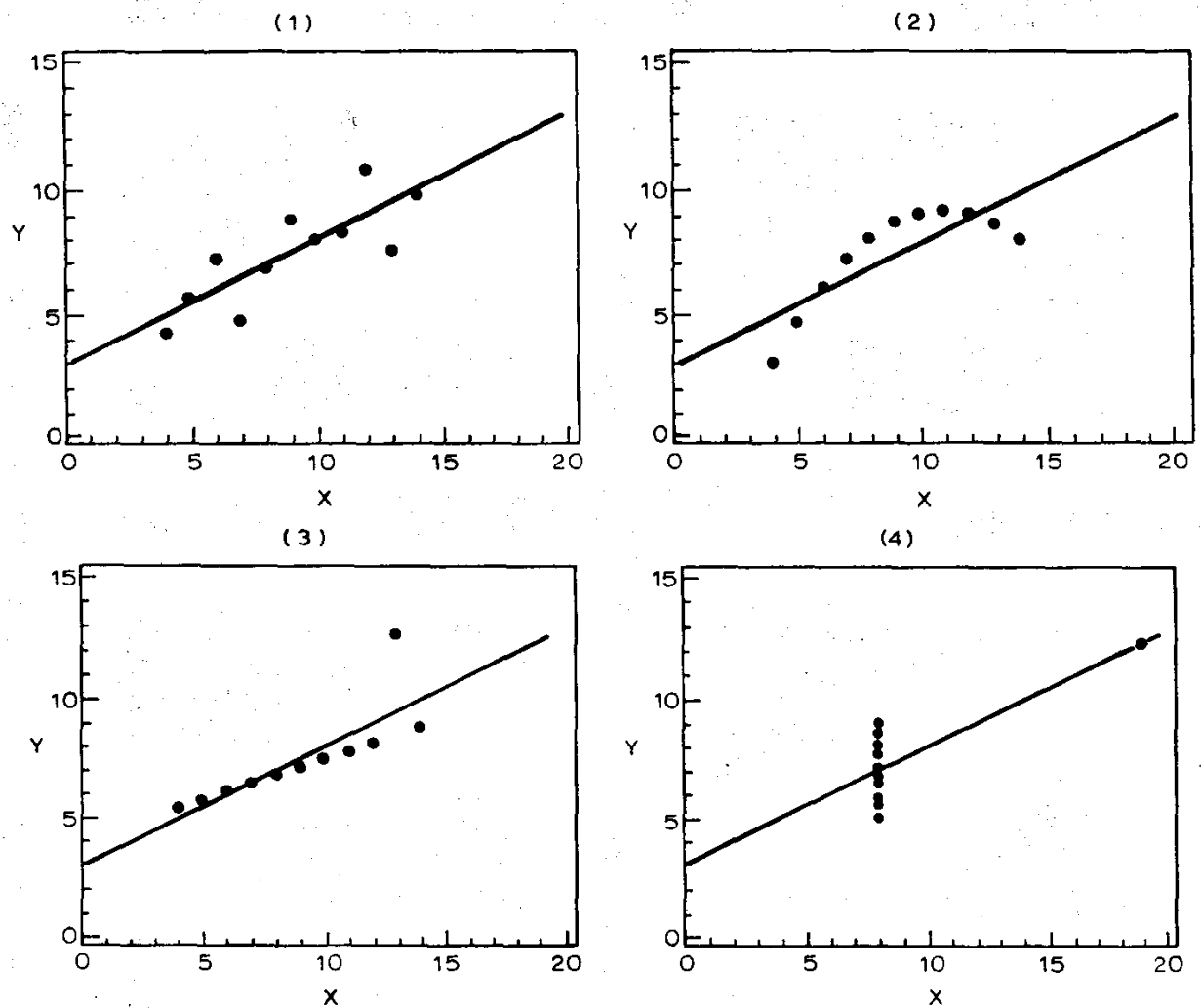
The only real solution to this largely unacknowledged problem (it is entirely ignored by most textbooks on statistics) is to develop models on one dataset (that is, select the statistically important variable) and then fit the models on a completely different dataset. This has important implications for analyses on social exclusion using the BHPS that is currently the only available longitudinal panel survey for the UK. It would be far better to develop models (find the ‘best’ variables) of social exclusion using different survey data and only then attempt to fit these models using the BHPS. To date this has not been done in any study of social exclusion in the UK.

