A guide to the Ghana Statistical Service's 2014 Integrated Business Establishment Survey Phase Two Public Dataset, version 1.

Andrew Kerr and Bruce McDougall[1], DataFirst, February 2021

Introduction

This guide serves as an introduction to the Integrated Business Establishment Survey (IEBS) 2014 phase two public dataset that has been released on DataFirst. We explain how the IBES census phases one and two were undertaken by the Ghanaian Statistical Service (GSS) and our own further sampling for the public release. We explain the non-response adjusted weights we constructed for phase two and detail possible mismatching of some firms between phase one and phase two. We then introduce the final cleaned dataset and document our methods used in creating the public data from the data we obtained from GSS.

Citation of the Phase Two data and this Guide

To cite the IBES phase two public release data please use:

Ghana Statistical Service and DataFirst. Integrated Business Establishment Survey 2014, Phase II [dataset]. Version 1. Ghana: Ghana Statistical Service and DataFirst [producers], 2021. Cape Town: DataFirst [distributor],2021. doi: https://doi.org/10.25828/7er4-cz67

If you have used this guide please cite it as

Kerr, A. and McDougall, B. (2021). "A guide to the Ghana Statistical Service's 2014 Integrated Business Establishment Survey Phase Two Public Dataset, version 1.

IBES 2014 Design

The Integrated Business Establishment Survey was an establishment census conducted by the Ghana Statistical Service (GSS) in 2014. IBES 2014 phase I collected data on 638 000 establishments in Ghana across all sectors. All non-household-based establishments were included, as well as any household-based establishments with a sign indicating their presence (GSS, 2015)[2]. Basic

[1] andrew.kerr@uct.ac.za

We thank the Ghanaian Statistical Service (GSS) for making the data available and to Anthony Krakah, Isaac Dadson and Jacqueline Anum from GSS for the extremely helpful assistance in understanding the IBES data. This release of the IBES phase 2 data has been funded by the Project for Enterprise Development in Low Income Countries (PEDL).

[2] In a separate paper also funded by PEDL as part of this project (Kerr and McDougall, 2020) we examine what a "census" of firms meant in the 2014 IBES as well as the 1962, 1987 and 2003 firm censuses that have been undertaken in Ghana.

information such as the industry, number of persons engaged, ownership and registration was collected.

After Phase I, Phase II was a roughly 5% stratified sample of the phase 1 firms undertaken in 2015 (GSS, 2017). To select this sample firms were stratified by industry, region and size. There were 10 regions, 101 industries and 6 firm size categories, but not all the industry-size-region groups had samples selected from them, since some had no firms in them. Neyman optimal allocation was undertaken to determine sample sizes within each stratum (GSS, 2017). All firms with 50 or more persons engaged were sampled with certainty. A simple random sample was undertaken in each stratum.

The firms in phase 2 were enumerated in far greater detail, including questions on assets, costs and revenues. Unlike phase 1, which involved a single standard questionnaire, phase 2 was collected using 9 different questionnaires depending on the activity of the establishment.

Public Release Data Sampling for Phase Two

As discussed above, phase 1 of IBES census of 638000 firms undertaken by GSS in 2014. GSS then sampled about 5% of these firms for phase 2, around 31000 firms. The full phase one data is publicly available on the GSS website[3] and can easily be merged into the phase two data from DataFirst using the ***beid*** firm identifier variable. GSS has allowed DataFirst to release a 40% sample of the phase two sample. This means that the dataset we are releasing contains 40% of the 5% phase two sample collected by GSS. [4] To select the firms for the public release we stratified on firm size, region and 1 digit ISIC code. There were 9 833 firms with 10 or more persons engaged in the realised GSS phase two sample and we sampled these firms with certainty. We then sampled 50% of those firms in the realised sample with 5-9 persons engaged (1746 firms), 10% of those with 2-4 persons engaged (755 firms) and 10% of those with 1 person engaged (343 firms). The final sample size is thus 12660 firms.

The design weight in the public release data set is thus a function of the design weight for GSS phase two and the probability of selection into our 40% public release sample. For firms with 10+ persons engaged this means that the design weight in the public release is simply the phase two weight provided by GSS. For those with 5-9 persons engaged it is the design weight multiplied by the inverse

[3] https://www2.statsghana.gov.gh/downloadpage.html

[4] The sample size in the public release is 40% of the number of firms in the sample design, which was 31 152. The realized sample size in phase 2 was 24 360 firms due to non-response, which we discuss in more detail below.

of the probability of selection (0.5), ie twice the design weight, and similarly for the smaller firms. In the next section we discuss how we adjusted the design weights for non-response.

Non-response Adjusted Weights

The GSS phase two report (GSS, 2017) gives the non-response rate as 78%. There was no non-response adjustment to the weights used in the GSS phase 2 report.[5] We have used the phase one sampling methodology to create non-response adjusted weights for the public release. The strata used by GSS to select firms for phase 2 were size group, industry and region combinations and GSS took a simple random sample within strata. We use a simple method to create non-response adjusted weights. The non-response adjusted weight is the number of firms in the stratum enumerated in phase 1 divided by the number of responding firms in the stratum in phase 2. This method assumes that non-response is "Missing Completely at Random" within strata (Lohr, 2010)[6].

