SME baseline concept note

Neil Rankin

30 January 2018

Research questions/outcomes

Phase 1 of the project aims for two outcomes:

1. An estimate of the number of SMEs in South Africa, potentially by some level of stratification (which could be things size, sector and location), and some confidence interval (or level of certainty) for this estimate.
2. A process or methodology for Phase 2 of the project which is a survey of SMEs. This will require a sample frame, questionnaire and surveying strategy.

A key outcome for this phase will be a publishable (in an academic journal) report or paper which documents the methodology and results for step 1 above.

Step 1. Estimating the number of SMEs

This is challenging since no comprehensive sample frame of the population exists in South Africa.

Definition

The initial key decision to make is the definition of an SME. There are a number of dimensions to this (my comments are in *italics* but are just starting points for the discussion):

- How do we define the size limit:
  - Is it by employment, revenue, assets or a combination of all three? *I think it should be one of these, probably employment but then we need to think about non-full time employees.*
  - Does this size limit vary by sector (like the dti's definition)? *No*
  - Do we have a lower size limit? *Not sure*
- How do we define the scope?
  - Should firms be formally registered, and what does this mean (business number, VAT, paying tax)? *Not sure*
  - Are we limiting to certain sectors? *Not sure*
  - How do we handle companies set up for things like rental income, or shell companies? *Maybe the definition is that SMEs need at least one employee*

Methodology

It probably makes sense to use a number of different methodologies to arrive at an estimate.

1. Draw from existing sources

- SARS-NT
- SARS tax records
- GEMS
- QLFS/QES
- Other StatsSA datasets such as the Survey of Employers and the Self-employed.
- NIDS
- Banks

- Chambers of commerce, business groupings

2. Some sort of 'macro' calculation

- Revenue in sector x is y, we make the assumption that it follows some distribution (which we base on global work, this will be some sort of CDF - cumulative distribution function - and we will need to vary assumptions), and then calculate SME contribution.

3. Draw on the sampling methodology work about estimating difficult to observe populations.

Much of this literature is based on human populations which move. (http://journals.sagepub.com/doi/pdf/10.4256/mio.2010.0014) provides a summary of this approaches in this area. Another sampling approach is 'snowball' sampling and its variants (including respondent-driven sampling), where respondents are asked to refer others from their networks. A summary can be found here (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3250988/). A challenge with this approach is that it can be biased, for example those with more contacts are more likely to enter the sample, and it relies on individuals revealing their contacts. There are also other approaches as the ones described here (http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.2009.00750.x/full).

Main activities and time frames

There are seven (I think) main areas of work. Under each I have tried to describe the activity, estimate the time required, and the team composition. I have roughly thought about this taking 4-5 month, since there is probably an initial set of tasks which need to happen first (and will probably take about 3 months) and then others which build on this. In terms of team requirements, apart from the internal SBP team, it would be useful to have at least one 'research assistant' to help with the data processing, some of the literature etc, and we will also need help with building the database to house the data.

1. Summary of South African datasets and existing datasets.

- Tasks:
  - Need to go through existing literature and datasets in South Africa and extract estimates of firms and employment, in total and by firm size, location and sector if possible.
  - For those datasets with underlying data need to investigate whether estimates can be calculated from this underlying data.
  - Write-up these findings.
- Time requirements:
  - Length will depend on team composition (see thoughts below) and the thoroughness (number of datasets reviewed). This could easily take two months worth of person time (i.e. 40 days).
- Team composition:
  - Up to three people: one who surveys the literature, a second to derive the estimates from the data, and Neil in an oversight role. The first two roles could be collapsed into one if we could find the right person. The second role will require some ability to work with data in a language like Stata or R.

2. Review of approaches to estimating populations

- Tasks:
  - Literature search to uncover approaches and identification of key ones
  - Read and sythesise key approaches
  - Write-up potential approach
- Time requirements:
  - Probably a period of six weeks (not full-time)
  - Two weeks literature scan, two weeks reading and write-up, two weeks for comments and review.
- Team composition:
  - Probably something which could be split between Brendon and Neil

3. Gathering macro data and producing estimates from this

- Tasks:
  - Acquire data to form the basis of 'top-down' population estimates (this could be done during the process of data collection in 1)
  - Undertake these estimates
- Time requirements:
  - Can probably be bundled with 1 and be ongoing over a three month period
- Team composition:
  - Similar team to 1

4. Constructing a list which would form the basis of a sample frame

- Tasks:
  - Obtain as many lists of firm contact details as possible
  - Verify that firms still operate and meet the criteria
  - Check for duplicates
- Time requirements:
  - On going over three months
- Team composition:
  - Requires someone to do some 'drudge' work of getting lists etc (and maybe inputting them into a database, see data management description below)
  - May also be useful to have someone with some programming skills since this might require scraping and/or automation of cleaning and organising the lists

5. Methodology

- Tasks:
  - Develop the methodology for Phase 2 - the actual survey
  - This will require drawing on the previous development work as well as SBP's 'institutional knowledge
- Time requirements:
  - Probably three months. The first one or two which overlaps and runs parallel to the initial development work
- Team composition:
  - 'High-level' SBP team

6. Instrument design

- Tasks:
  - Design the survey instrument
  - Obtain feedback on it
  - Pilot with some firms
  - integrate it with the data management system
- Time requirements:
  - Similar to the methodology. Can happen at the same time
- Team composition:
  - 'High-level' SBP team

7. Data management system

- Tasks:
  - Plan a system to manage the data (probably needs the lists, details on contactability and fit, and collected data once that is done)
  - Build the database structure
  - Build an interface or a way to get data in
- Time requirements:
  - Probably the initial three-four months of the project
- Team composition:
  - 'High-level' team to think of specifications

– Someone to construct it
– potentially someone to maintain it.