A Guide to the Ghanaian Establishment Panel Study

Bruce McDougall and Andrew Kerr, DataFirst, University of Cape Town[1]

Introduction

This document is a guide to the Ghanaian Establishment Panel Study (GEPS), a panel dataset that links manufacturing firms observed in the Ghanaian IBES 2014 firm census to observations of the *same* firms 11 years prior in the NIC 2003 firm census. Both censuses were conducted by the Ghana Statistical Service (GSS). The first part of this document describes our success in matching firms, the second provides a guide to using the panel dataset, and the third details how the panel was created.

GEPS was created using a fuzzy matching algorithm that finds firms with similar characteristics such as name, area, contact person or contact number. The broadest match criteria resulted in 751 firms matched across the two censuses. The information used to match firms across censuses was collected in phase 1 of both the 2003 and 2014 firm censuses. The structure of each census was that phase 1 collected a little information on all qualifying firms and then a subsample of firms were selected for phase 2, where substantially more data were collected.

The 751 firms that were matched between 2003 and 2014 were present in phase 1 of each census - this does not guarantee that they were interviewed in phase 2, since phase 2 was not a census but a sample. Only 222 of the 751 firms have 2003 phase 2 data, only 106 have 2014 phase 2 data and only 64 firms have phase 2 data from both 2003 and 2014. This means that firm-level production regressions, for example, can only be run on 328 observations in a pooled regression, and only 64 if a panel regression is used. This number may be even lower if there is item-level missing data for some key variables.

Another potential constrain is the quality of the firm matching. Users should familiarize themselves with how matching was done to ascertain how confident they are in the matched firms. We used a fuzzy algorithm based on firm name similarity and a host of other variables. Some of the matches are extremely strong – for example, a firm had the exact same name, industry, contact person's name, district and even telephone number in 2003 vs 2014. However, other matches were less strong, and there is the possibility of false matches.

Matching Rates

Given that 22684 firms in IBES 2014 claimed to be operating ("alive") in 2003 but only 751 firms were matched by ourselves the matching rate is 3.35%. This is extremely low. But we should note that matching rate depends on which denominator one uses. The obvious one of

[1] Andrew.kerr@uct.ac.za
We thank Tony Krakah from GSS as well as Francis Teal and Justin Sandefur for helpful comments and suggestions. This panel is an output of a project funded by the Project for Enterprise Development in Low Income Countries (PEDL).

22684 -the number of firms in IBES 2014 that reported being alive in 2003 - is more than the number of firms enumerated in the 2003 NIC. The reason for this is the substantially larger number of particularly small firms enumerated in IBES 2014 relative to NIC 2003, as documented in our paper "What is a firm census? An answer from Ghana 1962-2014." [2]

A more sensible alternative denominator can be obtained from the work of Davies and Kerr (2018), who used the 2003 NIC to draw a sample of 1000 firms and tried to interview these firms in 2013. We use the firm death numbers in that paper and assume that the firms that were not found actually died, following the results of Paufhausen and McKenzie (2019), who argued, based on analysis of several firm panel surveys, that Davies and Kerr's death rates calculated assuming the firms that were not found were not deaths were a massive outlier compared to other surveys. So if we assume the same death rate as in Davies and Kerr (2018), assuming all the not founds were deaths, then using these predicted number of deaths and survivors from NIC 2003 the match rate roughly doubles to 6.2%.

But if mostly larger firms are matched it is possible that the persons engaged in the matched firms still represent a relatively large fraction of all persons engaged in firms that were expected to survive (either matched or unmatched). When using the predicted survival rates and changes in employment in the matched firms as a predictor of the changes in unmatched firms we find that the persons engaged in matched firms are 22% of predicted persons engaged in surviving firms. But this is still a low matching rate. Researchers should thus use the panel with caution, and explore how matching varied by observable characteristics.

How to Use the Panel Datafiles

The Linking File

The fuzzy matching process resulted in a list of 751 IBES firms for which a match was found in NIC, stored in our panel linking file called geps-2003-2014-v1.dta. The link between NIC and IBES is created via two ID variables: the IBES firm ID called BID and the NIC firm ID for the matched firm in NIC called refnum. These variable names were chosen to match the names given by GSS.

