Commercial Income Data: Associations with Health and Census Measures

Philip McLoone

MRC Social & Public Health Sciences Unit
Occasional Paper No 7
May 2002
Commercial Income Data: Associations with Health and Census Measures

Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>2</td>
</tr>
<tr>
<td>Data &amp; Methods</td>
<td>3</td>
</tr>
<tr>
<td>Results</td>
<td>5</td>
</tr>
<tr>
<td>Income distribution</td>
<td>5</td>
</tr>
<tr>
<td>Variation by Local Government District</td>
<td>7</td>
</tr>
<tr>
<td>Variation by Postcode Sector</td>
<td>8</td>
</tr>
<tr>
<td>Associations with Health Measures</td>
<td>10</td>
</tr>
<tr>
<td>Conclusion</td>
<td>14</td>
</tr>
</tbody>
</table>
Summary

CACI Paycheck is advertised as providing estimates of gross income for small areas across the UK. The data consist of numbers of households falling into £5,000 income categories. The data cover Local Council areas, postcode sectors and full postcodes.

1. The methodological report provided by CACI do not explicitly demonstrate how the income estimates are derived. The estimates are based on consumer surveys that required individuals to report annual income.

2. The reported income data have been modelled by demographic and household composition information to provide ‘fitted’ distributions of income for each small area. When information on income for a particular geographical area was sparse CACI employed information from the census to achieve the fitted distribution. The distributions of income at the level of output area are therefore smooth. They represent idealised distributions (or probability distributions) of income within areas. The data do not provide reported income levels.

3. It is unclear whether the final income estimates rely on the fitted association with census information. If it is the case that they do then the derived income distributions are simply a function of census derived variables.

4. Paycheck is available for full postcodes in Britain. However the smallest geographical unit for which census information is made available is Output Area. It is therefore difficult to understand how CACI provide estimates for residents within full postcodes.

5. For postcode sectors the correlation between mean income and Carstairs measure of deprivation was −0.686. When correlations between Carstairs scores and the proportion of households falling below different income bands were explored the strength of the association peaked with the percentage of households falling below £10,000.

6. Comparison of the ability of Carstairs scores and Paycheck data to predict mortality and long-term illness (derived from the census) revealed that Carstairs ‘explained’ around 80% of variation in postcode sector mortality ratios compared to 70% explained by Paycheck. There was no added explanatory power when both Carstairs scores and income (categorised in several ways) were both employed within these analyses. There was no added effect of income above the effect of deprivation. Similar results were found when the analysis was repeated with limiting long-term illness.

7. Compared to an already available deprivation measure whose derivation is publicly available Paycheck performs less well in explaining the variations in postcode sector mortality ratios and a census measure of limiting long term illness.
Introduction

For a large part of the twentieth century the socio-economic position of individuals had often been derived through occupational grade (for example, the Registrar General's social classes). More recently populations resident in small geographical areas have had their socio-economic experience categorised by variables derived from the decennial census. Examples of this include Carstairs and Townsend deprivation scores, which are based on the characteristics of populations resident within postcode sectors, or electoral wards respectively. The Carstairs score is a simple index derived from four census variables. The variables employed in the derivation of the Carstairs index relate to levels of car ownership, unemployment, over-crowding and low social class within each postcode sector. A postcode sector is that geographical area covered by the first five digits of a postcode. An average postcode sector contains 5,000 residents, but there is large variation in population size which can range from around 50 to 20,000 residents.

When census deprivation measures are utilised in health research it is often assumed that they describe or indicate something about material disadvantage in society (either between individuals or between populations) and consequential health experience. Carstairs deprivation index attempts to locate areas (and the populations they contain) on a dimension which reflects the access that people have to material resources which permit individuals to play the roles, participate in relationships and follow the customary behaviour which is expected of them by virtue of their membership in society. Part of the rational of employing car-ownership and low social class within the Carstairs index is justified by suggestions that car-ownership is a useful surrogate for current income, and that low social class indicate earnings at the lower end of the income scale. In part, therefore, census deprivation measures might be expected to indirectly indicate some feature of the income differences that exist between areas. An important question therefore concerns whether conventional measures of socio-economic disadvantage are any better at measuring the effects of material disadvantage on health than that obtained by measuring income alone. A further question is to what extent variations in morbidity and mortality are a consequence of income differences, or a consequence of broader deprivation effects.

