Dirty data: longitudinal classification systems

Emma Uprichard

Abstract: Typically in longitudinal quantitative research, classifications are tracked over time. However, most classifications change in absolute terms in that some die whilst others are created, and in their meaning. There is a need, therefore, to re-think how longitudinal quantitative research might explore both the qualitative changes to classification systems as well as the quantitative changes within each classification. By drawing on the changing classifications of local food retail outlets in the city of York (UK) since the 1950s as an illustrative example, an alternative way of graphing longitudinal quantitative data is presented which ultimately provides a description of both types of change over time. In so doing, this article argues for the increased use of 'dirty data' in longitudinal quantitative analysis, a step which allows for the exploration of both qualitative and quantitative changes to, and within, classification systems. This ultimately challenges existing assumptions relating to the quality and type of data used in quantitative research and how change in the social world is measured in general.

Keywords: classification, classification systems, longitudinal quantitative analysis, York, food

Introduction

Longitudinal data analysis has been taking place for over three centuries (Menard, 2002). Over the past decade, however, within the social sciences in particular, quantitative longitudinal research has gained a growing currency and witnessed a flourishing level of interest. This is because of the increased public access to large-scale longitudinal surveys and the fact that longitudinal techniques have become more available in standard quantitative software packages (Hosmer *et al.*, 2008). At the heart of much of this work lies the problem of how to capture *trajectories*. Typically, such endeavours involve the combined attempt to understand: (a) *what* has changed, (b) *how* that thing has changed, and at least implicitly if not explicitly, (c) *why* that thing has changed the way it has. The onus, then, is often on developing retrospective historical descrip-

tions in an attempt to consider issues of causality, which in turn feed into developing prediction(s) about possible future trajectories.

Whereas a lot of attention has been paid to the difficulties in determining causal models in longitudinal research, this article focuses instead on the problems involved in developing *descriptions* of classifications over time. This is not to say that issues of causality and prediction are unimportant or that they can be easily disentangled from issues of description – as C. Wright Mills (1959: 170) argues, prediction and description 'are not to be sharply separated'. Rather, the view here is that to even begin exploring issues of causality within the complex social world, it is necessary to first trace a rich description of the changes to the classifications *as they exist*, that is, without any 'cleaning up', however problematic that may be.

In effect, the argument is a simple one: when it comes to exploring, describing, measuring quantitative change longitudinally, it is necessary to use 'dirty data' – a term used deliberately to argue against 'cleaning up' classifications. This contrasts with current practices in quantitative (and increasingly qualitative) longitudinal research, which tend to entail the more or less rigorous merging, recoding of 'suitable' categories that would seemingly facilitate measuring change longitudinally. Instead, by leaving the data alone, and using the categories in their raw form as much as possible, descriptions of both qualitative *and* quantitative changes to social classifications systems are rendered possible.

Analysing classifications over time is, after all, where the money is. Predicting changes and continuities to classifications – *not* to variables, but to polythetic *cases* – is what the global commercial enterprises, such as the large supermarket chains, insurance and credit card agencies, etc. are now doing with quantitative data collected via our transactions. Even quantitative software packages such as SPSS – re-branded by IBM as PASW in version 18, reflecting precisely this shift in 'Predictive Analytical Statistics' where classification and visualization are sold as its forte – are in on the game (see Uprichard *et al.*, 2008). Commercial enterprises are doing it; social scientists, however, are rather embarrassingly lagging behind and merely thinking or talking about it – see Savage and Burrows (2007) on the 'empirical crisis' within sociology specifically. As Byrne (2002: 35) suggests, 'the movement of systems of classes and indeed even more importantly, changes in classification systems over time, is exactly what should interest us most'. Abbott (2000: 299) goes as far as stating that a 'central challenge for the future of all social research is to figure out how to handle this category change without simply sweeping it under the rug'. In this article, category changes are *not* swept under the rug; they are moved to centre stage.

To be sure, there are plenty of quantitative techniques used to describe changing social objects of study over time. Crudely speaking, existing approaches vary according to the extent to which they focus on the variable(s) rather than the case(s) (see Byrne and Ragin, 2009). Many time-based statistical techniques, for instance, tend to zoom in on the change of one or more variables or the timing of events (Singer and Willett, 2003). Other procedures, such as agent-

based, cellular-automata or mathematical modelling (eg Schelling, 1971; Batty, 2005; Allen, 1997; Gilbert and Troitzsch, 1999) have been developed, which arguably also tend to simulate change in one or more variables, albeit in different ways. Then there are narrative approaches used to combine qualitative and quantitative methods (Elliott, 2005), and temporal qualitative comparative analysis (Caren and Panofsky, 2005; Chaplain *et al.*, 1999; Ragin and Strand, 2008). The list continues, but the point is that, whilst different in detail, the existing approaches tend to share a focus on describing quantitative change in one or more variables, without paying attention to the qualitative changes to those variables. As Abbott (2000: 205) notes, 'Work involving data over time often boils down to "trends and counts," with little sensitivity to historical context or contingency, and less to qualitative temporal fluctuations in the social categories or attributes analyzed.'

There is nothing wrong with describing 'trends and counts'. This may well be all we can meaningfully do and, as will be seen, is also the basis for what is proposed here. However, what needs to happen alongside such a description is an explicit exploration of category changes that may underpin that description vis-à-vis *what* is counted, *how* those things are counted, and in turn *why* the trends which are based on those counts are as they are. After all, irrespective of the approach employed, and whether or not the desired data(set) exists (which often it does not), data are themselves part and parcel of that changing object of study. Quantitative data are *always and necessarily* manifestations of the processes involved in describing a changing object of study; data do not exist in a vacuum independent of our knowledge any more than the objects they represent (see Desrosières, 1998). Therefore, whatever is described in whatever way it is described, a quantitative description says as much if not more about the method used as it does about the social world it reveals. Ideally, then, any chosen methodology must consider the nature of the abstractions whilst also attempting to maximize the probability that the object is adequately and accurately described and explored (see Reed and Harvey, 1996; Sayer, 2000). In turn, in order to develop any meaningful longitudinal description of change, it is necessary to consider the co-construction of the actual description obtained and the historical contingency involved in the construction of that empirical description (see Dupré, 2006; Desrosières, 1998, 2000, 2001; Hacking, 1999, 2002).