The file includes 28 other variables. Most important is the *rank* variable, which is a roughly ordinal measure of the quality of the match, ranging between 1 and 33 with 1 being the strongest match. Ranking the quality of the fuzzy matches is beneficial as the user can decide how good or bad the matches are that they are willing to work with. Users should be warned that not all the matches were high quality – as a rule of thumb we consider those beyond rank 25 weak. To evaluate this rank variable, see the discussion below in *How the Panel was Created* and refer to the table at the end which contains the criteria that were used to define a match at each rank.

The remaining variables in the linking file were those used to match the firms and these can also be referred to as a gauge of match quality. The *similscore* variable for example gives the similarity score between the matched firms' names in IBES vs NIC (ranging from 0.6 to 1). Any variable prefixed with *simil* is a similarity index where 1 indicates a perfect match. Actual firm names, locational data and contact information naturally could not be included in this public data file.

[2] Available online: https://www.csae.ox.ac.uk/papers/what-is-a-firm-census-in-a-developing-country-an-answer-from-ghana

Data Sources

There are four data files that can be merged into the GEPS linking file to create a panel: NIC phase 1, NIC phase 2, IBES phase 1 and IBES phase 2. It is up to the user to merge in what they want -if they prefer they can work with just phase 1 data to create a panel of 751 observations. As explained, if they merge in phase 2 for greater detail, this means a substantial decline in sample size– only 64 matched firms had phase 2 data for both time periods.

The full IBES 2014 phase 1 data can be found on the GSS website:
https://www2.statsghana.gov.gh/downloadpage.html

The 2003 NIC data (phase 1 and 2) is available on Francis Teal's website:
https://www.empiricalde.com/ghana-firm-census-1987-and-2003

IBES phase 2 is slightly more complicated. DataFirst was previously allowed to release a 40% sample and that data file is available on our open data portal. However, not all of the firms that we found good matches for in the fuzzy panel were included therein. We therefore decided to upload an additional file with phase 2 data for *all* 106 of the IBES phase 2 firms that had panel matches. This is distributed with GEPS on DataFirst's website:

https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/859

The user can use the firm identifiers to merge in data from the NIC and IBES phases 1 and 2.

How the Panel was Created

This section details how GEPS was created.

Notes on Firm Census Designs

There are several design features of the NIC and IBES that make it difficult to track the establishments that were surveyed in 2003 and survived to 2014 to be surveyed again. Firstly, the questionnaires did not contain any linking questions specifically for this purpose: firms in 2014 could not indicate if they were surveyed before, or if they had changed premises or names. A starting point is to restrict the IBES sample to firms established before 2004. For manufacturing firms, this represents 22684 of 99437 (or 24%) firms in IBES. Firms without starting year information in IBES were dropped.

The next natural step is to look for observations with the same firm name recorded NIC and IBES. Firm names can provide a major clue as to whether two observations in IBES and NIC are tracking the same firm in reality, especially if the names are very distinct.

A problem with looking solely at names is that this may contradict the way observations (called establishments) are defined in the census. Establishments in the NIC/IBES censuses are defined/differentiated by 1) their physical location and 2) the type of production they are engaged in. As a result, in both NIC and IBES, different observations can have the same name; indeed, there are many duplicated firm names across different observations. In a given year, the two observations could have the same name but different locations due to the same name arising coincidentally or possibly because these are branches of a franchise scattered throughout Ghana. Two separate observations can also exist at the exact *same* geographic location and have the same name but be recorded as two different enterprises because there are two different business activities going on. Naturally two observations with the same name might also differ in both location and activity, in which case it is highly likely that the name is the same by coincidence. As such, if we are to stay true to GSS's definition of an establishment and try to match these 1-to-1 between NIC and IBES, matching by name alone will create a high rate of false matches.

Difficulties due to Firm Names & Fuzzy Matching

Above we make the case that pairs of observations in NIC/IBES with the same firm name might not actually represent the same "establishment" as defined by GSS. Conversely, in many cases observations in NIC and IBES with slightly different names actually do represent the same establishment. This occurs if there is a slight misspelling, a respelling, or a reordering of the words/letters/initials that make up the firm name. There are surprisingly many variations that appear even for simple names. Many establishments are identical in every regard, except for a minor naming error.

