Measuring Income Mobility with Dirty Data

Frank A. Cowell
and
Christian Schluter

Contents

CASEpaper
CASE/16
November 1998

Centre for Analysis of Social Exclusion
London School of Economics
Houghton Street
London WC2A 2AE
CASE enquiries: tel: 0171 955 6679

Centre for Analysis of Social Exclusion

The ESRC Research Centre for Analysis of Social Exclusion (CASE) was established in October 1997 with funding from the Economic and Social Research Council. It is located within the Suntory and Toyota International Centres for Economics and Related Disciplines (STICERD) at the London School of Economics and Political Science, and benefits from support from STICERD. It is directed by Howard Glennerster, John Hills, Kathleen Kiernan, Julian Le Grand and Anne Power.

Our Discussion Papers series is available free of charge. We also produces summaries of our research in CASEbriefs. To subscribe to the series, or for further information on the work of the Centre and our seminar series, please contact the Centre Administrator, Jane Dickson, on:

Telephone:	UK+171 955 6679
Fax:	UK+171 242 2357
Email:	j.dickson@lse.ac.uk
Web site:	http://sticerd.lse.ac.uk/case.htm

Editorial Note

Frank Cowell is Professor of Economics at the London School of Economics and a member of CASE. Christian Schluter is a Lecturer in Economics at the University of Bristol and an associate of CASE. This paper was written as part of CASE's research on incomes and economic exclusion.

Acknowledgements

This work was partially supported by CASE, and by ESRC grant R000 23 7324. We gratefully acknowledge helpful research assistance by Chris Soares and Hung Wong.

Abstract

We examine the performance of measures of mobility when allowance is made for the possibility of data contamination. We find that "single-stage" indices – those that are applied directly to a sample from a multivariate income distribution – usually prove to be non-robust in the face of contamination. However, "two-stage" models of mobility – where the distribution is first "discretised" into income intervals and then a transition matrix or other tool is applied – may be robust if the first stage if appropriately specified.

1 Introduction

Economists and other social scientists are interested in the movement of personal incomes. Information extracted from individual income histories can be useful in drawing conclusions about the persistence of poverty, some aspects of "openness" of a society or the extent of economic opportunity. The standard tool for characterising this collection of information about personal income streams is the *mobility index*. However if a mobility index is to do this kind of job it is important that (a) the index be appropriately founded on ethical principles or reasonable axioms that capture the meaning of mobility, and (b) that it be reasonably reliable in the face of the imperfections that are inevitably present in even the most carefully collected data: even if one is reasonably confident about a data source, it is obviously inappropriate to assume that the data will automatically give a reasonable picture of the "true" picture of mobility. It is on issue (b) that we focus here: the purpose of this paper is to examine the relative performance of different types of mobility index when one makes allowance for dirty data.[1]

The issues that arise under (b) have an important role to play in the specification and selection of income-mobility indices. Unlike the case of other summary indices in applied welfare economics - such as inequality measures or Social-Welfare Functions there - is not really a good *a priori* case for one mobility index rather than another or one class of indices rather than another. Instead, most

[1]The formal analysis underlying the discussion below is presented in Cowell and Schluter (1998).

commonly-used mobility measures are essentially pragmatic. The behaviour in the presence of data imperfection can be one good guide to the choice of a pragmatic index.

There are several ways in which data imperfections might be introduced into a formal analysis of income mobility. One is to adopt a standard model of measurement error. An alternative - pursued here is to use a model of data contamination: a researcher may anticipate that, because of miscoding and other types of mistake, a proportion of the observations may not "really belong" to the data, and that including them in the working dataset may have a serious impact upon mobility estimates and comparisons. We will consider the performance of some important classes of mobility measures in the presence of this type of contamination.

The central question that we wish to address is whether the properties of mobility indices in conjunction with the characteristics of panel data can give rise to misleading conclusions about income-mobility patterns. Obviously if contamination is in some sense "large" relative to the true data then we cannot expect to get sensible estimates of mobility indices; but what if the contamination were quite small? Could it be the case that isolated "blips" in the data or extreme values could drive estimates of income mobility? We analyse this problem using methods of robust analysis that have become established in other fields.

