Editing and Imputing Administrative Tax Return Data

Charlotte Gaughan
Office for National Statistics UK

Overview

- Introduction
- Limitations
- Data Linking
- Data Cleaning
- Imputation Methods
- Conclusion and Future Work

Introduction

- An assessment of the feasibility of editing and imputing administrative tax return data to provide a substitute for survey data
- Current duplication of collection of key variables by The Office for National Statistics (ONS) and the National Tax Office (HMRC)
- The initial study focuses on turnover for the calendar year 2012. Turnover was chosen because of its importance to ONS, and 2012 for quality reasons

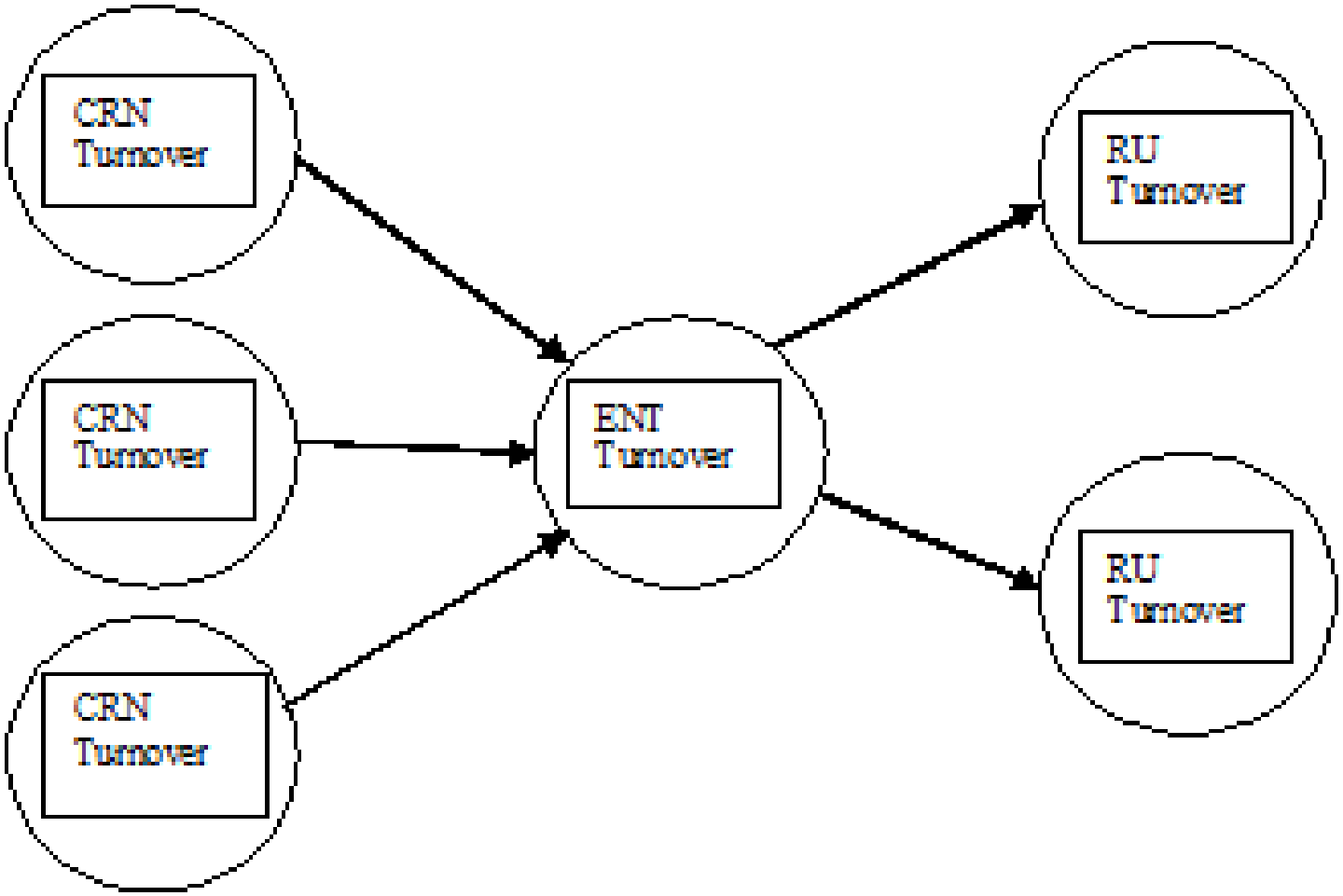
Limitations

- Tagging system used by companies to submit data. Only data which are tagged are available to use.
- Data are neither consistently nor reliably tagged
- Variables poorly populated e.g. ‘number of employees’ has a tag rate of around 1%.
- Legal issues surrounding the tax data limits our access to the data stored in London.
- Unable to query suspicious data with businesses
- The nature of the data i.e. data submitted for tax purposes may impact on the accuracy of the data compared to survey data.

Data Linking

- Companies submit tax data according to their own accounting period, the company returns were assigned to a calendar year based on the mid-point of their period of accounts.
- Data were then summed by company reference number (CRN) and by calendar year
- Data are submitted using a company reference number - in order to compare the data to ONS survey data, the HMRC data needed to be translated to RU level

Data Linking



Data Cleaning

- For the year 2012 around 0.66% of returns had all variables missing.
- Turnover had 8 % of values missing
- Estimated around 350,000 scaling errors
- 99.74% of the total turnover in the HMRC data attributed to the company with the largest turnover.

Data Cleaning

- Previous period data unedited, thus no previous period validation performed
- ONS data could be utilised at RU level which enables scaling error checks to be conducted using following formula:

650<(TaxTurnover /ONSTurnover)<1350

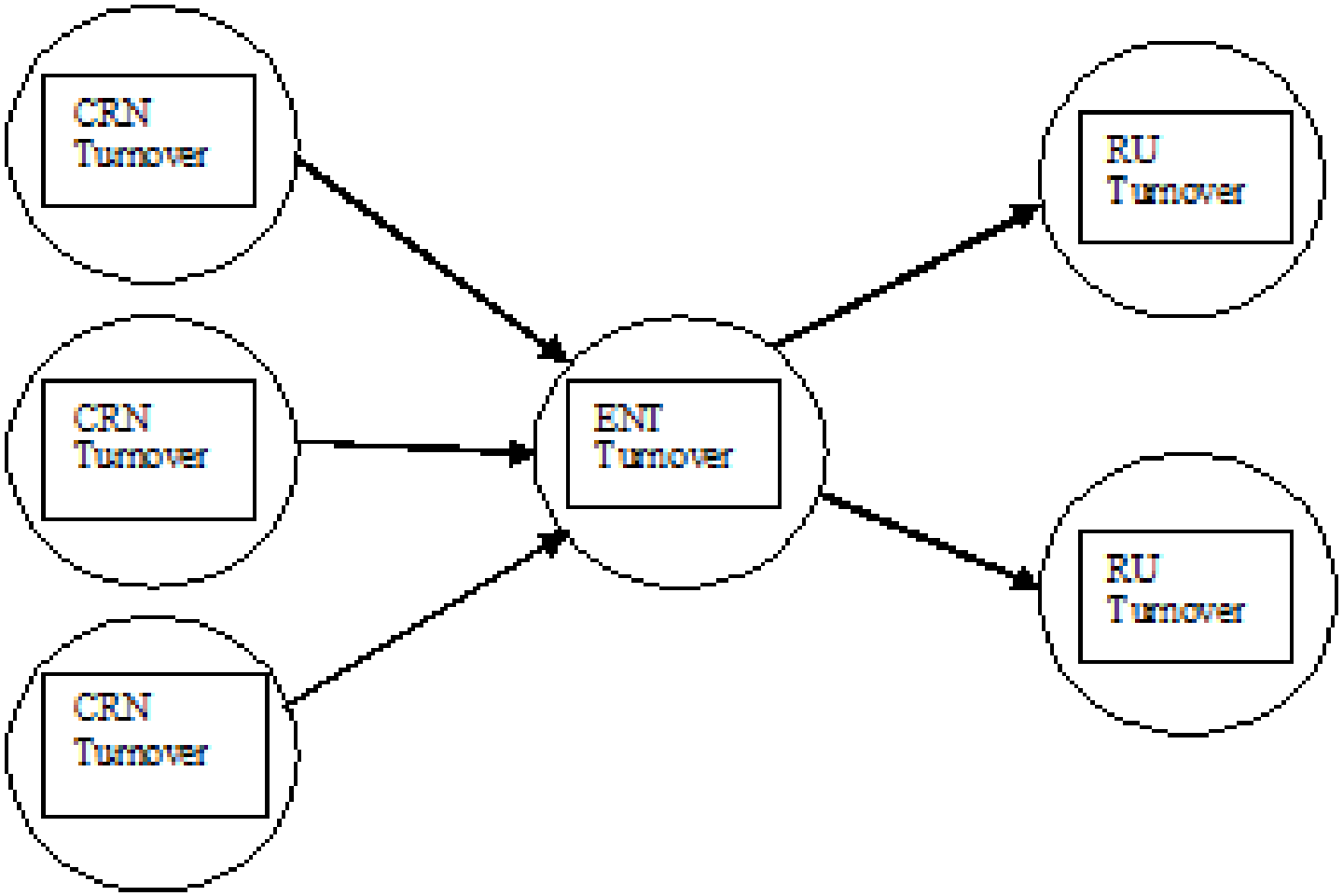
Data Cleaning

- The tagging system utilised meant that some data are submitted by companies that is not tagged – these data are not accessible to us
- Companies returned figures in currencies other than GBP, in 2012 0.4% of the companies did so - removed from the dataset
- Negative values were also removed from the dataset
- Foreign currencies and negative values were imputed

Imputation

- Imputation study required i) to provide a complete dataset to allow comparisons to ONS survey data ii) to ascertain whether the missing data in the HMRC dataset could be accurately imputed
- The dataset needed to be complete at the RU level; missingness at the CRN level did not in itself present a problem.

Data Linking



Imputation

- **All** CRN Turnovers are present at the ENT level; the CRN Turnover values are summed to create ENT turnover. The ENT value is then apportioned by RU employment to create RU Turnover.
- **Some** CRN Turnovers are present at the ENT level; the missing CRN Turnover is imputed using a median imputation and then summed with the CRN Turnovers which are present to create ENT turnover. The ENT value is then apportioned by RU employment to create RU Turnover.