In these cases, it is necessary to provide a 'fuzzy' match between the name variables that allows slight variation in firm names between NIC and IBES. We allow this fuzziness using the *matchit* command in Stata. Matchit does a pairwise comparison of every name in each dataset, ranks the similarity between them, and outputs the results. An appeal of Matchit is that it allows the user some control over the algorithm used to match names which can be tailored depending on the types of data being input. It also returns the similarity ranking score which is very useful for limiting the results to close matches. We used the ngram method to parse strings, with n=3.

Matching Procedure: Definition of a Match

Because establishments are defined by location and activity, matching should be done with some control for location and production type. Otherwise, we might erroneously link different branches of a franchise, or even completely distinct operations that happen to have the same or similar names. A downside of this approach is that we cannot allow an establishment to move location or change activity. However, this is a direct result of census definitions and cannot be undone. Trying to link all branches of a franchise (in different locations and even activities) will possibly give more matches but this is a different question to the panel we aim to create, which means linking analytically equivalent observations at a 1 to 1 level.

Matching Procedure: Ranking Matches

Matchit provides us a list of all the possible name matches for each unique firm in each dataset, which creates many duplicates of each original firm. We then merged in all the phase 1 information for these duplicated firms from NIC/IBES and looked for good matches in terms of names and the other variables between both NIC and IBES. We narrowed the matching process down by only considering those possible matches that had a name similarity score of at least 0.6.

We then use 33 sequential sets of criteria which define a "good" match. There are ranked by how good the match is: We start with the most specific (stringent) conditions and then loosen the criteria. The ranking variable is called rank and ranges from 1 to 33. The different sets of conditions used are provided in the table 1 at the end. This includes similarity (fuzzy) matching not only between firm names (a default condition), but also between suburbs, towns, and something we call "contact person", allowing extra room for misspellings in the data. Throughout the process we doubled checked the results to see that the matches seemed reasonable, adjusting the thresholds accordingly to avoid picking up the firms that seemed too dissimilar.

Final Processing and Results

The result of the above is a set of potential matches in NIC for each IBES firmid. We then keep the highest-ranking match for IBES firmid and drop the rest. The highest-ranking match is that with the best rank, and the best similarity across other variables in cases where two matches (or more) are tied in terms of rank. This process results in a unique best match in NIC for each IBES firmid in the dataset. We also decided to drop matches where two IBES firms were matching to the same NIC firm (66 cases). Many of these 66 did not seem like good matches in any case. The result is a 1 to 1 link between IBES and NIC. This unfortunately rules out merges or splits of a firm(/s) but has the benefit of providing an intuitively simple relationship.

There are 22684 IBES firms that were born before 2003. Of these, around 8000-10000 have firm names that are similar (>0.6 matchit similarity score) to names of firms in the NIC, depending on the matchit algorithm used. However, when one incorporates even quite a limited set of other variables that must match, the overall rate drops rapidly. The final numbers of IBES firms with any of the 33 match types was **751** firms.

Table of Matching Criteria

The matching table on the (landscape) page below details all the matches we allowed. Some examples illustrate the logic of table 1. The first and highest ranked match comes from matching telephone, cellphone, or fax numbers, as well as district, isic4 and start years (all being exactly equal). Note that because Ghana changed the directory listing between surveys we ignore the regional extensions (the beginning of the number) and only take the last 5 digits. In the data this provides only 27 matches. This is a very strong match as telephone numbers are designed to be unique and non-changing or repeating after the regional codes.

Rank 2 firms are the same but allowing for only a 2-digit ISIC match and for missing start year data. When we flag firms as rank 2, we skip the firms already flagged as 1, in order to preserve the ranking. This continues throughout the process. For both 1 and 2 I use the default similarity score minimum between firm names of 0.6. Rank 3 uses the same conditions as 2, but replaces an exact match between districts with a fuzzy match between suburbs.

Ranks 1-7 are based on varying conditions that include a telephone match. Ranks 8-16 are based on various conditions, including a postal box similarity match. Ranks 17-21 instead use a match between the contact person in NIC and the firm name in IBES. This is useful as many entrepreneurs named their firms after themselves in NIC. In 22-25 we use the same start year, ISIC and geographic information, but do not combine this with another "special" condition such as telephone, PO box or contact person. In 26-33 we sweep back through the special variables, allowing far more lenient conditions for the other variables. For example, several have no start year match condition, and some even have no geographic conditions. We consider these matches beyond 26 less reliable than those of a better rank.