There is a special difficulty associated with the problem of data contamination in the present context. Pragmatic approaches that are relatively easy to implement in other income distribution problems may be impractical in applica-

tions to issues such as the measurement of mobility. For example, in the analysis of income inequality, it may be appropriate to "trim" data by eye or by algorithm, but the types of rule-of-thumb treatment of outliers that could work well for a univariate problem are likely to be unwieldy in the case of multivariate distributions.

This practical difficulty underlines the importance of understanding the general properties of mobility indices when applied to contaminated data. Our approach has been to establish these properties for two broadly-defined types of index using a simple model of data contamination. Section 2 sets out the basic ingredients of the approach; sections 3 and 4 discuss the first of the two principal types of mobility indices; section 6 discusses the second type of index; section 7 concludes. Finally, some of the relevant literature is briefly surveyed in Appendix A, and a glossary of formulaæ for the measures discussed is given in Appendix B

2 The Approach

Imagine a video-recording of each person's income life-history. If income x is recorded period by period (for example annually) over a fixed time-span then for each individual we would have a multi-period profile of information that may be used as the basis for describing the pattern of personal income mobility within an economy. This can be represented as a vector $\mathbf{x} := (x_1, x_2, ..., x_T)$ where T is the number of periods. Mobility analysis is often described as though there were

just two periods ("before" and "after") so that the problem would be reduced to a bivariate analysis. However, as we shall see this is unnecessarily restrictive for the issues on which we wish to focus.

2.1 Income Distributions

What do we mean by an income distribution? If we were just to be concerned with a snapshot of the economy then this could just be taken as the standard concept from the statistical textbooks, a distribution function, F, where, for any value of income x, $F(x)$ gives the proportion of the population that has an income less than or equal to x. In the present context the basic concept with which we will work is the distribution of individual income profiles $\mathbf{x}$: this can be thought of as a multiperiod income distribution. From this it is straightforward to derive other income distribution concepts such as the cross-sectional income distribution for any one period, or the distribution of discounted lifetime income.

2.2 Mobility

Once we have the idea of a multi-period income distribution in mind the concept of a *mobility index* can be introduced. This is just a summary statistic of the multi-period distribution like measures of location and dispersion that can be used for single-period distributions. There are several alternative approaches to the specification of such indices, which need not detain us here. Specific types of

mobility indices are discussed in Sections 3 to 6 below; for the moment note that the class of indices M be resolved into two important subclasses:

- *Single-stage* indices attempt to make full use of information in a theoretical or empirical distribution F: they are, so to speak, estimated directly from the data.

- *Two-stage* indices are based on a "discretisation" of the distribution: F is pre-processed by converting it into a grouped distribution where the groups, or income intervals, may be exogenously imposed or may be related to statistics of the distribution itself.

For a particular multivariate distribution F we then wish to evaluate the mobility index $M(F)$. However, in most practical applications the "true" distribution will not be known *a priori* but must be estimated from some set of sample data. An estimator of $M(F)$ can then obtained by one of the following two approaches.

1. For the *non-parametric approach* one replaces F with the empirical distribution represented by the sample.

2. In the *parametric approach* one assumes *a priori* that income is distributed according to some pre-specified family of functional forms (for example the family of multivariate lognormal distributions). One then estimates the values of the parameters from the data to obtain one particular member

distribution of the parametric family. Mobility is then estimated using this distribution.

Here we will assume that a complete set of micro-data is available for the T periods, and we focus upon non-parametric methods.

2.3 Data Contamination

Because in practice a mobility index is usually estimated using a sample one should realistically expect that the data may be subject to contamination: for example the misreporting of weekly as monthly income, or the presence in the sample of data points that have been miscoded by the data transcriber (the classic decimal-point error). If one had reason to suspect that this sort of error were extensive in the data sets under consideration the problem of distributional comparison might have to be abandoned because of unreliability. However, it is possible that there might be a fairly serious problem of comparison even if the amount of contamination were fairly small, so that the data might be considered "reasonably clean".