These issues are discussed here specifically in relation to the changing classification of food outlets in the city of York (UK) between 1951 and 2001. Although the example of food outlets is quite specific, the methodological issues discussed are ones that are likely to be present when exploring change in other substantive topics. Perhaps a more politically loaded example might have been occupational, class, race and ethnicity classifications in the census. Whatever example is used though, classification, which gets used as variables (see Tryon's (1939) early work on clustering cases versus variables), need to be tracked over time; the qualitative changes *to* the classification or variables are just as important as the quantitative changes *within* them. Hence, many of the challenges involved in studying social change empirically are arguably methodological

ones rather than substantive ones to do with the nature of the object of study. This is not to say that the nature of the object of study is unimportant, but that there are methodological issues that relate to *longitudinal* social research, irrespective of what case is being examined, precisely because studying something that changes inevitably raises its own methodological issues. The argument, therefore, relates to any attempt to explore the social world over time and the many classification systems through which change happens (see Bowker and Star, 2002, on this issue).

Three points are worth noting before moving on to discuss the study. First, the discussion is much informed by Bhaskar's (1975, 1979) critical realist philosophy. There is plenty of scope to argue that other approaches might have been preferable. Indeed, there may be a good case to argue that what is presented is compatible with alternative philosophical perspectives or authors. Ultimately, the positions of authors arguing against Kant's transcendental argument about the extent to which we can deduce the world based on our (limited) knowledge of it will probably not be incompatible with many of the views presented here. Nevertheless, since this paper relates to the practice of studying social change empirically, the discussion places more value on how to measure change than it does on the value of understanding how change occurs. In turn, therefore, Bhaskar has been used because he deals explicitly with the empirical problems of knowing the social world, specifically by adapting a hard empiricist perspective to the study of society, and offers therefore some useful hinge-pins to the discussion here, since hard empiricism arguably still underpins the legacy of much variable driven quantitative analysis.

Secondly, in contrast to authors such as Hacking, Bowker and Star, and Dupré – who also inform this argument in important ways – here it is suggested that exploring changes in measurement and classification needs to be an *empirical* enterprise. It is not enough to merely reflect on classifications, which is arguably what these authors have done. Instead, the value of obtaining a rigorous empirical description of a changing classification system is underscored here. Third, what is presented here is a little different to the two major schools of thought that, Bowker (2002: 8) suggests, are typical of researchers exploring historical classifications. Whereas one approach involves using 'classifications available to actors at the time' (see, for example, Hacking (1995) on child abuse), the other approach is to 'use the real classifications that progress in the arts and sciences has uncovered' – the example he gives is 'Tort's (1989) work on "genetic" classification systems, which were not so called at the time, but which are of vital interest to the Foucauldian problematic'.

Here, the approach is more similar to Hacking's work on the history of child abuse, inasmuch as the food retail categories are collated *as they appear* in the directories, irrespective of whether or not they are meaningful now. However, the approach is also sensitive to the locality to which they refer. In other words, it has a foot in both of the major schools of thought insofar as the aim is to note both the changing nomenclature of the classifications, for example 'bakers' versus 'bakers and cakeries', whilst at the same time considering what these

classifications mean within the local context of York itself. The effort to consider the inter-dynamics between the changing classifications and the changing city is key, Bowker (2002: 10) argues, to understanding 'classification systems according to the work that they are doing – the network within which they are embedded'.

Category and frequency change in food retail outlets

In the discussion that follows, the obstacles, and importantly, the resolutions, involved in what first appeared to be a simple exercise of counting the change of food outlets over time are explicated. Thus, the argument is derived from the genuine difficulties that emerged as part of a wider study relating to understanding changes in food and eating practices since 1945 in the city of York and the UK. From the outset, certain changes relating to contemporary urban regions were assumed to have occurred, such as the shift from industrial small sector retail outlets to a more service-based economy that relates to reorganization of the global financial, cultural and political economy (see Mérenne-Schoumaker, 1996; Sassen, 1994; Zukin, 1995). For example, it was expected that the number of local inner-city bakeries had declined, whereas the number of local restaurants increased, and that the changes reflected a general qualitative change to the city and surrounding urban regions. However, *empirically* exploring how these general changes manifest themselves specifically with respect to the food retail outlets at the *local* level of York itself was deemed to be a fundamental part of understanding how change may manifest itself locally and through time.

In order to sketch the changing numbers of local food outlets, two key sources were used: *Kelly's Directories* and the better-known *Yellow Pages*; the latter replacing the former in 1975. Both sources require a fee for traders to be listed, and although there are issues that one needs to consider when using them to do with their content and reliability for reasons which are discussed below, they nevertheless provide an accessible and continuous source of local trade data (see Shaw, 1978; Timmins, 1978) and have therefore frequently been used as a way of studying local change (eg Jopson and Reeder, 2008; Shaw, 1984; Shaw and Alexander, 1994).

Although each directory appears annually, the UK national population census dates were taken as the same time-points used to explore the food retail industry, thus allowing the census to provide an additional source with which to understand other socio-economic changes at national, regional and local level levels. Thus, snap shots at 1951, 1961, 1971, 1981, 1991 and 2001 were taken. This entailed using the local *Kelly's Directories* for 1951, 1961, 1971 and the *Yellow Pages* for 1981, 1991 and 2001. (The *Yellow Pages* moved to a 'split year' publication in 1991, so instead of a 1991 publication, there was a 1991/1992 version, which was used as the 1991 snap-shot; likewise 2001/2002 is used to reflect 2001.)