In each of the 33 cases we visually inspected the data and set fuzzy thresholds at a point just above where clearly false firms began to be matched. For example, in rank 12 the minimum name similarity rises from 0.6 to 0.65.

Table 1: Minimum conditions that define a match

rank	start year match	isic code match	geographic match	name similarity min	fuzzy geographic match	other	Number of cases
1	Exact	4 digit	District	0.6	None	Telephone	25
2	Exact or missing	2 digit	District	0.6	None	Telephone	8
3	Exact or missing	2 digit	Suburb	0.6	Simil_subs > 0.2	Telephone	0
4	Within 1 year	2 digit	Town	0.6	Simil_town > 0.3	Telephone	4
5	Within 1 year	None	town or suburb	0.6	None (exact)	Telephone	10
6	None	2 digit	District and town or suburb	0.6	None (exact)	Telephone	28
7	None	2 digit	Town or suburb	0.6	Simil_subs > 0.5 or simil_towns > 0.3	Telephone	14
8	Exact	4 digit	District , suburb	0.6	Exact	Simil_PoBox > 0.5	11
9	exact or missing	2 digit	Town	0.6	Exact	Simil_PoBox >0.5	16
10	Within 1 year	2 digit	Town and Suburb	0.6	Simil_subs > 0.5 and simil_towns > 0.3	Simil_PoBox >0.5	13
11	Within 1 year	None	Town and suburb	0.6	Simil_subs > 0.2 and simil_towns > 0.6	Simil_PoBox > 0.5	5
12	None	2 digit	District, town and suburb	**0.65**	Exact	Simil_PoBox > 0.5	6
13	None	2 digit	District, town and suburb	**0.65**	Simil_subs > 0.2 and simil_towns > 0.6	Simil_PoBox > 0.5	21
14	Exact	4 digit	District	**0.65**	none	Simil_PoBox > 0.5	22
15	Within 1 year	2 digit	District, towns	0.6	Simil_towns>0.4	Simil_PoBox > 0.5	11
16	Exact or missing	2 digit	District	**0.65**	None	Simil_PoBox > 0.5	11
17	Exact or missing	4 digit	District, town and suburb	0.6	Simil_subs > 0.5 and simil_towns > 0.3	Simil_nameperson>0.2	8
18	Exact or missing	2 digit	District, town	0.6	Simil_town>0.3	Simil_nameperson>0.2	25
19	Exact or missing	4 digit	District	0.6	None	Legal form==1, Simil_nameperson>0.2	96
20	Within 1 year	4 digit	District, towns, suburb	0.6	Simil_subs > 0.5 and simil_towns > 0.3	Legal form==1 Simil_nameperson>0.2	11

21	none	4 digit	Districts, town, suburb	0.6	Simil_subs > 0.5 and simil_towns > 0.3	Legal form==1 Simil_nameperson>0.2	36
22	Exact or missing	4 digit	District, town, suburb	0.6	Exact	None	7
23	Exact or missing	4 digit	District, town, suburb	0.6	simil_subs > 0.3 simil_towns> 0.2	Legal form	25
24	Exact or missing	4 digit	District, town	0.68	simil_towns> 0.4	Legal form	51
25	Within 1 year	4 digit	District, town, suburb	0.6	Exact	none	11
26	Exact or missing	2 digit	None	0.7	None	Simil_nameperson>0.7* legal form	36
27	None	2 digit	None	0.6	None	Telephone, Legal_form,	44
28	Exact or missing	2 digit	District	0.6	None	Simil_PoBox>0.5	2
29	None	2 digit	Town or suburb	0.65	Simil_town > 0.5 simil_subs>0.4	Simil_PoBox>0.6 legal form	74
30	None	2 digit	District	0.65	None	Simil_PoBox>0.6 legal form	49
31	Exact	2 digit	District	0.6	None	Simil_nameperson >0.7* Legal form	1
32	None	2 digit	District	0.6	None	Simil_nameperson >0.8* Legal form	24
33	None	2 digit	None	0.6	None	Simil_nameperson >0.9*	46

Total matches all ranks: 751   Total matches ranks 1-25:47