Income in the October Household Survey 1994

by

Martin Wittenberg
About the Author(s) and Acknowledgments

Recommended citation


© DataFirst, UCT, 2008

DataFirst, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 5708,
Email: info@data1st.org/support@data1st.org
Income in the October Household Survey 1994

From DataFirst

Martin Wittenberg School of Economics and SALDRU, University of Cape Town January 2008

Contents

1 Introduction
2 The structure of the income data
3 Spikes in the “net income of employee” variable
4 Conversion rates between reporting periods
5 The impact of the imputations
6 Imputations in the “gross income from own activity” variable
7 Conclusion
8 Files to accompany this document

Introduction

Income dynamics in the post-apartheid era are of particular interest to economists, since they relate to the incidence of poverty and inequality and throw light on the operation of the labour market. The 1994 October Household Survey is strategically placed at the beginning of the post-apartheid era. Nevertheless the income information in the survey needs to be treated with considerable caution since much of it is imputed in ways which are questionable.

This document sets out what we have managed to discover about the process of imputation and ways of correcting for these imputations.

The structure of the income data

There are two key places in the questionnaire where income information is solicited. In Question 3.13 the “income from main job” is asked for; while in Question 3.19 the “gross income/turnover for all own account activities” is the subject of enquiry. In both cases provision is made for a Rand amount to be supplied, a series of categories are given and then the respondent is prompted to indicate whether the reporting period is per day, week, month or year.

This “raw” information, however, is not provided in the dataset. Instead the following four variables are given: “Net income of employee (Rand) (calculated per month)”, “Net income of employee (Code) (calculated per month)" both derived from Question 3.13 and “Gross income of employer (Rand) (calculated per month)” and “Gross income of employer (Code) (calculated per month)” both derived from Question 3.19. The calculations seem to have involved at least the following: conversion of daily, weekly and annual data to monthly; imputing information given only in brackets to point values; and correcting gross values to net values. The effect of the imputations has led to a particular structure of the data as supplied, in particular there are pronounced “spikes” within the data. We will discuss the “net income of employee” variable first, before turning to the income from own account activities.

Spikes in the “net income of employee” variable

The consequence of the imputations can be seen in the “spike plot” shown in Figure 1, which gives the information for the “net income of employees” variable. The placement of these pronounced spikes is odd, since they are not located at “nice” round numbers. Indeed the placement is bizarre to say the least: one would not expect around 30% of all individuals who reported being paid daily to have received precisely the same net monthly income of R1364. This is not an artefact of small numbers either: this spike corresponds to 172 individuals.

Figure 1 Spikes in the "Net income of employee calculated per month" variable

One potential culprit for these strange numbers might be the conversion from gross income to net income. Individuals that reported round numbers might have had these spikes shifted by corrections in the conversion process. Figure 2 shows, however, that this is not the case. Individuals that claimed to have reported net income, i.e. where no conversion was necessary, were the ones were the data is spiky. The conversion from gross to net income seems to have smoothed over the spikes instead of creating them.
Figure 2 The income spikes are evident only for people claiming to report net income

Looking at the spikes of the individuals that reported monthly income data (bottom panel of Figure 2) is quite suggestive – particularly when the income brackets are borne in mind. It seems as if there is a “spike” corresponding to each income bracket. The most logical explanation is that the spike corresponds to individuals that reported an income category only and that were then placed at these particular values. The “spiky” nature of the data therefore potentially allows us to separate out the point information from the categorical data: with the caveat that the process of correcting the data among those reporting “gross income” seems to have removed any spikes that might have been there initially.

We can see the impact of the conversion in Figure 3, where we have added the “income deductions” back to the net income variable to create an estimate of what the individuals would have reported originally. The spikes magically reappear. In fact, as Figure 4 shows (when compared to Figure 1) the spikes are precisely at the same points.

Figure 3 The spikes re-emerge when the deductions are added back to the "net income" figures.