A standard model of this type of problem is as follows. Suppose that the "true" multivariate distribution for which we wish to estimate mobility is F but, because of the problem of data-contamination, we cannot assume that the data actually observed have really been generated by F. What we actually observe instead of F is a mixture of it with some other "alien" distribution $(1-\varepsilon)F+\varepsilon H$

where H is a distribution representing contamination and ε (which lies between 0 and 1) represents the importance of the contamination in the mixture. Clearly if ε in (1) were large we could not expect to get sensible estimates of mobility indices; but what if the contamination were very small?

To address this question for any given mobility statistic M we can use an elementary version of this contamination model. Imagine that the contamination distribution is made up of a set of discrete "blobs" (point masses). In its simplest form we could take the case where there is just one such blob, a single false income observation at $\mathbf{z} := (z_1, z_2, ..., z_T)$. Use $H^{(\mathbf{z})}$ to denote the distribution that consists of just this blob; then instead of the true multiperiod distribution we actually see the mixture given by

$$F_{\varepsilon}^{(\mathbf{z})} = [1 - \varepsilon] F + \varepsilon H^{(\mathbf{z})} \tag{1}$$

For any given mobility index M we could obviously then work out the apparent amount of mobility using the contaminated distribution $M\left(F_{\varepsilon}^{(\mathbf{z})}\right)$. In fact this gives us an appropriate tool for assessing the impact on mobility estimates of an amount of contamination that is "small" in the sense that ε approaches 0. for any hypothetical value of $\mathbf{z}$ we could just differentiate $M\left(F_{\varepsilon}^{(\mathbf{z})}\right)$ with respect to ε and evaluate the result at $\varepsilon = 0$. This is what is known as the *influence function* (IF) for M. It gives us the influence on the estimator M of contamination at the point $\mathbf{z}$, and its value will depend upon the position of $\mathbf{z}$ with respect to

the position of the majority of the data. It indicates whether an estimate of mobility will be stable in the presence of a few "alien" observations in the income profile and, because the IF is the first-order term in the linear expansion of the asymptotic bias of the estimator it will also provide information about the bias of the mobility estimate. If, under the given model of data-contamination (1) IF is bounded for all possible points of contamination $\mathbf{z}$, then the mobility statistic M is said to be *robust*. Of course it is particularly interesting to know whether IF could in practice be unbounded. Typically, this problem of unboundedness can arise when components of $\mathbf{z}$ approach extreme values: in this case a single outlying observation in the income profile could drive the mobility estimate by itself.

Clearly it would be useful to know how the influence function will behave for various types of data contamination for a wide class of mobility indices. So in sections 3 to 6 we consider the problem of characterising IF for certain key types of mobility statistics M.

3 Stability indices

The first subclass of single-stage indices builds upon an extension of inequality analysis. Imagine that income inequality is evaluated for each of the cross sectional distributions $1, 2, ..., T$ and for the distribution of "time-averaged" income for each person's profile over the T periods; the average could be a simple

arithmetic mean, or some kind of weighted average using weights $w_1, w_2, ..., w_T$. If there were absolutely no mobility in the income distribution (although there might be overall income growth) then we would expect inequality in each period's cross-section $I(F_t)$ and inequality of weighted-average income $I(F_{\mathbf{w}})$ to be identical. This is the basis for the idea of a so-called "rigidity" or "stability" index: total income immobility is represented by the above case and departures from this extreme state are assessed using the (as yet unspecified) inequality index I. A typical stability index can be written

$$1 - \frac{I(F_{\mathbf{w}})}{\sum_{t=1}^{T} w_t I(F_t)} \; . \tag{2}$$

Of course each F_t (the cross-sectional distribution in period t) and $F_{\mathbf{w}}$ (the distribution of weighted average income) are derived from the joint distribution function F, and consequently, if the true F is not directly observable and we have to work with a contaminated distribution, these derived distributions will also be affected. Furthermore, because the mobility index (2) is defined as a function of the values of an inequality statistic for several derived distributions of F, its influence function will depend upon the influence function for the inequality index implemented for these derived distributions. Whether the influence function of the stability index is unbounded depends in part on whether the influence functions for the particular inequality measure I are themselves "badly behaved" in the sense that the IFs evaluated for these indices are unbounded. Partly it

depends on whether, in a sense, the "bad behaviour" of the top and bottom lines in (2) happen to cancel each other out. Apart from trivial cases of little practical importance - such as where the contamination just happens to rescale all incomes to the same extent - it is not self-evident whether such a convenient cancellation occurs. Particular instances of stability indices - essentially specific inequality measures - have to be checked individually.