   Once the basic logistics of the research design were chosen, the next stage was to examine what the categories were and to, quite simply, count the number of entries under each category. This proved to be a classifier's nightmare, or a classifier's dream, depending on which perspective is taken. Taking 'bakers' as a first example, the category changes, as well as the quantitative changes within each category, made what was first thought to be a simple exercise into one that signalled many difficulties to come in other categories also (see Table 1). In 1951, 'bakers' are classified under 'bakers & pastry cooks', of which there are a total of 98. In 1961, as expected, the number of entries listed under this same category decreases, to just 66. In 1971, however, the category, 'bakers & pastry cooks', is no longer available, but 'bakers' now is; this time there are 64 listed entries, just two entries fewer than the decade before, suggesting that these same organizations were possibly the same entries in the previous two decades also despite the different wording of the categories. In 1981, the category 'bakers' disappears and is replaced with 'bakers & confectioners', a category which lasts until 2001, going from 31 units in 1981, up to 35 in 1991 and then back down to 32 in 2001.

   Similar issues are present with regards to 'green grocers' over the same time period. However, instead of a total of four different categories appearing between 1951 and 2001, there are a total of ten categories over the same time (see Table 2), making the category change more complicated, as well as the difficulty of measuring change, greater. The changes to 'cafés and restaurants' (see Table 3) are similar, albeit greater still. In 1951, there is just one category, namely 'cafés, restaurants and caterers', which has morphed into seven categories by 2001, with another eight other categories appearing in between; fifteen categories are altogether created over the total time period, each supposedly accounting for a different kind of business at the local level.

   Note that the categories have been ordered in each table quite specifically. Just as in the visualization of boxplots it is necessary to first rank the cases

Table 1: *Category and frequency changes to 'bakers' in York, 1951–2001*

'BAKERS'	1951*	1961*	1971*	1981**	1991/92**	2000/01**
Bakers & Pastry Cooks	98	66				
Bakers			64			
Bakers & Confectioners Retail				31	35	32
Bakers & Confectioners Sundriesmen				1		

*York Directories; **Yellow Pages

Table 2: *Category and frequency change in '(Green)grocers' in York, 1951–2001*

'(GREEN)GROCERS'	1951*	1961*	1971*	1981**	1991/92**	2000/01**
Grocers, tea and coffee provision merchants	280					
Fruiterers (retail)	90					
Greengrocers (retail)	59					
Grocers		236				
Fruiterers & Greengrocers		66	53	44		
Grocers – Retail (e.g. Jacksons)				65		
Grocers – Licensed				18		
Greengrocers & Fruit Sellers					41	16
Grocer & General Stores					41	
Grocers & Convenience Stores						42

Table 3: *Category and frequency change in 'Cafés & Restaurants' in York, 1951–2001*

'CAFES/RESTAURANTS'	1951*	1961*	1971*	1981**	1991/92**	2000/01**
Cafes, Restaurants and Caterers	54					
Restaurants, Cafes & c.		31				
Snack Bars		2		2		
Fried Fish Shops			47			
Restaurants			28	64	97	126
Fish & Chip Shops & Restaurants				45	42	36
Cafes & Cafeterias				11	32	
Take Away Food Shops				11	39	
Coffee Bars				9		
Take Away Food						75
Cafes						41
Sandwich Shops & Delivery						14
Pizza Delivery & Take Away						8
Internet Bars & Cafes						1

according to absolute values of the particular variable to be plotted, so too have the categories been assigned a particular row based on the year they appear and the frequency of entries within them. The first row in each table, for example, is a 1951 category, namely whichever has the highest frequency of entries listed within it. Thus, in Table 2, 'Grocers, tea and coffee provision merchants' has 280 entries in 1951, the highest frequency of all the included categories in that year, so it is also the first row in that table. For each year, then, as far as was possible, the order of the categories in each table also reflects something about the frequency within particular categories. (Note that the ordering of cells needs to be a bit flexible, since the degrees of freedom decrease within the table with every previous entry, ie the positioning of rows for each column is determined by both the frequency and sequence in which a category appears, but each category is itself determined by all the antecedent cells.)

The careful reader will already detect three patterns within each of these tabular stories of change. First, categories present in 1951 do not tend to exist any more by 2001. Second, some categories are created in one year but then disappear immediately afterwards and become replaced by others that then in their turn also 'die' and become replaced by other newer ones as time goes on. Third, visually, a kind of 'step' pattern is observed between the categories over time, which is shaded in the tables to make it more visible – more about this shortly. At this point, as a purely exploratory exercise, and also because it became apparent that the 'death' and 'birth' of categories was relative to all the categories within each year as well as across the years, it was decided to place all the categories between 1951 and 2001 in one table – see Table 4 below. From this table, a total of 46 categories are seen dying, adapting, emerging all together in relation to one another.

What is interesting, although perhaps not surprising, is that these three patterns are even more visible when exploring *all* the categories. Now, we not only observe that out of the nine categories that appear in 1951, only one – namely 'Butchers' – survives through to 2001. No other category shows such longevity. 'Fishmongers' is next in terms of category durability, appearing in 1961 and lasting through to 2001. The next longest surviving categories include 'restaurants', which lasts for four successive decades, appearing for the first time in 1971, whereas 'Fish & chip shops and restaurants' lasts for three successive decades. All the rest die out quickly after just one or two decades.