Conversion rates between reporting periods

Figure 4 The spikes for individuals reporting "gross income" are at the same points as for those reporting net income
Before we analyse the location of the spikes a bit further, it is important to consider the way in which Statistics South Africa seems to have converted between reporting periods. In the case of daily data the pattern emerges when one looks at a tabulation of the actual values. Ignoring spikes at strange values, there is a periodic pattern of spikes at multiples of 110. This kind of pattern would be produced if the original “round” Rand incomes (i.e. spikes at multiples of five and ten) had been multiplied by 22. This multiplication factor in turn makes sense, since there are roughly twenty-two working days in every month.

```
.tab salary_r if income_i==1
```

In the case of the weekly data the pattern is harder to pick up, but there is a pattern nonetheless with spikes at 130, 260, 390, 520 and so on. This pattern would be produced if round incomes had been multiplied by \(\frac{4}{3}\). This multiplier makes sense if one remembers that there are 52 weeks in a year and also 12 months. But 52/12 is equal to \(1\frac{1}{3}\).

**The location of the spikes**

The strongest confirmation that these numbers indeed give the conversion rates comes by regraphing the spikes in terms of the original reporting periods, i.e. dividing the daily income figure by 22 and the weekly one by \(\frac{4}{3}\). In the resulting spike plot (see Figure 5) the spikes are now at precisely the same position.

```
Income per Month| Freq | Percent | Cum. |
----|------|--------|-----|
  0 |  15  |  2.38  |  2.38 |
 33 |   1  |  0.16  |  2.54 |
 70 |   1  |  0.16  |  2.7  |
110|   9  |  1.43  |  4.13 |
132|   5  |  0.79  |  4.92 |
147|   1  |  0.16  |  5.08 |
150|   1  |  0.16  |  5.24 |
154|   1  |  0.11  |  5.35 |
170|   1  |  0.16  |  5.51 |
172|   2  |  0.32  |  6.83 |
176|   8  |  1.27  |  8.1  |
186|   1  |  0.16  |  8.26 |
210|   1  |  0.16  |  8.41 |
220|  31  |  4.92  | 13.33 |
223|   1  |  0.16  | 13.49 |
242|  16  |  2.54  | 16.03 |
264|  15  |  2.38  | 18.41 |
280|   1  |  0.16  | 18.57 |
286|  17  |  2.7   | 21.27 |
300|   1  |  0.16  | 21.43 |
308|   7  |  1.11  | 22.54 |
330|  20  |  3.17  | 25.71 |
336|   1  |  0.16  | 25.87 |
360|   1  |  0.16  | 26.03 |
374|   1  |  0.16  | 26.19 |
396|   2  |  0.32  | 26.51 |
400|   1  |  0.16  | 26.67 |
418|   1  |  0.16  | 26.83 |
440|  13  |  2.06  | 28.89 |
453|   1  |  0.16  | 29.05 |
484|   3  |  0.48  | 29.52 |
528|   2  |  0.32  | 29.84 |
550|   5  |  0.79  | 30.63 |
580|   1  |  0.16  | 30.79 |
600|  21  |  3.33  | 34.13 |
682|   2  |  0.32  | 34.45 |

In the case of the weekly data the pattern is harder to pick up, but there is a pattern nonetheless with spikes at 130, 260, 390, 520 and so on. This pattern would be produced if round incomes had been multiplied by \(\frac{4}{3}\). This multiplier makes sense if one remembers that there are 52 weeks in a year and also 12 months. But 52/12 is equal to \(1\frac{1}{3}\).

**The location of the spikes**

The strongest confirmation that these numbers indeed give the conversion rates comes by regraphing the spikes in terms of the original reporting periods, i.e. dividing the daily income figure by 22 and the weekly one by \(\frac{4}{3}\). In the resulting spike plot (see Figure 5) the spikes are now at precisely the same position.

