## 1. Research Data

This document describes the data management tasks associated with producing the empirical data sets underpinning the thesis “Risk and Resilience in Scottish Charities”. The PhD project, and data collection, began on 01 October 2013.

### 1.1. Transforming administrative data for use in social science research

The crucial task in using administrative data for academic research purposes is to transform the data so it is amenable to statistical analysis. The data should be processed in such a way that the final dataset looks (structurally) and behaves (statistically) like social survey data. Common steps in processing administrative data include: data cleaning; coding of variables; handling of missing cases and values; matching or linking with other datasets; and creating derived variables. The salient difference relates to the nature of the data collection stage. Survey data collection is informed and driven by statistical and research concerns. Administrative data, as previously discussed, is collected for other purposes. This places greater emphasis on consulting the metadata and guidance held by the administrative data collector in order for the dataset to be transformed to meet the needs of the researcher.

*Data provenance*

There are four main sources of charity administrative data utilised in this research:

* Scottish Charity Register – is the public register, created and maintained by OSCR, of all organisations registered as Scottish charities. This includes charities registered by OSCR and those registered under different regimes such as UK Inland Revenue.
* Annual Returns – is a mandatory reporting requirement for all active Scottish charities (and some that are inactive but have charitable assets that require monitoring to ensure they are still used for charitable purposes). The annual return collects financial information about a charity and must be submitted, along with a set of accounts, within nine months of the end of an organisation’s accounting year. Charities with annual gross income of less than £25,000 complete only the annual return, while charities with annual gross income of £25,000 or more also submit a supplementary monitoring form that captures more detailed financial information.
* Investigations data – these are records captured by the Monitoring & Compliance team at OSCR of complaints and subsequent regulatory action against charities; Chapter Four deals primarily with these data.
* Financial Exceptions data – these are records that capture instances of financial vulnerabilities and improprieties in charities’ annual returns; Chapter Six deals primarily with these data.

Each of these datasets was acquired from OSCR at different times over the course of the research. Table 3.1 below provides further information on each of these sources of raw data (i.e. before any data processing has occurred), in particular their format, size and coverage (cases, variables and time periods). Thanks to OSCR’s use of a consistent unique identifier – Scottish Charity Number – each of the datasets have the potential to be linked to each other. The first two data resources act as the base datasets for the analyses, with the others used as a source of further information. The data are generally of high quality and often the most reliable and valid measures of interesting constructs that we have; the fields contained in these datasets have remained consistent in definition and recording over the time period.

**Table 3.1.** Charity administrative data properties

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Properties | Scottish Charity Register | Annual Returns | Investigations | Financial Exceptions |
| Observations | 44,888 | 165,310 | 2,666 | 31,182 |
|  |  |  |  |  |
| Charities | 44,888 | 27,371 | 1,802 | 7,164 |
|  |  |  |  |  |
| Variables | 45 | 95 | 15 | 11 |
|  |  |  |  |  |
| Size | 204 MB | 785 MB | 1.2MB | 5 MB |
|  |  |  |  |  |
| File type | .csv | .csv | .csv | .csv |
|  |  |  |  |  |
| Time period | 2006-2014 | 2003-2014 | 2006-2014 | 2006-2015 |
|  |  |  |  |  |
| Format | Cross section | Panel | Panel | Panel |
|  |  |  |  |  |
| Data linkage potential | Yes | Yes | Yes | Yes |
|  |  |  |  |  |
| Administrative use | Public register of all charities that are or have been active in Scotland | Monitor use of charitable resources | Investigate instances of charity misconduct and take necessary regulatory action | Monitor use of charitable resources and identify financial vulnerabilities and improprieties |
|  |  |  |  |  |
| Research use | All analyses | All analyses | Chapter Four  Chapter Seven | Chapter Six |

*Data cleaning*

The raw data described above needed considerable processing in order to be suitable for research purposes. The data management process entails the following sequential steps:

1. Converting raw data to Stata .dta files;
2. Identify and remove invalid observations;
3. Identify and remove invalid or unnecessary variables;
4. Identify and remove problematic duplicate observations (i.e. two or more observations with the same values for all variables or values for one variable that should differ across observations);
5. Identify and correct invalid values for variables (e.g. incorrect charity numbers, negative annual gross income);
6. Converting text variables to numeric and recoding variables (e.g. collapsing categories of a variable);
7. Creating derived variables;
8. Data linkage.

Certain data management tasks – such as the inclusion of observations in the statistical models, derived variables, data linkage and invalid values for independent and dependent variables – are best understood in the context of the topic under investigation and thus are described in the relevant analytical chapters. For now, we focus on the data management steps necessary to transform the administrative data into statistically usable datasets, regardless of the research topic. Table 3.2 below summarises the application of the data management process for each of our datasets.