Note that 'Confectioners (retail)' also lasts four decades, but this also signals a possible issue with the 1971 data as a whole, inasmuch as it seems that this particular category has been missed out for just this year. It would certainly make sense within the context of the city of York and its industrial history in chocolate making with *Rowntree's* (now *Nestlé–Rowntree*, from *Nestlé*'s take-over in 1988) and *Terry's* (which finally closed in 2005) playing a key part in the lives of families across several generations, that 'Confectioners', as a category, is relatively robust.

The same 'hole' in the 1971 data is present for at least another two categories. There are only two 'Snack Bars' in 1961 and again in 1981, suggesting that there

Table 4: *Category and frequency changes to food retail outlets in York, 1951–2001*

	1951	1961	1971	1981	1991/1992	2000/2001
Grocers, tea and coffee provision merchants (retail)	280					
Confectioners (retail)	156	80		21	16	
Butchers	122	99	76	62	56	33
Bakers & Pastry Cooks	98	66				
Fruiterers (retail)	90					
Fried Fish Dealers	70	56				
Green Grocers (retail)	59					
Cafes, Restaurants and Caterers	54					
Fishmongers, Poulterers etc	28					
Grocers		236				
Fruiterers & Greengrocers		66	53	44		
Restaurants, Cafes & c.		31				
Fishmongers		16	11	4	5	3
Snack Bars		2		2		
Poulterers		1		1		
Health Food Stores		1				
Bakers			64			
Fried Fish Shops			47			
Restaurants			28	64	97	126
Pork Butchers			27			
Grocers – Retail (ie Jacksons)				65		
Fish & Chip Shops & Restaurants				45	42	36
Bakers & Confectioners Retail				31	35	32
Grocers – Licensed				18		
Dairies				17	11	4
Take Away Food Shops				11	39	
Supermarkets (1 Tea Cos)				11	36	26
Cafes & Cafeterias				11	32	
Coffee Bars				9		
Frozen Food & Freezer Centres				6		
Delicatessens & Cooked Meats				5		
Health Food Shops				4	9	6
Bacon & Ham Curers and Suppliers				1		
Bakers & Confectioners Sundriesmen				1		
Greengrocers & Fruit Sellers					41	16
Grocer & General Stores					41	
Dellcatessens					7	10
Freezer Centres					5	4
Pick your Own Fruit & Vegetables					2	0
Take Away Food						75
Grocers & Convenience Stores						42
Cafes						41
Sandwich Shops & Delivery						14
Pizza Delivery & Take Away						8
Internet Bars & Cafes						1
Organic Foods						1

were probably still at least two snack bars in 1971, but they were just not counted. Similarly, there is one entry under 'Poulterers' in 1961 and 1981, but the entry does not appear in 1971. Another issue is raised with respect to 'Health food store' recorded in 1961 and then 'Health food *shop*' in 1981, 1991 and 2000/1, as if it was missed out in 1971, but then picked up again in 1981 under a slightly (perhaps careless) adjustment to the wording, suggesting possibly a preference for British terminology of a 'shop' as opposed to the more American sounding 'store'. Merging these categories was considered, but for the purposes of continuing the initial exploratory examination of the classifications themselves, it was decided to leave them all exactly as they appear in the directories and *Yellow Pages*.

Now, it is possible that these are data errors. Some categories were perhaps just forgotten about, double counted, or carelessly renamed (see Shaw, 1978 for further examples of this in relation to the trade directories). Data entry mistakes do happen and this is perhaps all that is observed here also. These 'issues' may or may not be highly significant. There is really no way of knowing. But this is precisely why it is safer to leave them in *as they are* rather than trying to clean, merge, code, recode, or measure some new 'made up' category which is *not* there. 'Dirty data' are arguably more valid than 'cleaned up' data, if only because they are what we *do* know *was* recorded and counted, irrespective of the extent to which they are interpreted as meaningful. Even where data entries are presumed or known to be 'dirty' errors, there is value there too. Like patterns in missing data, patterns in data entry errors over time may be interesting. They may for instance, appear more frequently in some years than others. 1971 here, for example, marks itself out from the other years precisely because of its possible data entry errors, signalling perhaps the start of the rift of change that was to become more visible later on. Dirty data provide clues to the context in which the data were constructed and it is in accessing this context that the numbers themselves can be better understood.

In any case, it is worth noticing the 'shape' of the shading. Like the histogram, it is the *overall* distribution of the data that is perhaps more important than the precise details of each row. 'New' categories appear each decade. In fact, at each decade, at least as many categories emerge as disappear. 1981 has the most categories. Again, whether or not this is significant is difficult to tell, but what is interesting is that this is also the year in the census data that shows important socio-economic 'ripples' of change (Uprichard, 2005), suggesting that around this time, change was taking place locally across multiple dimensions. By 2000/1, there are nineteen categories compared to just nine in 1951.

It is possible that there is something about the category itself that makes some survive longer than others. One of the anonymous reviewers of this article, for instance, pointed out that 'it seems to be broader, more general, categories that survive' and that we might expect this 'since they are going to find it easier to collect members than more narrowly defined categories, *ceteris paribus*'. How one defines whether a category is 'broader' or 'more narrow' is debatable, but the point remains an interesting one for of course if there is a pattern in the

level of abstraction to do with whether categories survive or do not survive over time, that would surely have implications on how one might try to predetermine possible future change and continuity as well. Be this as it may, there is certainly a need to do more comparative work within other cities in the UK and elsewhere to know what all this may mean.

That said, if we think of the 'food retail industry in York' as the case and the categories of food retail outlets as the variables, then this quantitative 'growth' in the number of categories from nine in 1951 to nineteen in 2001 – but a total of 46 categories between 1951–2001 – may point not only to the need for polythetic descriptions of the case, but that more attributes are actually necessary the more complex the case becomes. Conversely, the qualitative continuity in certain categories is of interest also. Why, for example, is it that 'Butchers' is the only surviving category throughout 1951–2001? Would this be the same in other cities? Is there something particular about the business that makes it meaningful to the city itself?