```
Income per Month| Freq | Percent | Cum. |
----|------|--------|-----|
  0 |  15  |  2.38  |  2.38 |
 33 |   1  |  0.16  |  2.54 |
 70 |   1  |  0.16  |  2.7  |
110|   9  |  1.43  |  4.13 |
132|   5  |  0.79  |  4.92 |
147|   1  |  0.16  |  5.08 |
150|   1  |  0.16  |  5.24 |
154|   1  |  0.11  |  5.35 |
170|   1  |  0.16  |  5.51 |
172|   2  |  0.32  |  6.83 |
176|   8  |  1.27  |  8.1  |
186|   1  |  0.16  |  8.26 |
210|   1  |  0.16  |  8.41 |
220|  31  |  4.92  | 13.33 |
223|   1  |  0.16  | 13.49 |
242|  16  |  2.54  | 16.03 |
264|  15  |  2.38  | 18.41 |
280|   1  |  0.16  | 18.57 |
286|  17  |  2.7   | 21.27 |
300|   1  |  0.16  | 21.43 |
308|   7  |  1.11  | 22.54 |
330|  20  |  3.17  | 25.71 |
336|   1  |  0.16  | 25.87 |
360|   1  |  0.16  | 26.03 |
374|   1  |  0.16  | 26.19 |
396|   2  |  0.32  | 26.51 |
400|   1  |  0.16  | 26.67 |
418|   1  |  0.16  | 26.83 |
440|  13  |  2.06  | 28.89 |
453|   1  |  0.16  | 29.05 |
484|   3  |  0.48  | 29.52 |
528|   2  |  0.32  | 29.84 |
550|   5  |  0.79  | 30.63 |
580|   1  |  0.16  | 30.79 |
600|  21  |  3.33  | 34.13 |
682|   2  |  0.32  | 34.45 |

In the case of the weekly data the pattern is harder to pick up, but there is a pattern nonetheless with spikes at 130, 260, 390, 520 and so on. This pattern would be produced if round incomes had been multiplied by \(\frac{4}{3}\). This multiplier makes sense if one remembers that there are 52 weeks in a year and also 12 months. But 52/12 is equal to \(1\frac{1}{3}\).

**The location of the spikes**

The strongest confirmation that these numbers indeed give the conversion rates comes by regraphing the spikes in terms of the original reporting periods, i.e. dividing the daily income figure by 22 and the weekly one by \(\frac{4}{3}\). In the resulting spike plot (see Figure 5) the spikes are now at precisely the same position.

![Figure 5 Converting the monthly data back to the "original" form produces spikes at the same location](http://data1st.com.uct.ac.za mediawiki/index.php/Income_in_the_October_Household... 2008/09/23)
This evidence suggests fairly strongly that individuals that merely ticked a category without providing a Rand income were placed at these values within the brackets and this amount was then converted to a monthly figure, using the appropriate conversion rate. That begs, of course, the question why these particular values should have been used for the imputation in the first place.

That question turns out not to have any easy answers. It seems natural to suppose that these numbers would be based on the means or medians of those individuals actually supplying point estimates within each category. Nevertheless trying all sorts of permutations of these (even geometric means) we could not reproduce these numbers.

The impact of the imputations

Since the spikes are well defined it is straightforward to remove them. This may remove also the odd individual who coincidentally gave precisely that figure, but this error is likely to be negligible. Taking out the spikes has a dramatic effect. Table 2 shows fully 54% of all the observations at one or other of the spikes! The same table also shows that there is a systematic negative relationship between reporting actual information and income received. Note that category 12 is not the highest income bracket (R33 000+) had a mean monthly income of R6231, instead of the R4292 assigned them by the imputation algorithm. The net effect of these different biases for individuals with annual income in the top bracket (R33 000+) had a mean monthly income of R660 or R440 as the appropriate imputed value, rather than the R1364 actually imputed) figures shown in Table 3 show significant differences. For instance the mean income among daily paid workers that fell into the lowest bracket was R30 and controlling for clustering) would reject the idea that the two means are equal. Table 4 shows that there is a noticeable impact of the imputation process on average numbers.