This is not too demanding because there are only a few inequality measures (or families of measures) that are considered as serious candidates for use as the index I in (2). The two principal candidates are:

- *The Gini Coefficient.*
- *The Generalised Entropy Indices.* This broad class includes measures that are ordinally equivalent to (and that have similar statistical properties to) the Atkinson inequality indices and the coefficient of variation.

see Appendix B for the relevant formulas.

Many inequality indices are inherently nonrobust (Cowell and Victoria-Feser 1996), and the two above in particular are indeed so. Furthermore it can be shown that this nonrobustness is not a phenomenon which somehow cancels out in the top and bottom of the fraction in (2). The so-called "stability" indices, are in fact all unstable!

4 "Distance" and related measures

A second principal subclass of single-stage indices interprets mobility in terms of "distributional change" (Cowell 1985) and typically focuses upon measures that incorporate a concept of distance between incomes. As far as the measures' properties in the face of contaminated data are concerned they can be treated in the same manner as the approach of section 3. The distributional-change approach requires restriction to a two-period interpretation of mobility: we will label the two periods $(t-1, t)$. Imagine that someone defines the "distance" $D(x_{t-1}, x_t)$ between the two periods' incomes for a particular individual: mobility may be thought of as some kind of average over the population as the income distribution evolves from $t-1$ to t. There are several commonly-used indices that employ a notion of aggregating the "distance" between individuals' incomes in the two distributions.

- *The Hart index* incorporates the concept of distance that is implicit in the use of the variance of logarithms.

- *The Fields-Ok Index* uses a distance concept is based on the absolute differences of logarithms.

- *The King index* introduces a concept of changing ranks within distributions as well as distance. Furthermore, following Atkinson (1970), King derives axiomatically a social-welfare function consistent with the proposed

mobility measure.

However, all of these, as well as a more general class of distance-based measures based on the generalised entropy concept can be shown to be non-robust. The next section discusses whether this matters.

5 Simulation

We have seen that the single stage measures introduced in sections 3 and 4 are non-robust. *In principle* they might be extraordinarily sensitive in that an infinitesimal amount of contamination in the wrong place could cause the value of the index to be biased away from the value it would adopt for the uncontaminated distribution. It remains to establish how important this issue is likely to be in practice.

To investigate this we could have taken a set of panel data and manipulated some of the observations. However, there is always the danger that some results may be specific to the dataset chosen, and it would clearly be more illuminating to be able to examine systematically the sensitivity of the simulation results to changes in the characteristics of the underlying distribution. Given that our purpose is to examine the behaviour of practical tools, rather than to discuss case studies of particular examples of income mobility, it makes sense to use an experimental "dataset" over which one has some control, but which is not too far away from the sort of numbers one might encounter in practice. We therefore carried

out a simulation on an artificial distribution that has characteristics similar to actual data.

Our baseline distribution was a bivariate lognormal with parameters that would be of the same order of magnitude as empirical estimates for the Michigan Panel Study of Income Dynamics. The PSID income concept used was log annual, unequivalised, real, post-tax, post-benefit income in 1989. These considerations suggested the use of simulated data where marginal distributions were given by Lognormal$(10.25, 0.5)$: the two parameters are respectively the mean and variance of log-income in the assumed distribution. A number of values for the correlation coefficient on log-income were used in the experiment. For a further point of reference it may be interesting to note that if a lognormal were fitted to the BHPS data (annual real net income equivalised using the McClements' scale) for 1991 the result would be closer to Lognormal$(9.5, 0.34)$ and the correlation coefficient on log-income for 1991/92 would be about 0.7.