Pig farming and pork butchers are indeed renown in the Yorkshire region, and after speaking with a couple of local butchers to ask them whether they had any comments about the local meat industry, it quickly became clear that there was a sense that the meat industry in general had changed considerably, now making it increasingly difficult for them to 'specialize' in particular meat products (such as 'pork' or 'poultry' which are other meat-related categories in 1971, and in 1961 and 1981 respectively). There was also a point made about butchers in York being very much a family enterprise, which had therefore maintained a sense of continuity in ownership as the family business was passed down the generations. Of course, two conversations are by no means sufficient to make confident interpretations, but examining category changes raises other questions, which are precisely about understanding the *local context* in which the categories are situated. Using 'dirty data' allows for the historical contingency of the qualitative changes *to* the categories to be investigated, which ultimately underpin the quantitative changes also – precisely the kind of approach that authors such as Abbott (2000), Bowker and Star (2002) and Hacking (1995, 1999, 2002) advocate.

A word of caution, however. Let us be very clear about the quality of the data used here: neither the *Kelly's Directories* nor the *Yellow Pages* are perfect records. On the one hand, they share similar, but not identical geographies, although the geographical changes are relatively small. Technically speaking, though, the available data on York does not allow for detailed like-for-like longitudinal comparisons between the time-points. Yet even if the geographies were comparable, they probably would not be as useful as might appear anyway, since the administrative geography does not reflect the growing city anyway. Indeed, when it comes to studying change of a dynamic entity, it is oxymoronic to strive for like-for-like comparisons anyway. The very fact that something is understood as *dynamic* implies that over time it is also understood to have changed, so seeking for like-for-like comparisons of a changing thing arguably needs to be abandoned in longitudinal research anyway. On the other hand,

there is no way of knowing that the listed entries represent *all* the outlets under any category. It must be assumed that there are sampling errors within each directory, but then we also need to assume that these errors are of a similar kind in each directory, making the 'dirty data' nevertheless meaningful. However imperfect they are, both sources provide the most accessible source and to this extent they offer similar advantages and disadvantages.

Whilst it is important to be aware of the various issues within 'dirty data', from an ontological and epistemological perspective, *the data as they are available* remain informative. The key is to move away from the notion of the variable as a measure of a thing in itself and instead towards the idea of the variable as a *trace* – a left-over mark, if you like – of that dynamic thing which it measures (Byrne, 2002). Byrne explains:

> The dynamic systems which are our cases leave traces for us, which we can pick up. We can, as it were, track them, and we can infer from the traces what it was that left them. From the traces we can reconstruct a version of the real entities *and* of the relationships among those entities *and* of the emergent forms which are the product of and the producers of the relationships among those entities. We can glimpse the entities and the systemic relationships among the entities. (Byrne, 2002: 36)

Note though that there are time and space issues intrinsic to 'variate traces'. Byrne does not say anything explicit on this, even though they are implicit in his work elsewhere. Three are important to us here. First, there is always a 'disjuncture' between the temporality of the data and its production. In other words, the category is likely to only ever be recorded *after* it has come into being. So even though a new category might 'appear' in the Directories, we can be sure that people were using this category at ground level in York beforehand, even if we cannot determine when it first appeared. Of course, it may be that some categories were used more widely than others, but the point here is what is counted becomes a category. Second, it is worth paying attention to the temporality of traces, for example duration – this might show 'grooves' in which particular trajectories are 'stuck'. Alternatively it could mark a possible future phase shift. When the last butcher closes in York, this will mark a quantitative shift in butchers, but it will also mark a symbolic qualitative shift in how food retail outlets in York have changed. Third, the changing spatiality intrinsic to the traces is important. That is, the geography of each of the categories is not insignificant; even the geography of the resource is significant. The York Directories, for instance, which were in use between 1947 and 1975, and focus primarily on York ward, were replaced by the Yellow Pages, which include many other surrounding areas too, reflecting the changing nature of what was deemed to be 'local' or 'accessible' to the York resident. In other words, as the city of 'York' grew so did the 'pages' from which local residents accessed services and facilities; bigger and thicker sources also mean the bigger and more diverse geography of 'York'.

What we have here, then, as we do with any measures over time, is a complex interaction going on between the thing that is counted, how that thing is per-

ceived, how it changes in absolute terms and how it is perceived to have changed – at the same time the context in which that thing is changing (or not) may or may not be changing as well. Conceptualizing measurable tangible change as Byrne's 'variate traces' is helpful in grasping some of this complexity, especially with regards to understanding what it is that is measured when measuring quantitative change. In some ways, it has been phrased in different ways since Plato's allegory of the cave, for at its heart lies the tension between the epistemological versus the ontological complexity of the world and our efforts to relate one to the other – or more accurately perhaps, it has to do with the extent to which one believes it is possible to abstract one from the other. Byrne's point is reminiscent of other authors (eg Cilliers, 2001; Bhaskar, 1998), who each argue, albeit in different ways, that to think of the epistemological realm (which in this case has to do with quantitative descriptions) and the ontological as independent of one another, is a mistake. This is because, as Cilliers (2001) spells out, there is 'a constant to and fro between them in which our models and, especially in the case of the human sciences, the world itself is transformed'. Cilliers continues:

> Boundaries are simultaneously a function of the activity of the system itself, and a product of the strategy of description involved. In other words, we frame the system by describing it in a certain way (for a certain reason), but we are constrained in where the frame can be drawn. The boundary of the system is therefore neither purely a function of our description, nor is it a purely natural thing. (Cilliers, 2001: 141)