Univariate methods

Many analyses of social exclusion indicators venture little beyond the production of contingency tables. However, most people are not good at extracting any useful information from numerical tables and most published results do not follow the simple rules that can be used to improve the ‘readability’ of tables (see Tufte, 1983; Cleveland, 1993).

Statisticians have repeatedly argued that much of the data and statistical analyses presented as tables would be better presented graphically. Over 30 years ago, Anscombe (1973) argued that graphs are essential to good statistical analysis. He illustrated this point effectively by presenting four scatter plots showing very different relationships between variables y and x . All four data sets had similar statistical properties but in three of the four cases the statistical analysis would mislead a researcher who failed to graphically display the data (Franke, 1997).

Figure 6.1: Anscombe's illustration of the need to visualise data



Source: Anscombe, F. J. (1973) Graphs in Statistical Analysis. *The American Statistician*, 27, 17-21
Reprinted with permission from *The American Statistician*. Copyright 1973 by the American Statistical Association. All rights reserved.

In Figure 6.1 the first plot shows the stereotypical scatter of points around the fitted regression line, the second shows a parabolic relationship between the variables, the third shows a linear relationship between y and x for all but one data point, and the fourth shows that the fitted regression line is completely determined by a single outlying observation. Yet the basic regression results for all four data sets are identical. They have the same regression coefficient and intercept, same R^2 , and same significance level. Unfortunately, even statisticians often fail to practice what they preach and frequently resort to publishing results as tables when graphical methods would be preferable (Gelman et al, 2002).

The ‘best’ statistical methods to use for the analyses of social exclusion indicators will often be graphical methods, particularly where changes over time and dynamic processes are being measured using multiple indicators/multivariate data. There are a wide range of statistical software packages that have been designed to help with the visualisation and exploration of these kinds of data, for example BMDP/SPSS Diamond (Franke, 1997) However, these often expensive commercial software packages may be superseded in the near future by free software such as Gapminder (see www.gapminder.org/) and similar developments. It should be noted that when analysing data using visual methods Mark et al (2002) found “that people who worked in groups were more correct in their answers for objective questions, based on searching a large dataset” and that “groups were more accurate in their results for a free data discovery task...”. We suggest that given the right visualisation system, groups do better than individuals in finding more accurate results.

Multivariate methods

Most multivariate analyses of social exclusion data are undertaken using General Linear Model (GLM) techniques. Regression analysis (including logistic regression), analysis of variance, discriminant function analysis and canonical correlation are all versions of the GLM. Regression analysis was first developed by Legendre (1805) and Gauss (1809) to fit data on the orbits of astronomical objects. At the end of the 19th century it was used by Yule (1899) to investigate the causes of changes in pauperism in England. This was one of the first studies to use multivariate statistical techniques to try to establish causal relationships in the social sciences – to study the ‘social physics’ of poverty (Freedman, 2002).

Yule’s analysis of census data used only four variables (pauperism = out relief ratio + proportion of the old + population) and ‘showed’ that providing ‘out relief’ caused pauperism. With the benefit of 20:20 hindsight and modern knowledge it is obvious that Yule’s model was hopelessly mis-specified – it omitted many important variables that should have been included, for example, the availability of work. This illustrates the problems of model selection and model bias discussed previously, and is a reminder that in a hundred years time our causal models of social exclusion may also seem equally over-simplistic. The usual GLM methods make a number of

assumptions about social exclusion data that are rarely correct (for example, data are continuous, normally distributed, etc).

Social exclusion indicators are often designed to measure the lower tail of a normally distributed variable that measures the degree to which people are excluded from obtaining goods and services. Thus, social exclusion indicators give special attention to the ‘worst off’ section of the population. The statistical problem is that most of the population is not excluded, so there will be a large proportion of zero scores. Common regression techniques that require variables with a normal distribution, like OLS (ordinary least squares), will most likely produce biased and unreliable estimates with these kinds of data (Breen, 1996). However, social exclusion measures can be looked on as censored variables that represent an underlying unobserved normally distributed variable.