There are two main types of contamination that may then be modelled within this bivariate framework. Type 1 is that of the "rogue profile": both components of the income profile (x_{t-1}, x_t) are simultaneously contaminated for particular observations in the data-set. Type-2 contamination may be thought of as the "blip" problem: contamination may afflict individual components of the profile. The experiments simulated "decimal-point contamination". This means that a proportion of the observations are recorded as being 10 times larger (in our case) or smaller than they should be: it is one of several typical manual recording errors

	correlation=0.50				correlation=0.75			
contam:	*2.5%*	*5%*	*7.5%*	*10%*	*2.5%*	*5%*	*7.5%*	*10%*
"Stability"indices								
GE(-1)	0.9575	0.9344	0.9223	0.9133	0.9746	0.9615	0.9539	0.9488
	(0.0131)	(0.0117)	(0.0108)	(0.0096)	(0.0093)	(0.0081)	(0.0073)	(0.0068)
GE(0)	0.9328	0.9042	0.8908	0.8816	0.9614	0.9456	0.9377	0.9327
	(0.0123)	(0.0094)	(0.0076)	(0.0061)	(0.0079)	(0.0060)	(0.0047)	(0.0040)
GE(1)	0.9055	0.8811	0.8726	0.8677	0.9456	0.9326	0.9272	0.9245
	(0.0156)	(0.0109)	(0.0089)	(0.0075)	(0.0100)	(0.0073)	(0.0058)	(0.0052)
GE(2)	0.8954	0.8856	0.8846	0.8849	0.9374	0.9337	0.9315	0.9316
	(0.0350)	(0.0304)	(0.0295)	(0.0255)	(0.0255)	(0.0238)	(0.0205)	(0.0196)
Gini	0.9626	0.9437	0.9341	0.9275	0.9803	0.9706	0.9655	0.9622
	(0.0072)	(0.0054)	(0.0042)	(0.0033)	(0.0042)	(0.0031)	(0.0024)	(0.0020)
"Distance"-based indices								
King	1.2146	1.2361	1.2383	1.2386	1.3805	1.4718	1.4843	1.4870
	(0.0632)	(0.0178)	(0.0128)	(0.0104)	(0.1839)	(0.0853)	(0.0592)	(0.0514)
Hart	0.8019	0.6655	0.5811	0.5112	0.7982	0.6643	0.5765	0.5106
	(0.0552)	(0.0478)	(0.0425)	(0.0360)	(0.0642)	(0.0542)	(0.0457)	(0.0414)
Fields-Ok	1.0017	0.9995	1.0008	1.0001	1.0005	0.9999	0.9999	1.0002
	(0.0334)	(0.0329)	(0.0333)	(0.0331)	(0.0347)	(0.0347)	(0.0342)	(0.0349)

Table 1: **Bias in mobility indices resulting from type-1 contamination**

found in practice.

Table 1 reports the experiment for the first type of contamination in a sample of size 500 where the contaminated observations range from 2.5% to 10% of the sample.

shows the contaminated mobility estimate as a ratio of the true value (so an unbiased entry would have the value 1.0000). The figures in parentheses show the standard errors of the estimate. As the top part of the table shows the stability indices based on GE-measures or the Gini index can exhibit substantial downward bias (4 to 13 percent) if the correlation coefficient of the log-income process is low; if the correlation is higher, the bias is reduced (the bias worsens