Table 2 Proportion of data imputed by income bracket

<table>
<thead>
<tr>
<th>Bracket:</th>
<th>Monthly</th>
<th>Daily</th>
<th>Weekly</th>
<th>Annual</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1-R99</td>
<td>52</td>
<td>1364</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>R100-R199</td>
<td>143</td>
<td>3146</td>
<td>620</td>
<td>12</td>
</tr>
<tr>
<td>R200-R499</td>
<td>308</td>
<td>6776</td>
<td>1335</td>
<td>26</td>
</tr>
<tr>
<td>R500-R999</td>
<td>726</td>
<td>15972</td>
<td>3146</td>
<td>61</td>
</tr>
<tr>
<td>R1000-R1999</td>
<td>1386</td>
<td>30492</td>
<td>6006</td>
<td>116</td>
</tr>
<tr>
<td>R2000-R3999</td>
<td>2679</td>
<td>59398</td>
<td>11609</td>
<td>223</td>
</tr>
<tr>
<td>R4000-R7999</td>
<td>4993</td>
<td>109846</td>
<td>21636</td>
<td>416</td>
</tr>
<tr>
<td>R8000-R15999</td>
<td>10244</td>
<td>225368</td>
<td>44391</td>
<td>854</td>
</tr>
<tr>
<td>R16000-R32999</td>
<td>20509</td>
<td>451198</td>
<td>8872</td>
<td>1709</td>
</tr>
<tr>
<td>R33000+</td>
<td>15800</td>
<td>1133600</td>
<td>23167</td>
<td>292</td>
</tr>
</tbody>
</table>

The impact of the imputations

Given that there are so many missing values, one might assume that it would be beneficial to impute. Nevertheless the procedure actually adopted in this data set is unlikely to be useful. It is inconceivable that the distribution of daily, weekly, monthly or annual data will be the same within each bracket. Indeed the empirical (non-imputed) figures shown in Table 3 show significant differences. For instance the mean income among daily paid workers that fell into the lowest bracket was R30 and the median only R20. Converted to monthly figures (by the factor of 22) this would give R660 or R440 as the appropriate imputed value, rather than the R1364 actually assigned in the data set. While the daily and weekly values are over-imputed, the annual data is likely to be underimputed. Individuals with annual income in the top bracket (R33 000+) had a mean monthly income of R6231, instead of the R4292 assigned them by the imputation algorithm. The net effect of these different biases for the different categories is difficult to predict. Redoing the imputation but using the category means separately for each reporting period leads to a slight increase in mean monthly “net income” from R1844 to R1529.

Table 3 Distribution of reported income within the bracket, by reporting period

<table>
<thead>
<tr>
<th>Income category</th>
<th>Daily</th>
<th>Weekly</th>
<th>Monthly</th>
<th>Yearly</th>
</tr>
</thead>
<tbody>
<tr>
<td>R Mean</td>
<td>30.2</td>
<td>70.9</td>
<td>64.55</td>
<td>7 Mean</td>
</tr>
<tr>
<td>Median</td>
<td>20</td>
<td>73.62</td>
<td>70</td>
<td>Median</td>
</tr>
<tr>
<td>Imputed</td>
<td>8.371</td>
<td>0.242</td>
<td>0.379</td>
<td>Imputed</td>
</tr>
<tr>
<td>a</td>
<td>582</td>
<td>1.617</td>
<td>203</td>
<td>a</td>
</tr>
<tr>
<td>Median</td>
<td>124.9</td>
<td>149</td>
<td>150</td>
<td>8 Mean</td>
</tr>
<tr>
<td>Median</td>
<td>120.5</td>
<td>350</td>
<td>152</td>
<td>4 Mean</td>
</tr>
<tr>
<td>Imputed</td>
<td>8.5</td>
<td>0.5</td>
<td>0.369</td>
<td>Imputed</td>
</tr>
<tr>
<td>a</td>
<td>10.5</td>
<td>1.357</td>
<td>829</td>
<td>a</td>
</tr>
<tr>
<td>Mean</td>
<td>278.4</td>
<td>297.7</td>
<td>292.1</td>
<td>9 Mean</td>
</tr>
<tr>
<td>Median</td>
<td>200</td>
<td>296.1</td>
<td>336</td>
<td>10 Mean</td>
</tr>
<tr>
<td>Imputed</td>
<td>0.571</td>
<td>0.529</td>
<td>0.312</td>
<td>10 Imputed</td>
</tr>
<tr>
<td>a</td>
<td>7</td>
<td>1.966</td>
<td>2.985</td>
<td>a</td>
</tr>
<tr>
<td>Mean</td>
<td>500</td>
<td>653.6</td>
<td>748.5</td>
<td>10 Mean</td>
</tr>
<tr>
<td>Median</td>
<td>500</td>
<td>600</td>
<td>780</td>
<td>20 Mean</td>
</tr>
<tr>
<td>Imputed</td>
<td>0</td>
<td>0.595</td>
<td>0.555</td>
<td>20 Imputed</td>
</tr>
<tr>
<td>a</td>
<td>7</td>
<td>1.289</td>
<td>1.734</td>
<td>a</td>
</tr>
<tr>
<td>Mean</td>
<td>1400</td>
<td>1368</td>
<td>1441</td>
<td>20 Mean</td>
</tr>
<tr>
<td>Median</td>
<td>1400</td>
<td>1389</td>
<td>1428</td>
<td>20 Median</td>
</tr>
<tr>
<td>Imputed</td>
<td>0.8</td>
<td>0.6</td>
<td>0.654</td>
<td>20 Imputed</td>
</tr>
<tr>
<td>a</td>
<td>5</td>
<td>7</td>
<td>0.453</td>
<td>a</td>
</tr>
</tbody>
</table>