In Britain there is little information about how levels of income varies between the populations of small areas, or about the relationship between income and measures of health. The reason why this is the case is that income data are not routinely collected or in the case of our income tax collecting body (the Inland Revenue) rarely made available to researchers. Focused studies, which have utilised self-reported income, acknowledge the inherent limitation that income is often poorly reported. Given the paucity of information about income it was with interest that during 1997 a commercial company called CACI launched a product called PayCheck. The company claims that this provides an estimate of household earnings for every postcode in the UK. Their data modelled gross income and cover income support and welfare.

The purpose of this report is to provide an account of CACI's household income data, it's perceived limitations and potential utility in the analysis of variation in measures of health in Scotland when compared with a commonly used census deprivation measure.
**Data & Methods**

CACI is an international information technology products and services company. Founded in 1962, the company have specialised in developing and integrating systems, software, and simulation products in support of government and commercial groups world-wide. CACI Marketing Systems has been providing demographic and consumer information products such as ACORN (A Classification of Residential Neighbourhoods) for nearly 30 years. The company offers a number of demographic and market analysis databases in a variety of formats: reports, maps, electronic media, software, books, and lists.

In 1997 CACI produced a product called PayCheck which provided an estimate of household income for every postcode in the United Kingdom. The data modelled gross income before tax and covered income from a variety of sources, including income support and welfare. The exact nature of statistical methodology employed by the company to create this product is unclear. Commercial sensitivity mean that the methodology is not explicitly or exhaustively described in their literature. However, it was ascertained that the data used to produce the imputed household income data came from various sources which included lifestyle data, data from the census, the company’s own population estimates, ACORN (the company’s own geo-demographic classification) and market research data.

The base data employed to create estimated household income for postcodes was derived from CACI’s Lifestyle Plus database. This database comprises several million records of which a high proportion contain a figure for annual household income. The lifestyle database appears to have been created via a variety of organisations who have been collecting self completed ‘lifestyle’ questionnaires from individuals who have returned guarantee certificates for ‘white goods’ (i.e. fringes, cookers etc) or have been completed as a result of direct mail to specific individuals.

The company reported that this database contained information for the majority of postcodes in the UK. A detailed account of the numbers and coverage involved was not provided. Because the lifestyle information is self reported there were concerns that systematic variation in response rates between social classes would be apparent. However, the company reported that the spread of respondents was more representative nationally than would have been expected from surveys of this type. This still raises the possibility that the information returned was not representative. The company stressed that responses were not under-represented by respondents in more affluent areas.

The PayCheck data were created by applying a series of weights to each lifestyle record. The weights, which were applied, depended on ACORN type, postal area and household income. Using the combination of this information estimated means, standard deviations and distributions of household income were produced for any area for which a reasonable amount of information was available.

To provide an estimate of income for those postcodes without reasonable amounts of information, and enhance the data produced above, an income model was generated from the existing lifestyle data. When this was not available comparable data from the 1991 census was utilised. This information included marital status, age, age of family, number of children, dwelling type, tenure, number of rooms, number of cars, employment status, ACORN type and postal area.
In essence the PayCheck data expresses the fitted relation between income and the variables listed above when applied to data derived from the census. Equivalent census counts were extracted at small area level for the whole of the UK and the constructed formula was then applied to each area and used to create a set of modelled incomes across the UK. The result was an income directory covering the whole of the UK.

During the later part of 1999 CACI launched a second PayCheck report. This is now described as the largest and most detailed survey of gross household incomes (including income support and welfare) ever undertaken in the UK. The ability of the company to readily update their income estimates presents considerable advantages over traditional deprivation measures, which due to their reliance on the census are generally updated every 10 years.

The data obtained from CACI consisted of a record for each census (1991) output area. An output area is the smallest geographical area for which census information is made available. There were just over 38,000 output areas in the 1991 census and on average they contained 150 residents (approximately 60 households). Each record of the dataset contained 21 categories of income in £5,000 monetary bands, starting from less than £5,000, £5-10,000, £10-15,000 up to £100,000+. Each record provided the number of households falling into each income category. Because the data are imputed each income category did not necessary contain whole numbers of household units. In essence the distribution of income provided by CACI represents a probability distribution. That is to say it provides the probability that a household selected at random within a particular area will fall into a particular income band.

The information provided by CACI was useful in obtaining mode & median income for areas. However, to obtain a measure of mean income, a judgement had to be made about which value of income to take for each income category. Income bands are £5,000 wide. For the purpose of this report the mid-point of each income range was used to estimate mean income. The exception to this was at the extreme categories of income (<£5,000 and £100,000+) when the overall shape of the income distribution was extrapolated. This provided an estimate of the average income levels within these extreme categories, and allowed for a more reliable estimate of mean household income for each area to be calculated.