In other words, the measures are representations *both of the system and a representation of our description of it*. This is a crucial point and is central to this argument. Simply put, it is important to think about the data as re-presentations of the re-organization of social life. That said, Bhaskar's epistemological logic about measurement is also worth citing here, precisely because it affirms the sequential ordering of what variables are or are not representing:

> To be is *not* to be the value of a variable; though it is plausible (if, I would argue, incorrect) to suppose that things can only be *known* as such. For if to be were just to be the value of the variable we could never make sense of the complex processes of identification and measurement by means of which we can sometimes represent some things as such. *Knowledge follows existence*, in logic and in time; and any philosophical position which explicitly or implicitly denies this has got things upside down. (Bhaskar, 1998: 29, italics added)

That is, in order to obtain a description of a thing, that thing first needs to exist, and then we need to become sufficiently aware of it in order to know about it in order to measure its frequency as a variable. Conversely, a variable needs to exist first as a variable in order to know more about it and to understand what it represents. This is not to say that knowledge and existence are not recursively constituted or that variables do not have impact on the things that they measure. Rather, it is to underscore the ontological dimension of both variables and the things that they describe and that in order for a variable to come into being, the thing it describes necessarily is already perceived to exist. Of course, once a

variable comes into being then it too takes on both an ontological and episte-mological dimension which impact on both the ontology and the epistemology of the thing that it describes, and so on. But either way, knowledge follows existence.

The reason Bhaskar's point about existence preceding knowledge is impor-tant to bear in mind with respect to methods of understanding change in the social world is because it assumes that despite a constant recursivity between a thing and how it is measured, particularly over time, that relationship is asym-metrical. For example, the reason 'bakers' were counted was that they were there to count in the first place (and not because they were counted that they existed). This is the case, even though once 'bakers' were formally counted, it was also possible to know more about them. In other words, there is always an assumed temporal direction between existence and knowledge, between a thing and the variables that describe it. Hence the time-lag between variate traces and the thing it represents. Like the tracks in the sand signalling the passage of a particular kind of animal, variables are necessarily better at telling us what has happened (in the past) compared to it happening (in the present) or what will happen (in the future). Using variables for predictive purposes becomes a tall order once the temporal nature of the variable is acknowledged.

Guérois and Paulus (2001) raise this issue in an interesting paper about the history and chronology of the main urban geographies in French census and administrative statistics (eg *l'unité urbaine* or *l'aire urbaine*, etc.). They argue that the timing of when new terms and concepts become officially recognized coincides – *and is a direct consequence of* – what is happening directly at ground level, often as a widespread phenomenon across cities and urban regions throughout France. Similarly, the records of change in the food retail outlets reflect something about what was currently happening at the ground level around that time.

As Desrosières (2000: 176) puts it in relation to the construction and selection of variables, the categories were 'chosen and constructed because they were 'judged socially to be of social importance'. In other words, at a given moment in history society judges that 'something' is an issue for intervention by that same 'society' through action taken by its public authorities and its institutions.' There is much value, therefore, in employing 'dirty data' to explore longitudinal change precisely because they also contain within them information about political judgements of the actors that were involved in its construction. The importance of actors remains an implicit component in this kind of quantitative data analysis.

Re-thinking longitudinal quantitative research

The dialectical relationship between the empirical description of the object of study (the realm of epistemology) and the actual object of study (the realm of ontology) is arguably an intrinsic component of any social investigation. It is

worth, therefore, discussing these issues a little further before bringing this article to an end. Certainly, with regards to developing quantitative longitudinal descriptions, these debates tend to be played out in the epistemological-onto-logical interface relating to 'measurement', 'classification', and of course the subsequent operationalization of the 'variable' and the 'case'. The point being that quantitative data are socially constructed, and just as for any social construction, attention needs to be paid to the processes involved in the construction of quantitative data.

This is not a new argument; it has been well presented elsewhere (eg Bateson, 1984; Cicourel, 1964; Hindess, 1973; Irvine *et al.*, 1979). However, even though it is a well-known perspective, it is strangely overlooked in *longitudinal* quantitative research. Yet when all the issues involved in quantitatively describing change over time are put together, it becomes a wonder that any internal validity within measurement systems are left at all. For instance, the number system and the property to be measured need to 'match up'. That is, there needs be an isomorphism between a number system and a property to be measured whereby 'a one to one relationship must be obtained between certain characteristics of the number system involved and the relations between various quantities (instances) of the property to be measured' (Togerson, 1958, cited in Cicourel, 1964: 10–11). For example, to say that there are *x*-many 'bakers' assumes a one-to-one relationship between the number system used to count the number of 'bakers' (usually this is the base ten number system, but it could be another) and that this number system can work adequately in relation to what it may mean to count 'bakers'. It is meaningful to employ integers to reflect 'bakers' since it is not possible to have a fraction of a 'baker'. Simple counts, therefore, become one of the most powerful tools to any longitudinal quantitative researcher.

Yet social scientists tend to have a very bad habit of ignoring the fact that number systems also need to be appropriate to the things they measure and focus predominantly on the social construction of that thing. This may be an acceptable omission, except of course that most classifications do not change in the same way as the number systems that measure them. Most of the time, when a classification changes, it rarely does so in this 'crisp' way suitable to the integer number system used to measure them. Instead, most classifications go through a transitional phase where they are 'fuzzy'. Thus, it is difficult to place 'all bakers' in a 'baker' category; some bakers may be a 'kind of bakers, but also sell takeaway food, such as fresh sandwiches'. Likewise, census classifications have evolved over time making it rather difficult to measure key changes with many of the core demographic categories. As Marsh *et al.* (1988: 854) note, the core issue in calculating change between censuses is being 'able to distinguish real trends from artefactual change brought about by altering census definitions'.