Notes:
1. The statistics are weighted using the Statistics South Africa released person weights.
2. The income categories are as per reporting period, e.g. category 2 is R1 to R99 either per day, per week, per month or per year.
3. The statistics are calculated over individuals who were not located at one of the spikes, i.e. pre-imputation.
4. The statistics provide the mean and median of the point data, proportion of data set located at spikes and sample size of the cell as a whole (i.e. imputed and non-imputed).

An increase of 3% in mean income (among those reporting) does not seem an awful lot, although the discrepancy is sufficiently large that naive hypothesis tests (not controlling for clustering) would not reject the idea that the two means are equal. Table 4 shows that there is a noticeable impact of the imputation process on average incomes among people who report daily incomes and those reporting annual income. Indeed the difference is statistically significant in the latter case.

<table>
<thead>
<tr>
<th>Imputed</th>
<th>100</th>
<th>610</th>
<th>990</th>
<th>1389</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.857</td>
<td>0.75</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Imputed</th>
<th>100</th>
<th>610</th>
<th>990</th>
<th>1389</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.857</td>
<td>0.75</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This evidence suggests fairly strongly that individuals that merely ticked a category without providing a Rand income were placed at these values within the brackets and this amount was then converted to a monthly figure, using the appropriate conversion rate. That begs, of course, the question why these particular values should have been used for the imputation in the first place.

That question turns out not to have any easy answers. It seems natural to suppose that these numbers would be based on the means or medians of those individuals actually supplying point estimates within each category. Nevertheless trying all sorts of permutations of these (even geometric means) we could not reproduce these numbers.

The impact of the imputations

Since the spikes are well defined it is straightforward to remove them. This may remove also the odd individual who coincidentally gave precisely that figure, but this error is likely to be negligible. Taking out the spikes has a dramatic effect. Table 2 shows fully 54% of all the observations at one or other of the spikes! The same table also shows that there is a systematic negative relationship between reporting actual information and income received. Note that category 12 is not the highest income bracket (R33 000+) had a mean monthly income of R6231, instead of the R4292 assigned them by the imputation algorithm. The net effect of these different biases for individuals with annual income in the top bracket (R33 000+) had a mean monthly income of R660 or R440 as the appropriate imputed value, rather than the R1364 actually assigned in the data set. While the daily and weekly values are over-imputed, the annual data is likely to be underimputed. Individuals with annual income in the top bracket (R33 000+) had a mean monthly income of R6231, instead of the R4292 assigned them by the imputation algorithm. The net effect of these different biases for the different categories is difficult to predict. Redoing the imputation but using the category means separately for each reporting period leads to a slight increase in mean monthly “net income” from R1844 to R1529.
Imputations in the “gross income from own activity” variable

The general points from the discussion on the “net income of employee” variable carries forward to the “gross income from own activity” variable. There are still spikes which seem to be due to imputations from the bracket data, and there are still conversions from daily, weekly and annual data that need to be taken into account. The situation is, however, somewhat more messy as Figure 6 indicates. Here we have again converted the “calculated” values back to the original ones, using the same conversion rates as for the “net income of employee” variable. These conversion rates seem to be the right ones, because the spikes again line up precisely.