Associations between the income data, Carstairs measure of area deprivation, standardised mortality ratios (ages 0-64, 1991-95) and self-reported long-term illness (derived from the 1991 census) were then explored across all postcode sectors in Scotland. Poisson multilevel regressions of both health measures with Carstairs scores, average income and the proportion of households falling below each income band were also compared.
Results

Income distribution
As a preliminary it is useful to describe the distribution of the income data. Figures 1 & 2 describe the distribution of household income across Scotland as estimated by PayCheck. The histogram (Figure 1) shows a large skew, with 12 percent of households having a gross income of less than £5,000. 5% of households had an income in excess of £45,000 with 0.1% having an income of over £100,000. The mode income was between £10-15,000 and the estimate of average income was £18,385.

An interesting feature of Figure 1 is the apparent smooth distribution of income. Either side of the mode of the distribution in Figure 1 can be described by a different quartric exponential distribution. Although the underlying distribution of histogram is difficult to quantify, it appears to be a mixture distribution generated from the individual distributions of income imposed at output area level (or possibly at the full postcode level). Even at output level the distribution of income is probably a mixture distribution resulting from the modelling of income at postcode. Figure 3 shows the smooth nature of income at output area level for several areas with small populations. Analysis of these curves suggested that distributions from the exponential quadratic family (for example the Normal distribution) fit the data rather well – this might suggest that at postcode level the distribution employed to model income was the Normal distribution.

Figure 1 – Estimated distribution of household income in Scotland
Figure 2 - Cumulative distribution of household income in Scotland

Figure 3 Distribution of income for selected Census output areas. The figure illustrates the smoothness of the income distribution within very small areas (with small populations).
Variation by Local Government District
Figure 4 explores the data by local government district (LGD) and plots the average household income within each local government district (as defined prior to 1996) against the average Carstairs score for that district. Whilst it may be argued that LGD are internally too heterogeneous a unit to be used in this manner the graph does reveal some interesting features. The first is that in general there is an inverse association between the two measures. The second is the curvi-linear relationship between both measures. As the average Carstairs score increase over the value of around 2 there is no further reduction in average income level. Figure 5 shows this relationship in further detail by detailing the association between the proportion of households with an income below 4 thresholds (£5,000, £15,000, £25,000 and £35,000) and the corresponding Carstairs score for an area. It is obvious from the picture that the most linear association is experience at a threshold of less than £5,000. Each successive category of income shows increasing curvature, partly as a result of the fact that most households in the most deprived areas tend to fall below the higher cut-off levels.

Figure 4 – Average household income in Local Government districts plotted against Carstairs scores
Variation by Postcode Sector

The correlation of average income with Carstairs scores at postcode sector level was -0.686 (Spearman -0.748). Figure 6 plots average household income against Carstairs scores for each postcode sector in Scotland. Each point on the graph represents a postcode sector. The graph displays greater variation in average income towards the most affluent end of the graph, and overall there is a curvilinear relationship between both measures. It is apparent that towards the more deprived end of the chart there are a few unusual postcode sectors – deprived areas with high average incomes.

Figure 7 models this relationship more formally and shows the fitted relationship between deprivation scores and average income within each local government district in Scotland. It is apparent that the fitted lines for each LGD converge towards the deprived end of the distribution. However the graph shows that greater variation exists between local government districts in terms of the PayCheck income data at the more affluent end of the distribution.

The structure of the PayCheck data allows an exploration of the correlation between deprivation and the proportion of households falling below different income bands within postcode sectors. Figure 8 shows that as the income threshold increases the strength of the association rises slightly until a value of <£10,000 is reached, after which the association between both measures rapidly decays, although there is still an association even when the percent of households falling below £95,000 is employed. This figure reveals a smooth decay in the association between the two measures. This is a result of the modelling process used to construct the income data.
**Figure 6** Mean household income in postcode sectors plotted against Carstairs scores

**Figure 7** Relation between Carstairs scores and mean household income in different local government districts
Associations with Health Measures

The income data were examined in relation to standardised mortality ratios and long-term illness ratios. Figure 9 plots age and sex standardised mortality ratios (SMRs) against household income for each of the postcode sectors in Scotland. This figure reveals a large degree of scatter, much of which is transient in nature (random variation). For example, those postcodes with an SMR of zero which are shown on the horizontal axis simply result from the lack of deaths during the time period considered. The line fitted through the graph superimposes the average relationship between both measures, and this is clearly curvilinear. In contrast when the relationship between mortality and deprivation is shown (Figure 10) the relationship is clearly more linear. Figure 10 does however hint at a slight curvature, but whether this is a true reflection of reality is open to speculation. In any event the fitted line does not exhibit the gross curvature seen in Figure 9. Limiting long-term illness (LLTI) displayed similar types of association with each measure.