If this is the case, they can therefore be analysed as metric variables using a Tobit Regression Model. The Tobit Model assumes a relationship between the latent (unobserved) normally distributed variable y^* and the actually observed outcome measure y . The latent dimension y^* is equal to the observed dependent variable y if y exceeds the cut-off point for censoring, which in this case is zero. If y^* is below the cut-off point, y is equal to the cut-off value, that is, zero. The model assumes a linear relationship between the independent variables and the latent dimension y^* based on a likelihood function that estimates a linear model for the metric data and a probability model for the censored data. The estimated Tobit regression coefficients are linear and additive in relation to the latent response continuum, not in relation to the observed measures.

Multiple causation

An additional problem with the multivariate analyses of social exclusion data is that a multiple causation framework implicitly underlies these analyses. In other words, there is assumed to be a complex ‘web of causation’, with numerous interconnected risks and protective factors, which results in social exclusion. A similar multi-causal framework is also commonly found in many other subjects, for example, epidemiology (Krieger, 1994). However, there is little theoretical evidence to believe that a causal web is the correct framework explaining social exclusion – there is

currently no detailed theory that would lead to this belief. Instead there is detailed multivariate statistical method and an implicit assumption in the existence of a causal web.

The primary problem with this situation is that both the multivariate methods and the nature of available survey data tend to focus analyses at the individual and household level. However, the causes of social exclusion of population groups may be more than the sum of the causes of social exclusion at the individual or household level. For example, it may not be possible to explain the social exclusion of refugees or travellers from the multivariate analyses of the characteristics and circumstances of individuals. There may well be structural and hierarchal effects and the primary causes of the social exclusion of population groups and of individuals may be different.

Spatial data analysis

There is increasing evidence that a 'one size fits all' set of policies to tackling the problem of social exclusion can be both ineffective and inefficient. Exclusion is not a result of a single set of processes that effect all groups equally wherever they may live. There may be considerable and significant variations in the causes of exclusion depending on who you are and where you live and in order to be effective and efficient polices need to be tailored to address local needs.

One of the most confusing issues with measuring social exclusion is to identify the correct level of analysis. In particular, should the analysis be carried out at the individual, household/family, population group or area level? A considerable amount of research has been carried out at both the household and area level. Unfortunately, some researchers have not realised that analysis at these different levels requires the use of different statistical techniques and therefore much of the area-level research has been of relatively poor quality.

Area-level data tend to violate two basic assumptions of most statistical techniques:

1. *All observations are independent*: the answer one person gives to a question should not affect or influence the answer another person gives to the same question.
2. *Measurement errors are normally distribute*: it is impossible to measure anything perfectly. There will always be some error of measurement. What matters is there is no systematic bias in the measurement error.

Area-level data often violate these assumptions due to the following.

Spatial autocorrelation

Areas next to each other are likely to be more similar than areas further away. Spatial autocorrelation can be defined as the clustering pattern in the spatial distribution of a variable that is due to the very fact that the occurrences are physically close together, that is, that they are in geographical proximity. They are not independent of each other, but are linked. The data are spatially dependent.

Spatial autocorrelation is widespread: rich people move to areas where other rich people live; disease can spread from one neighbour to another, etc. If the values in a poverty or health 'cluster' are more alike than would be due to random processes, there exists a positive autocorrelation; if they are less alike than would occur through random processes, there exists a negative autocorrelation.

Modifiable area effect

All area boundaries are artificial social constructs that are usually a result of 'political' or 'religious' influences, for example, electoral divisions, parishes, etc. If the areas change so will the strength of the association between any two variables such as poverty and health.

This change in association between variables with the selection of different areal units can call into question the reliability of results. The effect of the selection of areal units on analysis is termed the modifiable areal unit problem (MAUP). It is formally defined as:

... a problem arising from the imposition of artificial units of spatial reporting on continuous geographical phenomenon resulting in the generation of artificial spatial patterns. (Heywood et al, 1998).

The MAUP had been most prominent in the analysis of socioeconomic data. Such areal data cannot be measured at a single point, but must be contained within a boundary to be meaningful. For example, it is not possible to measure the percentage of infant deaths or low-income households at a single point; this percentage must be calculated within a defined area. It is the selection of these artificial boundaries and their use in analysis that produces the MAUP.

The effects of the MAUP can be divided into two components: the scale effect and the aggregation or zonation effect. The scale effect is the variation in numerical results that occurs due to the number of zones used in an analysis. The aggregation or zonation effect is the variation in numerical results arising from the grouping of small areas into larger units. The use of small areal units has a tendency to provide unreliable rates because the population used to calculate the rate is smaller. On the other hand, using larger areal units will provide more stable rates but may mask meaningful geographic variation evident with smaller areal units.