**Table 3.2.** Data management process: charity administrative data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Step | Scottish Charity Register | Annual Returns | Investigations | Financial Exceptions |
| File conversion | .dta | .dta | .dta | .dta |
|  |  |  |  |  |
| Invalid observations | 2 dummy records | n/a | n/a | n/a |
|  |  |  |  |  |
| Invalid or unnecessary variables | Yes | Yes | Yes | Yes |
|  |  |  |  |  |
| Remove duplicate observations | 1,777 observations with invalid duplicate legal name | 9,549 invalid duplicate observations | 8 invalid duplicate observations | n/a |
|  |  |  |  |  |
| Check missing values | 2 variables with 100 percent missing values | 2 variables with 100 percent missing values | No variables with 100 percent missing values | No variables with 100 percent missing values |
|  |  |  |  |  |
| Invalid values | Discussed in Chapter Four | 345 observations with invalid combinations of charity number and annual return year | Discussed in Chapter Four | 740 observations |
|  |  |  |  |  |
| Recode variables | Yes | Yes | Yes | Yes |
|  |  |  |  |  |
| Create derived variables | Yes | Yes | Yes | Yes |

The raw data contained in the Scottish Charity Register lists records for 44,888 organisations. For the purposes of analysis however, many of these charities must be removed. First, two Scottish Charity Numbers (SCN) correspond to dummy charities created by OSCR for the purposes of testing their administrative systems (SC000036 and SC000107) and were dropped from the dataset. Second, a number of records had duplicate values for the legal name of the organisation (2,457); while many of these are valid duplicates (e.g. individual Church of Scotland charities), a large number of them had the value ‘Deleted’ (1,777). Charities with this legal name were dropped from the dataset as they correspond to organisations that were never recognised by UK Inland Revenue when they created the Charity Index – a precursor to the Scottish Charity Register – in 1992. Third, there were a large number of organisations that OSCR removed from the Index when they took over as the Registrar in April 2006. These charities never completed an annual return in the early years of OSCR’s regime and thus were removed from the Register. These organisations were easily identifiable as they had missing values for key administrative fields (e.g. constitutional form); this resulted in a further 12,371 charities being removed from the dataset. The result of this data cleaning is a dataset containing 30,738 valid Scottish charities. In the Annual Returns data 9,549 observations had duplicate values for every variable and were dropped from the dataset. As the data are in panel format (long), each observation should be uniquely identified by a combination of charity number and annual return year; however, 690 observations contained duplicate values for these key variables and 345 records were dropped.[[1]](#endnote-1) In the Investigations data 8 observations had duplicate values for every variable and were dropped from the dataset. The Financial Exceptions data contained issues with the presence of invalid duplicates. 740 records relating to charities with more than the valid maximum number of observations per year (32) were dropped from the dataset. A further 1,504 observations were dropped as they referred to obsolete financial exception codes (i.e. not recorded from 2012 and only applied to a small subset of the sector).

### 1.2. Survey and case studies

*Data provenance*

The type of survey instrument used in this study is the self-completion questionnaire. As it was not possible to send the survey to the population of charities active over the study period, the sample was restricted to those individuals that receive OSCR’s monthly newsletter (OSCR Reporter).[[2]](#endnote-2) At the time the survey was first sent by email by OSCR’s Communications team (09/06/2015), this included 6,355 individuals. However some of these recipients neither worked with or for Scottish charities (e.g. accountants and media workers), therefore the sample was further restricted to individuals listed as charity trustees (2,414), paid charity workers (1,074) and volunteers (612). This resulted in a final sample of 4,100 individual subscribers to the newsletter. The final number of responses stood at 420, a response rate of 10.2 percent of the newsletter subscribers; this accounts for roughly 1.8 percent of the population of Scottish charities at the time. As we lack data on their characteristics, we do not know if our sample is representative of the charities that subscribe to the newsletter.

The representativeness of our sample to the population of charities with regards to organisation size is described in Table 3.3 below.[[3]](#endnote-3) The survey sample is overrepresented with respect to larger charities and vice versa, and thus nonresponse bias is highly likely to be present in our sample. It is plausible that smaller organisations are less likely to subscribe to the newsletter or engage with OSCR in general except for complying with mandatory reporting requirements.

**Table 3.3.** Representativeness of survey sample to population of Scottish charities

|  |  |  |  |
| --- | --- | --- | --- |
|  | % | |  |
| Annual gross income | Survey | Scottish Charity Register | Representativeness |
| £0 | 0 | 3 | Slightly under |
| £1 - £24,999 | 31 | 54 | Considerably under |
| £25,000 - £99,999 | 20 | 20 | Representative |
| £100,000 - £499,999 | 25 | 13 | Over |
| £500,000 - £999,999 | 8 | 3 | Slightly over |
| £1,000,000 - £9,999,999 | 13 | 5 | Over |
| £10,000,000 + | 3 | 2 | Slightly over |
| Total | 100 (N=420) | 100 (N=22,313) |  |

Note: Percentages rounded to nearest whole number and thus columns may not sum to 100. Scottish Charity Register excludes charities for which there are no financial data (n=1,650).