Figure 6 In the “gross income from own activity” variable there seem to be two spikes per bracket

In this case, however, there seem to be two imputation spikes per bracket. For instance in the lowest category there are spikes at R62 as well as at R71. The first of these is identical to the spike in the “net income of employee” variable while the latter seems specific to this variable. In the next bracket there are spikes at R130 and R143. The former is new whereas the latter is identical to a spike we have seen before.

In Table 5 we give a listing of the “spikes” we observe in the data set. As before all the strange spikes can be obtained by simple transformations from the monthly spikes. Not all these potential spikes are actually observed in the data. For the higher income categories (above R1000 per reporting period) there is generally only one spike per interval.

Table 5 Income spikes in the "gross income from own account activities" variable

<table>
<thead>
<tr>
<th>Reporting period bracket:</th>
<th>Daily</th>
<th>Weekly</th>
<th>Monthly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>common</td>
<td>new</td>
<td>common</td>
</tr>
<tr>
<td>R1-R99</td>
<td>1364</td>
<td>562</td>
<td>269</td>
</tr>
<tr>
<td>R100-R199</td>
<td>3146</td>
<td>2860</td>
<td>820</td>
</tr>
<tr>
<td>R200-R499</td>
<td>8736</td>
<td>7504</td>
<td>1335</td>
</tr>
<tr>
<td>R500-R999</td>
<td>15972</td>
<td>4838</td>
<td>3146</td>
</tr>
<tr>
<td>R1000-R1999</td>
<td>30492</td>
<td>29150</td>
<td>6006</td>
</tr>
<tr>
<td>R2000-R3999</td>
<td>58938</td>
<td>57332</td>
<td>11609</td>
</tr>
<tr>
<td>R4000-R7999</td>
<td>110846</td>
<td>112266</td>
<td>21636</td>
</tr>
<tr>
<td>R8000-R15999</td>
<td>255368</td>
<td>230186</td>
<td>44391</td>
</tr>
<tr>
<td>R16000-R31999</td>
<td>451198</td>
<td>489742</td>
<td>88872</td>
</tr>
<tr>
<td>R32000-R63999</td>
<td>113300</td>
<td>999526</td>
<td>223167</td>
</tr>
<tr>
<td>R64000-127999</td>
<td>300556</td>
<td>35641</td>
<td>1289167</td>
</tr>
<tr>
<td>R128k+</td>
<td>95450000</td>
<td>1289167</td>
<td></td>
</tr>
</tbody>
</table>

Notes:
1. The columns labelled “common” list the spikes that are common with the “net income of employee” spikes. The columns labelled “new” indicate spikes that seem specific to the variable.
2. Cells with dark shading indicate that this value is not observed in the data at all.
3. Cells with light shading indicate that this value is observed, but the “spike” is not a local maximum, or it represents a negligible fraction of the values in that bracket.

It is difficult to make sense of this pattern. It seems clear why there would be a “new” set of imputations. We wouldn’t expect the point estimates from the “net income of employee” variable to be good approximations to the values of the typical “gross income from own account activities”. It is hard to see why there are any spikes in common with the other set of imputations, let alone how there could be two different imputations for values in the same bracket.

Although it is difficult to understand the process that generated these imputations, it is fairly easy to identify the imputations themselves. It is therefore also easy to remove the imputed values and so separate the categorical information from the proper point estimates.
Conclusion

It is evident that there are different levels of manipulation that led to the creation of the two income variables. However it is dubious to put daily and annually paid individuals at the same point within each bracket as monthly paid workers. The existence of the spikes, however, alerts us to the existence of the problem and enables us to correct appropriately.

Files to accompany this document

OHSIncome.do [1] This creates imputation flags for the two income variables plus it generates new imputations, different by reporting period.

Category: OHS 1994

- This page was last modified 09:24, 17 January 2008.
DataFirst is a research unit at the University of Cape Town engaged in promoting the long term preservation and reuse of data from African Socioeconomic surveys. This includes:

- the development and use of appropriate software for data curation to support the use of data for purposes beyond those of initial survey projects
- liaison with data producers - governments and research institutions - for the provision of data for reanalysis
  - research to improve the quality of African survey data
  - training of African data managers for better data curation on the continent
  - training of data users to advance quantitative skills in the region.

The above strategies support a well-resourced research-policy interface in South Africa, where data reuse by policy analysts in academia serves to refine inputs to government planning.