SMRs and limiting long term illness ratios were modelled independently against income at certain income bands; against average income; and finally modelled with Carstairs scores as the explanatory variable. Figure 11 illustrates the proportion of variation explained as a result of these models. For SMRs the figure shows that little improvement was achieved in the percentage of variation explained as different income thresholds were employed. There was a suggestion of a slight peak for the proportion of household falling below £10,000 and then a gradual fall. When average income was employed the results were similar to those achieved by using income bands. In contrast, Carstairs deprivation scores performed rather well – explaining around 80% of the systematic
Figure 9 Standardised mortality ratios* for postcode sectors plotted against average income.

Figure 10 Standardised mortality ratios for postcode sectors plotted against Carstairs scores.

*Ages 0-64. 1990-92
variation in postcode sector mortality. A similar pattern was seen for the relationship with long term limiting illness, although the proportion of variation explained was slightly smaller across each income group. Carstairs deprivation score was also slightly poorer at explaining variation in this health measure compared to mortality.

Investigating these patterns further, Figure 12 shows for each of the models presented in Figure 11 the percentage of unexplained variation which appears to cluster around larger local government district levels. The figure reveals two rather striking features. The first is that there appears to be larger local government district variation in limiting long term illness compared to mortality and this is not being adequately dealt with by either income measures or the Carstairs deprivation measure. Secondly Carstairs scores appears to explain differences in mortality between local government districts in mortality better than the income data. The percent of unexplained variation in mortality attributable to local government district differences was around 10%. This is one reason why Carstairs scores appear to predict mortality better than the income data.

**Figure 11** Approximate % of variation in postcode sector standardised mortality ratios (SMRs) and limiting long term illness ratios (LLTI) accounted by income bands, mean income & Carstairs scores
Figure 12 % of unexplained variation in SMRs and LLTI associated with local government district differences
Conclusion

This paper provides a summary of CACI’s PayCheck database, and a comparison of the behaviour of this information with Carstairs measure of deprivation when applied two health related measures - limiting long-term illness and mortality. Because of commercial sensitivities CACI, and marketing companies in general, are reluctant to provide detailed methodological information about how they derive their data, whether that is ACORN groupings or in this instance income data. As such the methodology used to derive the income information is not available for scientific scrutiny or rigorous evaluation. In evaluating the data, it was apparent that the overall distribution of income appeared rather 'contrived' in the sense that the information provided contained no raw data but rather expressed fitted relationships between census data and other survey data.

Compared to an already available deprivation measure, whose derivation is publicly available, the income data performed less well in explaining the proportion of variation in postcode sector mortality ratios and a census measure of limiting long term illness. Apart from this observation, the income data seemed to behave as expected in the types of relationships they demonstrated. Curvilinear associations between income and measures of health are well documented – and these are the types of relationship found in the analyses presented here. Why this type of relationship exists is open to question but has serious implications with regard to how associations between health and income are interpreted. Some commentators, for example, have tried to interpret the nature of such associations in causal ways and this has led to discussions about the effects of absolute versus relative disadvantage. For example there is a school of thought, which suggest that for modest increases in income a large decrease in mortality will result among those with low incomes.

On the other hand deprivation scores show more linear associations with health measures. Why should that be the case? In some sense deprivation scores are also 'contrived'. Although the rationale for their creation has been couched in terms of theoretical relationships to material circumstance and wealth, there remains a suspicion – especially with regard to Carstairs scores - that they simply utilise what best predicts mortality. We thus end up in a tautological circle from which it is difficult to escape. While deprivation scores have their political utility in highlighting differences between the most well off and least well off sections of society, they probably explain too much variation in mortality without ‘explaining’ anything at all. They leave little room to explore those other characteristics and processes of populations which may have relevance to health. Partly this is due to the limitations of ecological analyses in general. The PayCheck income data explain less variation in mortality and long term illness but one could argue that they perform better than deprivation scores in the sense that they leave more room to explore these other issues.