	correlation=0.50				correlation=0.75			
contam:	*2.5%*	*5%*	*7.5%*	*10%*	*2.5%*	*5%*	*7.5%*	*10%*
"Stability"indices								
GE(-1)	0.9968	1.0090	1.0240	1.0402	1.0130	1.0402	1.0663	1.0929
	(0.0140)	(0.0136)	(0.0140)	(0.0137)	(0.0109)	(0.0118)	(0.0133)	(0.0130)
GE(0)	0.9919	0.9937	0.9978	1.0020	1.0155	1.0330	1.0463	1.0586
	(0.0124)	(0.0114)	(0.0104)	(0.0096)	(0.0087)	(0.0081)	(0.0079)	(0.0078)
GE(1)	1.0090	1.0063	1.0019	0.9960	1.0501	1.0640	1.0665	1.0662
	(0.0144)	(0.0150)	(0.0149)	(0.0146)	(0.0109)	(0.0123)	(0.0131)	(0.0131)
GE(2)	1.0795	1.0707	1.0543	1.0353	1.1592	1.1616	1.1468	1.1289
	(0.0195)	(0.0160)	(0.0170)	(0.0173)	(0.0228)	(0.0159)	(0.0172)	(0.0172)
Gini	0.9904	0.9833	0.9789	0.9745	1.0021	1.0037	1.0041	1.0045
	(0.0090)	(0.0091)	(0.0088)	(0.0082)	(0.0063)	(0.0069)	(0.0068)	(0.0070)
"Distance"-based indices								
King	1.0718	1.0593	1.0658	1.0584	1.2522	1.2503	1.2458	1.2287
	(0.1130)	(0.1114)	(0.1088)	(0.1070)	(0.1623)	(0.1500)	(0.1525)	(0.1534)
Hart	1.1048	1.1864	1.2426	1.2821	1.3138	1.5533	1.7123	1.8551
	(0.0716)	(0.0750)	(0.0780)	(0.0785)	(0.0965)	(0.1059)	(0.1161)	(0.1232)
Fields-Ok	1.0750	1.1534	1.2289	1.3073	1.1159	1.2382	1.3525	1.4772
	(0.0343)	(0.0348)	(0.0355)	(0.0352)	(0.0353)	(0.0351)	(0.0378)	(0.0375)

Table 2: **Bias in mobility indices resulting from type-2 contamination**

with a reduction in the lognormal dispersion parameter). The lower part of the table shows that the bias for two of the distance-related measures can be very large: the King index is biased upwards and the Hart index downwards. This phenomenon persists even where the underlying log-income correlation is high.

The Fields and Ok index appears to perform extremely well in this case, but in a "blip" experiment it performs as badly, or worse than, the King index - see Table 2. The reason for this special behaviour is that, when one works out the influence function for the Fields and Ok index in this second case, simultaneous similarly-sized perturbations of x_{t-1} and x_t will effectively cancel each other out, a phenomenon that is absent from the "blip" model.

6 Transition matrices and related techniques

Income mobility is inherently a complex process, and the attempts at measuring mobility usually involve some attempt at simplifying the underlying model of the process; this *a priori* simplification then has consequences for the way in which sample data are to be handled. The simplifications usually involve "discretisation" of the process, in one or both of two aspects - in state space and in terms of time. The time discretisation is implicit in the discussion of Section 2 where time is treated as distinct periods rather than as a continuous flow.

Two-stage mobility indices involve discretisation of the state space. The transition matrix approach is a standard example of the two-stage approach and permits discussion of a richer pattern of income mobility than can be embodied within a single class of stability or distance-based indices. It might be thought that, as with the distance-based single-stage measures, the two-stage approach makes sense only for cases where $T = 2$; but there is no reason *a priori* why this should be so.

The essential components of the approach are as follows. One specifies a set of income classes (or "bins") into which observations from an empirical distribution are sorted . For simplicity we assume that the set of bins is the same for both periods, although this is not essential to the argument. The *transition probabilities* may then be expressed as the probability that an individual with income in bin i in period $t-1$ will have income in bin j in period t. The *transition*

matrix is formed of these probabilities and the mobility index is then expressed as a function of the transition matrix.

There are two types of issue that concern us here: the general characteristics of the function that is applied to the transition matrix and the specification of the bins. These correspond to two basic components of the impact of a small amount of contamination on the mobility estimate: (i) the effect on overall mobility of a small variation in any one transition probability and (ii) the impact on a given transition probability of the assumed contamination. Component (i) is typically uncontentious: it would be very perverse to specify a mobility criterion that was wildly sensitive to some small change in a transition probability, and we are not aware of any such measures in common use that would have this property. Component (ii) deserves more discussion, and needs to be considered in the light of two alternative practical ways of specifying the "bins".

Exogenous bins. If the income values for the interval boundaries are fixed independently of the data then it is straightforward to show that the influence function for each estimated transition probability is independent of $\mathbf{z}$, the assumed point of contamination. This means that the entire transition matrix must be robust.