Other area problems

There is a long list of other potential statistical problems in analysing area-level data, including the ecological fallacy that can occur where assumptions about individuals are made from the analysis of area-level data (see Robinson, 1950). For example, in many countries 'poor' areas suffer from high crime rates and this often leads people to assume that 'poor' people suffer high levels of crime. However, in many countries it is the 'richer' people living in 'poorer' areas that are often the victim of many crimes.

Statistical solutions

A number of different statistical solutions have been developed that to varying degrees can overcome these area-level analysis problems. Among them are Poisson regression, multi-level modelling, generalized regression (GREG), synthetic estimators, composite estimators (along with modifications for logistic regression and 'ecological effect'). However, the 'best' statistical method currently available for

analysing spatial data is probably geographical weighted regression (GWR). GWR is a relatively new modelling technique for spatial analysis that allows local, as opposed to global, spatial models to be calculated and interesting variations in relationships to be measured and mapped. GWR is based on the logical idea that relationships between variables measured at different locations might not be constant over space, for example, lack of available public transport might be a more important cause of exclusion in rural areas than in inner-city areas. However, almost all statistical techniques assume that relationships between variables are constant and do not change from one area to another. If relationships do vary significantly over space, then serious questions are raised about the reliability of traditional, global-level analyses.

GWR is not restricted to simple global analyses in which interesting local variations in relationships are ‘averaged away’ and unobservable; GWR allows these local relationships to be measured and mapped. This modelling approach challenges many of the global statements of spatial relationships that have been made in the academic and policy literature.

Fotheringham et al (2002) explain how GWR differs from the ‘usual’ OLS ‘global’ regression model as follows. A global regression model is given by:

$$y_i = a_0 + \sum_k a_k x_{ik} + \varepsilon_i$$

In the calculation of this model, one parameter is estimated for the relationship between each independent variable and the dependent variable and this relationship is assumed to be constant across all geographic areas in the study region. The estimator for the parameters in this model is:

$$a = (X^t X)^{-1} X^t y$$

where a represents the vector of global parameters to be estimated, X is a matrix of independent variables with the elements of the first column set to 1, and y represents a vector of observations on the dependent variable. GWR is a relatively simple

technique that extends the traditional regression framework by allowing local rather than global parameters to be estimated so that the model is rewritten as:

$$y_i = a_0(u_i, v_i) + \sum_k a_k(u_i, v_i) x_{ik} + \varepsilon_i$$

where (u_i, v_i) denotes the coordinates of the i th point in space and $a_k(u_i, v_i)$ is a realisation of the continuous function $a_k(u, v)$ at point i (Brunsdon *et al*, 1996). That is, we allow there to be a continuous surface of parameter values and measurements of this surface are taken at certain points to denote the spatial variability of the surface. Note that the global model is a special case of the GWR model in which the parameter surface is assumed to be constant over space.

In the calibration of the GWR model it is assumed that observed data near to point i have more of an influence in the estimation of the $a_k(u_i, v_i)$'s than do data located farther from i . In essence, the equation measures the relationships inherent in the model *around each point i* .

Conclusions

This review is about the multi-dimensional analyses of social exclusion. The underlying statistical framework is a sub-system variables approach in which social indicators are identified as output and input measures which can be used to judge 'progress' towards the social goal of inclusion. A causal model is implicit in this work that will facilitate the identification of policy interventions that will reduce exclusion. The 'best' type of statistical analysis to use depends on both the nature of the survey data available and the purpose of the analysis. In many circumstances when trying to analyse dynamic data, visualisation of the data may be more successful than formal statistical modelling. When undertaking causal modelling it is important to analyse multiple data sets rather than just a single survey, particularly when the 'correct' model is unknown.

Notes

¹ Crockett (2006, p 6) notes that prior to version 12, SPSS (Statistical Package for the Social Sciences) was incapable of correct weighted data analysis resulting in spurious statistical significance. Although SPSS now produces correctly weighted analysis for

descriptive measures (using the complex samples module) this is still not the case for type multivariate approaches.

² If you are not a physicist it is inadvisable to dwell on this point for too long.