The survey was divided into five main sections and a concluding page. Section 1 contained the introductory questions which captured basic demographic data about the individuals and the charity they represent (i.e. the role of the individual in the charity; the size, sector and legal form of the organisation). Section 2 requested data about a charity’s understanding of risk, specifically what terms it associates with this concept. Section 3 focused on the value a charity ascribes to risk and in particular risk management, including whether it supports operations, strategic planning or demonstrating accountability. Section 4 explored the behaviour of a charity in response to risk, with a particular interest in risk management practices. Section 5 examined the types of risks faced by charities. The final section captured the organisation’s Scottish charity number and any other thoughts the individual had with regards to the survey (see the appendices for this chapter to view the survey questions). The sections and content of the questionnaire map to the dimensions contained in the organisation level of the contextual framework outlined in Chapter Two of the thesis (Figure 2.3). Where possible, questions were derived (and slightly amended in some cases) from existing surveys of risk in the charity sector (Karlsson, 2012; Zurich, 2015). Therefore, the degree of reliability and validity of the questions is dependent on the source material, as well as the predictive or explanatory power of our measures of the different dimensions of risk in future studies.

An issue particular to our survey is the potential inability of the respondent to answer accurately about the organisation they are involved with – is it correct to say that a volunteer understands their charity’s formal and informal risk management procedures and tools? Therefore, it should be borne in mind that this study assumes that the answers the respondents give are congruent with those of the organisation as a whole.

*Data cleaning*

The data management process outlined in the previous section was also implemented for the survey data. There was very little data cleaning in comparison to the administrative data. No cases were dropped from the sample, and there were no duplicates and or concerns regarding excessive missing data. The one variable that did require substantial, largely manual cleaning was the one that captured a respondent’s Scottish Charity Number. As this was a free-text response, the following issues became apparent: respondent’s provided invalid values (e.g. SC071524, SC0123456 are invalid as Scottish Charity Number is sequential and there are only some 45,000 charities on the Register at time of writing); did not provide any value; or simply made a statement about their unwillingness to share this information. Out of 420 responses, 407 provided what appeared to be valid Scottish Charity Numbers. In order to examine our respondent’s investigation and regulatory intervention history, the survey responses were linked to Chapter Four’s data using the Scottish Charity Number: 339 of the 407 responses with valid data for this variable were matched in this way.[[4]](#endnote-4) Responses to the free-text question in the survey – What are the three most significant risks facing your charity? – were coded using the Charity Commission’s risk categories (see Table 2.1 in Chapter Two). Where there was a case for a response to be categorised as more than one type of risk, a judgement was made and applied consistently to all other scenarios where this occurred. For example, one respondent listed the following as a significant risk:

Coming to the end of government strategy focussing on our area of work, meaning future funding is likely to be reduced.

This could be categorised as either an external risk (as there is a change in government policy) or a financial risk (as the charity is likely to suffer a reduction in funding from this stream of work). The decision was made to categorise as external as this represents the locus of the risk.

The case studies were selected from the group of matched charities and efforts were made to include organisations of different sizes and regulatory histories (i.e. whether concerns about a charity’s conduct has been reported to OSCR in the past). 25 charities were contacted between 16 August 2016 and 10 October 2016 to elicit their participation in a face-to-face semi-structured interview and three agreed to the request: a third sector interface (TSI), cancer support charity and housing support organisation. The purpose of this data collection phase was to delve deeper into the concepts analysed in the survey and to explore whether charities felt burdened or encumbered by OSCR’s regulatory approach – see the appendices for this chapter for the topic guide used to structure the interviews. The interviews were conducted with senior members of the organisation – two managing directors, a treasurer and long-standing volunteer – at the charities’ premises and lasted between 45 and 63 minutes. Each interview transcript was listened to three times prior to uploading to NVivo for coding. The data were analysed using the framework approach advocated by Ritchie, Spencer and O’Connor (2003) and involved the following steps: drafting summary notes for each interview; developing a concept-driven, broad coding scheme derived from the topic guide; coding each transcript according to the scheme; collating codes under key themes; and interpretation and writing up of findings. Interview data were augmented with analysis of the organisations’ Trustee Annual Reports (TAR), a compulsory document submitted alongside the annual return and accounts that contains narrative information on governance, impact and provision of public benefit by the charity. Specifically, the reports were examined to see if they contained information on the risk management practices and policies of the organisations in question. Finally, all three sources of data – interviews, TARs and survey responses – were compared and contrasted to produce rounded, comprehensive accounts of the nature, understanding and impact of risk in these charities.

1. It is not known whether duplicates are valid (i.e. as the result of a change in a charity’s accounting year end date) or data entry errors; the duplicates were dropped using the *duplicates drop* command in Stata. [↑](#endnote-ref-1)
2. OSCR does possess email addresses for the principal contact of a charity but was unable to share due to data confidentiality issues. It was decided that the manual collection of email addresses for each charity was too onerous a task and thus an alternative sample-construction strategy was sought. [↑](#endnote-ref-2)
3. A copy of the Scottish Charity Register was downloaded from OSCR’s website on 17 September 2015 and used to compare to our sample. Though not perfectly overlapping, it is highly unlikely that the charities in our sample are not present on the Register a maximum of three months after completing the survey. [↑](#endnote-ref-3)
4. It is unknown why 68 of these responses did not match but it is probable that the data management process applied to Chapter Four’s data was a significant factor. [↑](#endnote-ref-4)