Endogenous bins. However, fixing the income boundaries of the bins *a priori* is perhaps rather unusual. It is more common to link the bin boundaries to a proportion of some statistic of the distribution, for example to a proportion of the mean or to one of the quantiles. The expression for the influence function

will now involve terms that are related to the sensitivity of the boundaries to contamination in the data. It is intuitively clear - and straightforward to show formally - that, unless the bin boundaries are parametrised as robust statistics such as functions of quantiles, the transition probabilities estimator suffers from an unbounded influence function. However, given that deciles or other quantiles are known to be robust statistics, then we have the positive result that transition matrices computed on the basis of these statistics will indeed be robust.

Our analysis of two-stage mobility criteria then has two specific conclusions:

1. The robust choice of income classes implies robust estimates of the transition probabilities.

2. The choice of the mobility index from this class of indices is effectively irrelevant from the view point of robustness, and should be guided by other considerations.

7 Concluding Remarks

We have seen that in the presence of data contamination commonly used "single-stage" mobility measures usually behave rather differently from appropriately designed two-stage models of mobility. The problem with single-stage indices comes partly from attempting to make the responsive to all income movements, wherever they may occur on the income scale, partly from the sensitivity of the

functions used in evaluating mobility, be they of the form of adapted inequality measures or distance measures.

The two-stage approach deals with these things separately. In stage 1 we process information: a non-linear function filters out information from parts of the income range; in particular extreme values may be filtered if the data "bins" are function of robust statistics of the distribution. In stage 2 the evaluation and weighting jobs can be performed "safely" by a large number of intuitive and formal algorithms that correspond to different concepts of mobility amongst discrete income or status levels.

References

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory 2*, 244–263.

Bound, J., C. Brown, G. J. Duncan, and W. L. Rodgers (1989). Measurement error in cross-sectional and longitudinal labor market surveys: Results from two validation studies. Working Paper 2884, NBER.

Bound, J. and A. B. Krueger (1989). The extent of measurement error in longitudinal earnings data: Do two wrongs make a right ? Working Paper 2885, NBER.

Cowell, F. A. (1985). The measurement of distributional change: an axiomatic approach. *Review of Economic Studies 52*, 135–151.

Cowell, F. A. (1995). *Measuring Inequality* (Second ed.). Hemel Hempstead: Harvester Wheatsheaf.

Cowell, F. A. (1998). Measurement of inequality. In A. B. Atkinson and F. Bourguignon (Eds.), *Handbook of Income Distribution*, Chapter 2. Amsterdam: North Holland.

Cowell, F. A. and C. Schluter (1998). Income mobility - a robust approach. Distributional Analysis Discussion Paper 35, STICERD, London School of Economics, London WC2A 2AE.

Cowell, F. A. and M.-P. Victoria-Feser (1996). Robustness properties of in-

equality measures. *Econometrica 64*, 77–101.

Fields, G. S. and E. A. Ok (1997, January). A subgroup decomposable measure of relative income mobility. Working paper, Cornell University.

Geweke, J., R. C. Marshall, and G. A. Zarkin (1986). Mobility indices in continuous time Markov chains. *Econometrica 54*(6), 1407–1423.

Hampel, F. R. (1968). *Contribution to the Theory of Robust Estimation*. Ph. D. thesis, University of California, Berkeley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association 69*, 383–393.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley.

Hills, J. R. (1998). Does income mobility mean that we do not have to worry about poverty? In A. B. Atkinson and J. R. Hills (Eds.), *Exclusion, Employment and Opportunity*, Chapter 3, pp. 31–55. LSE, Houghton Street, London: Centre for Analysis of Social Exclusion, STICERD.

Huber, P. J. (1986). *Robust Statistics*. New York: John Wiley.

King, M. A. (1983). An index of inequality: with applications to horizontal equity and social mobility. *Econometrica 51*, 99–116.

Maasoumi, E. and S. Zandvakili (1986). A class of generalized measures of

mobility with applications. *Economics Letters 22*, 97–102.

Maasoumi, E. and S. Zandvakili (1990). Generalized entropy measures of mobility for different sexes and income levels. *Journal of Econometrics 43*, 121–133.

Shorrocks, A. F. (1978). Income inequality and income mobility. *Journal of Economic Theory 19*, 376–393.