Consequently, in practice, Cicourel argues, measuring the social world involves measurement by fiat, instead of literal measurement. He explains the two ways of measuring the social as follows:

> *Literal* measurement refers to an exact correspondence between the substantive elements and relations under study and the ordered elements and relations of the measurement system. Measurement by *fiat* is an arbitrary or forced correspondence between elements, relations, and operations. (Cicourel, 1964: 225–226)

Although literal measurement is by far the most preferable kind of measurement, it is measurement by fiat that takes place in (and in relation to) the social world. This is not because social scientists are inherently 'bad measurers', but rather the nature of the social world itself renders it difficult to do otherwise. Taking the example of simple frequency counts, before even being able to assign a number to 'bakers', the classification 'baker' itself needs to be identified. Of course, identifying 'bakers' is more than simple nomenclature; it requires a set of (often legal) standards. As Hacking argues, classifications and standards are 'two sides of the same coin': 'classifications are containers for the descriptions of events – they are an aspect of organizational, social and personal memory – whereas standards are procedures for how to do things – they are an aspect of acting in the world' (quoted in Bowker, 2002: 7).

Making valid and reliable longitudinal comparisons over any meaningful length of time becomes, then, extremely problematic, even at the level of acquiring simple changes in counts. Classifications are in and of themselves insufficient in determining either the nature of the isomorphism required to measure them or the descriptions and mechanisms that helped to produce them. The classification 'baker', for example, is not sufficiently precise to indicate how to measure 'change in the number of bakers'. After all, 'change in the number of bakers' depends on the classification ('bakers') and standards that litigate what 'bakers' are both real in their consequences and in themselves. Hence, to measure 'change in the number of bakers' requires the definition of 'baker' to remain static over time, which in turn assumes that all things being equal, to be a 'baker', whatever that might involve, is both epistemologically *and* ontologically constant, which in the context of a dynamic social world is unlikely.

The changing classifications of any object of study themselves become meaningful, then, so long as it is recognized that: (a) the number system used to count a particular classification is itself a social construct, which (b) is only appropriate to that object of study if there is also an isomorphism between (c) the construct of that classification and the number system itself. In addition, following Bowker and Starr (2002), (d) the ways in which these classifications work and how they are interpreted is subject to historical change, which is (e) also regulated by standards which maintain or force changes to those classifications. All these points work together and have a number of implications for what is measured, how measurement is interpreted, and the re-ordering of the social world.

All that said, it is worth bearing in mind Bhaskar's (1998: 29) point that 'Knowledge follows existence, in logic and in time' and that however classifications are constructed, the ways in which they emerge and are subsequently made

to hold has much to do with their ontological properties as well. Thus, following Bhaskar (1979), it is because 'bakers' exist at all and are real in their consequences that they *can* be classified and maintained according to a set of standards and that they can also be identified and counted over time (though that they can be handled in this sort of way may be a contingently necessary condition for our knowledge about 'bakers'). However, the meanings associated with the classification and the underlying object, 'bakers', or any other classification or object for that matter, is subject to historical change as well. In turn, therefore, in order to examine something as relatively simple as 'change in number of bakers', it is also necessary to consider other related classifications that may or may not have helped to create, maintain or destroy them too. As Abbott (2000: 226) puts it, 'looking at the social world not in terms of simple variables but in terms of complexes of variables may be not only more important than we think, but may be all we are able to do'.

Conclusion

Using 'dirty data' to explore qualitative changes to classification systems and quantitative change within classifications over time is meaningful for two main reasons. First, it is at that level of abstraction that further patterns of change and continuity may be empirically captured. Second, any patterns within the trajectories of the categories themselves may point to shifts in the rate of change, the momentum involved in that change, and dimensions of change that may or may not be path dependent. 'Dirty data' – that is, using the categories as they exist *without* any recoding or merging etc. – is meaningful when exploring longitudinal change precisely because they allow for both quantitative and qualitative changes to be explored *together*. Leave the data and categories alone, even if this creates more mess in practice!

The upshot of all of this is that we need to rethink the kind of social research we are doing. As Abbott wrote over a decade ago:

> A quiet revolution is underway in social science. We are turning from units to context, from attributes to connections, from causes to events. The change has many antecedents: the exhaustion of our old paradigm, our inherent desire for change, the power of computers. It also has many consequences: new areas for empirical work, new methodologies, rediscovery of important old theories. (Abbott, 1995: 93)

We might add to this list that there is a shift to focus on the case rather than the variable and using variables differently to classify cases. As Byrne (2002: 127) suggests, 'It is as much a matter of reinterpreting what we are doing as of doing something differently'. The methodological revolution in the social sciences is here and it is already happening. Let us no longer be quiet about it, but rather discuss it openly and honestly. The social world has changed and so too must our ways of knowing change with it.

Acknowledgements

References

Abbott, A., (1995), 'Sequence analysis: new methods for old ideas', *Annual Review of Sociology*, 21: 93–113.

Abbott, A., (2000), 'Reflections on the future of sociology', *Contemporary Sociology*, 29(2): 296–300.

Allen, P., (1997), *Cities and Regions as Self–Organising Systems: Models of Complexity*, Amsterdam: Gordon and Breach Science Publishers.

Bateson, N., (1984), *Data Construction in Social Surveys*, London: Allen and Unwin.

Batty, M., (2005), *Cities and Complexity: Understanding Cities with Cellular Automata, Agent–Based Models, and Fractals*, London: MIT Press.

Bhaskar, R., (1975), *A Realist Theory of Science*, London: Version.

Bhaskar, R., (1979), *The Possibility of Naturalism*, 3rd edn, London: Routledge.

Bhaskar, R., (1998), 'Philosophy and scientific realism', in M. Archer, R. Bhaskar, A. Collier, T. Lawson and A. Norrie (eds), *Critical Realism: Essential Readings*, London: Routledge.