- **None** of the CRN Turnovers are present; Turnover is imputed directly at the RU level.

Imputation – Strata

- In scenario 2, CRN level strata were created by summing number of CRN's per ENT level for size. Combined with a 2 digit SIC at ENT level for industry.
- In scenario 3, imputation strata were created based on employment and SIC; employment level data were available via ONS data

Imputation

- Median imputation
- Trim mean imputation (top and bottom 5% and top and bottom 10%)
- Ratio of means imputation
- $y_i^* = R \times x_i$

$$R = \frac{\frac{1}{n}\sum_{class} y_i}{\frac{1}{n}\sum_{class} x_i} = \frac{\sum_{class} y_i}{\sum_{class} x_i}$$

Imputation

- A simulation study was conducted, whereby a simple random sample of 10% was taken and imputed with each of the methods above.
- Tested for Bias and Accuracy

$$\frac{\sum_{\text{class}} (y_i^* - y_i)}{\sum_{\text{class}} y_i} \qquad \frac{\sum_{\text{class}} (y_i^* - y_i)^2}{\sum_{\text{class}} y_i}$$

Conclusion and Future work

- The results are currently being finalised, and should be published by winter 2015.
- Limitations to study potentially reduce the benefit of utilising administrative data
- The study also has the scope to evaluate other variables such as sales revenue and purchases. Recommended before the dataset is used for national statistics.