Shorrocks, A. F. (1993). On the Hart measure of income mobility. In M. Casson and J. Creedy (Eds.), *Industrial Concentration and Economic Inequality*. Edward Elgar.

A Notes on the Literature

Because the subject matter of this paper is fairly technical it may be useful to give a brief overview of some of the relevant literature. The problem of measurement error in mobility analysis is discussed in Bound et al. (1989) and Bound and Krueger (1989).The alternative approach to the "dirty data problem" - that of modelling contamination using the concept of robustness - is based upon the work of Hampel (1968, 1974), Hampel et al. (1986), Huber (1986): their insights have been applied to a wide range of statistics with economic and statistical applications. The relationship between the measurement error approach and the robustness approach to imperfections in the data is discussed in Cowell (1998) in the context of income inequality.

The principal developments of stability analysis are attributable to Shorrocks (1978) and Maasoumi and Zandvakili (1986, 1990). For a general discussion on the use of inequality measures see Cowell (1995). The Hart mobility index is discussed extensively in Shorrocks (1993); the other distance-based indices are introduced in Fields and Ok (1997) and King (1983).

As the text stresses, two stage mobility indices do not, in principle, have to be discussed in terms of the simplified two-period model, though this makes the analysis very convenient of course. One of the few authors who has attempted to deal with multiperiod generalisations of the two-stage concept is Hills (1998). The modification of the approach to continuous time is discussed in Geweke et al.

(1986).

B Formulas

B.1 Inequality indices

In what follows let G be some univariate (single-period) income distribution, and define the generalised mean μ_α as

$$\mu_\alpha(G) = \int x^\alpha dG(x), \tag{3}$$

We can then write μ for the (arithmetic) mean, such that $\mu(G) := \mu_1(G)$. Also define $Q(G;q)$ as the qth quantile for the given distribution G: this is the smallest income x_q such that, for distribution G, $100q\%$ have an income x less than or equal to x_q. Formally $x_q = Q(G;q) := \inf\{x : G(x) \geq q\}$.

The Generalised Entropy class of indices is then given by

$$I_{\mathrm{GE}(\alpha)}(G) = \frac{1}{\alpha^2 - \alpha}\left[\frac{\mu_\alpha(G)}{\mu(G)^\alpha} - 1\right] \tag{4}$$

where the functional α (a real number anywhere between $-\infty$ and $+\infty$) is the sensitivity parameter of the index. For α large and positive the index is sensitive to changes at the top of the income distribution, for α negative the index is sensitive to changes at the bottom of the distribution. At $\alpha = 0$ and $\alpha = 1$ (4)

adopts the form of the so-called Mean-Log-Deviation index and the Theil index respectively. The Gini coefficient can be written as the functional

$$I_{\text{Gini}}(G) = 1 - 2\frac{\int_0^1 \int_{\underline{x}}^{Q(G;q)} x dG(x) dq}{\mu(G)} \tag{5}$$

where $Q(G;q)$ is the qth quantile, defined above.

B.2 "Distance-based" mobility indices

Let x_{t-1} and x_t denote an individual's income in two consecutive periods. The Hart index is formally defined as

$$M_{\text{Hart}}(F) := 1 - r(\log x_{t-1}, \log x_t) \tag{6}$$

where $r(.)$ is the correlation coefficient. The Fields-Ok index is based upon a distance concept using the absolute differences of logarithms:

$$M_{\text{FO}}(F) = c \int \int |\log x_{t-1} - \log x_t| \, dF(x_{t-1}, x_t). \tag{7}$$

King's index can be expressed as

$$M_{\text{King}}(F) = 1 - \left[\frac{\int \int \left(x_t e^{\gamma s(F,\mathbf{x})} \right)^k dF(\mathbf{x})}{\mu_k(F_t)} \right]^{\frac{1}{k}} \quad k \square 1, \; k \neq 0, \; \gamma \geq 0 \tag{8}$$

where $s(F; \mathbf{x}) := \frac{|x_t - Q(F_t; F_{t-1}(x_{t-1}))|}{\mu(F_t)}$is the "scaled order statistic" which captures reranking.