Bowker, G., (2002), 'How things work', paper posted at the eScholarship Repository, University of California, available at: http://repositories.cdlib.org/ucsdcomm/bowker.

Bowker, G. and Star, S.L., (2002), *Sorting Things Out: Classification and its Consequences*, Cambridge, MA: MIT Press.

Byrne, D., (2002), *Interpreting Quantitative Data*, London: Sage.

Byrne, D. and Ragin, C. (eds), (2009), *Handbook of Case–Based Methods*, London: Sage.

Caren, N. and Panofsky, A., (2005), 'TQCA: a technique for adding temporality to qualitative comparative analysis', *Sociological Methods & Research*, 34: 147–172.

Chaplain, M., Singh, G. and McLachlan, J., (1999), *On Growth and Form: Spatio–Temporal Pattern Formation in Biology*, Chichester: Wiley & Sons.

Cicourel, A., (1964), *Method and Measurement in Sociology*, New York: Free Press.

Cilliers, P., (2001), 'Boundaries, hierarchies and networks in complex systems', *International Journal of Innovation Management*, 5: 135–147.

Desrosières, A., (1998), *The Politics of Large Numbers*, Cambridge, MA: Harvard University Press.

Desrosières, A., (2000), 'Measurement and its uses: harmonization and quality in social statistics', *International Statistical Review*, 68: 173–187.

Desrosières, A., (2001), 'How real are statistics? Four possible attitudes', *Social Research*, 68(2): 339–355.

Dupré, J., (2006), 'Scientific classification', *Theory, Culture and Society*, 23: 30–32.

Elliott, J., (2005), *Using Narrative in Social Research*, London: Sage.

Gilbert, N. and Troitzsch, K., (1999), *Simulation for the Social Scientist*, Buckingham: Open University Press.

Guérois, M. and Paulus, F., (2001), 'Commune centre, agglomération, aire urbaine: quelle pertinence pour l'etude des villes?', paper presented at *Géo–évènement* conference, Paris, April.

Hacking, I., (1995), 'The looping effects of human kinds', in D. Sperber, D. Premack and A. J. Premack (eds), *Causal Cognition: A Multidisciplinary Debate*, New York: Clarendon Press.

Hacking, I., (1999), *The Social Construction of What?* Cambridge, MA: Harvard University Press.

Hacking, I., (2002), 'How "natural" are "kinds" of sexual orientation?' *Law and Philosophy*, 21: 335–347.

Hindess, B., (1973), *The Use of Official Statistics in Sociology: A Critique of Positivism and Ethnomethodology*, London: Macmillan.

Hosmer, D., Lemesow, S. and May, S., (2008), *Applied Survival Analysis: Regression Modeling of Time–to–Event Data*, 2nd edn, New Jersey: John Wiley and Sons.

Irvine, J., Miles, I. and Evans, J. (eds), (1979), *Demystifying Social Statistics*, London: Pluto Press.

Jopson, J. and Reeder, A., (2008), 'An audit of Yellow Pages telephone directory listings of indoor tanning facilities and services in New Zealand, 1992–2006', *Australian and New Zealand Journal of Public Health*, 32(4): 372–377.

Marsh, C., Arber, S., Wrigley, N., Rhind, D. and Bulmer, M., (1988), *'Research policy and review 23': The view of academic social scientists on the 1991 UK Census of Population*, London: Economic and Social Research Council.

Menard, S., (2002), *Longitudinal Research*, London: Sage.

Mérenne-Schoumaker, B., (1996), *La Localisation des Services*, Paris: Nathan.

Mills, C.W., (1959), *The Sociological Imagination*, London: Oxford University Press.

Ragin, C. and Strand, S., (2008), 'Using qualitative comparative analysis to study causal order: comment on Caren and Panofsky', *Sociological Methods and Research*, 36: 431–441.

Reed, M. and Harvey, D.L., (1996), 'Social science as the study of complex systems', in L.D. Kiel and E. Elliott (eds), *Chaos Theory in the Social Sciences*, Ann Arbor: University of Michigan Press.

Sassen, S., (1994), *Cities in a World Economy*, Thousand Oaks, CA: Pine Forge Press.

Savage, M. and Burrows, R., (2007), 'The coming crisis of empirical sociology', *Sociology*, 41(5): 885–899.

Sayer, A., (2000), *Realism and Social Science*, London: Sage.

Schelling, T., (1971), 'Dynamic models of segregation', *Journal of Mathematical Sociology*, 1: 143–186.

Shaw, G., (1978), 'The content and reliability of nineteenth-century trade directories', *Local Historian*, 13(4): 205–209.

Shaw, G., (1984), 'Directories as sources in urban history: a review of British and Canadian material', *Urban History Yearbook 1984*: 36–44.

Shaw, G. and Alexander, A., (1994), 'Directories and the local historian III: directories as sources in local history', *Local History*, 46: 12–17.

Singer, J. and Willett, J.B., (2003), *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*, New York: Oxford University Press.

Timmins, G., (1978), 'Measuring industrial growth from trade directories', *The Local Historian*, 13(4): 349–351.

Tryon, R., (1939), *Cluster Analysis: Correlation Profile and Orthometric Analysis for the Isolation of Unities of Mind Personality*, Ann Arbor: Edward Brothers.

Uprichard, E., (2005), 'Studying complex places: change and continuity in York and Dijon', unpublished PhD thesis, Department of Sociology, Durham University: Durham, UK.

Uprichard, E., Burrows, R. and Byrne, D., (2008), 'SPSS as an inscription device: from causality to description?' *Sociological Review*, 56(4): 606–622.

Zukin, S., (1995), *The Culture of Cities*, Oxford: Blackwell.