Table 1 shows total employment and total number of firms when using the phase 1 data. It also shows the same totals with the same phase 1 total persons engaged variable when estimated using the phase 2 firms only with either the weight provided in the data from GSS or the non-response adjusted weight created by ourselves. The GSS weight provided with the data substantially underestimates the total number of firms and the number of persons engaged from phase 1. Our non-response adjusted weight underestimates the total number of persons engaged by around 10% but, by design, gets the number of firms almost correct (when adding in the roughly 6000 firms in strata where no firm responded in phase 2).

A possible explanation for why the non-response adjusted weight underestimates total employment is that non-response rates within the largest size strata (50 or more persons engaged) are higher for the larger firms in this stratum. The confidence intervals are very small for both phase 2 estimates, due to the large number of strata with substantial heterogeneity in firm size across strata and because the largest firms are sampled with certainty. This indicates that the underestimate of firms and persons engaged when using the weight provided by GSS is very unlikely to be due to sampling error and is most likely due to the lack of a non-response correction.

[5] This was confirmed to us by GSS staff and can be seen in our analysis below.

[6] There are 470 strata with a total of 6180 phase 1 firms where no firms responded in phase 2. 80% of these firms were in 3 industries- public administration, secondary education and primary education. We thus decided to ignore these strata and so they are not represented in the final phase 2 dataset. There are almost no public administration "firms" enumerated in phase 2 at all. GSS staff said this was because relevant data could be collected from publicly available data sources. This means that public administration is not covered by phase 2.

Tables 2 and 3 replicate Tables 3.3 and 3.4 from the IBES phase 2 summary report. They also include the same statistics estimated with the non-response adjusted weight. The statistics estimated using the weight provided with the data are almost identical to those in the summary report. Those estimated using the non-response adjusted weight are substantially larger. Revenues, costs, and profits are 56, 59 and 55% higher than when using the weights provided in the data.

As noted above, GSS only allowed DataFirst to take a 40% sample of the phase 2 firms for the public release data. The final phase two weights (design and non-response adjusted) in the public release data are thus those described above together with a further adjustment to take into account that only 40% of the firms were selected for this public release of phase 2. There is no adjustment for firms with 10 or more persons engaged since all firms of this size interviewed in phase 2 were selected by ourselves for the public release data.

Changes between Phase One and Two and Matching Errors

One important issue in the phase 2 data is that there are matched firms with phase 1 and phase 2 data that look like different phase 1 and phase 2 firms have been merged together, despite having the same identifiers. We have not attempted to solve this issue, but we do make some suggestions below for some simple robustness checks.

Table 4 shows firm size categories for phase 1 and 2. Strangely the reference period of total persons engaged in phase 2 was 10 months **before** the phase 1 reference period, even though phase 2 was conducted a year after phase 1. The phase 1 reporting date was August 2014 whilst it was June 2013 for phase 2. The unweighted median change in persons engaged was zero, the unweighted mean was 3 but the 5^{th} percentile was 17 and the 95^{th} was 44.

When we carefully explained some of the larger changes in employment this revealed that *some* of these changes are **probably not** the result of a merge of the wrong firms. For example, a trade union listed 1500 persons engaged in phase 1 but then only 4 in phase 2, with the correct industry classification in both phase 1 and 2. One sensible interpretation is that phase 1 was measurement error, and that the trade union incorrectly gave the number of union *members* in phase 1, rather than the persons engaged by the union. When looking at the 47 “firms” that changed more than 5 size categories, 47% were churches, which may have incorrectly listed their congregant numbers in phase 1, rather than the number of persons engaged by the church.

There are also many changes in industry classification between phase 1 and 2, some of which look reasonable and some of which look like the wrong firms were merged together. Around 15000 out of the roughly 24000 firms in phase 2 did not change industry, with the mean change being .4 of an

industry category (which could be between 1 and 101)[7]. When looking at both size and industry only 7209 firms out of 24000 do not change size category and industry. If we allow for the fact that more than 1 year separated the date which the phase 2 and phase 1 employment questions referred to by allowing firm size category to change 1 up or 1 down, then 12 500 firms had the same industry and a *similar* size category. If we allow industry stratum to change by plus or minus 5 (there were 102 industry categories) then nearly 16000, or two thirds of firms had relatively similar industry and size in phase 1 and phase 2.

Despite being able to explain some of the changes between phase 1 and phase 2, the industry and employment changes do suggest that there are at least some phase 1 and 2 "matches" that are actually different firms. This merging problem will affect weighted estimates from phase 2 because large firms from phase 1 with a probability of selection for phase 2 of 1 and thus design weights of 1 will have been merged with small firms from phase 2 that have very small output, revenues or costs and should have larger weights, and thus their contributions are understated. Similarly, small firms from phase 1 with quite large weights (reflecting their low probability of selection) seem to have been merged with large firms from phase 2, that have large output, revenues, or employment, and this will be overestimated when the big weight of the small firm is applied to this large firm output or employment. If this is equally likely for small and big firms (as seems to be the case from the table), then perhaps this may not result in substantial biases overall, but it is important to note it and it may well affect some analysis.

As an example, 1.5% of the 1 person engaged firms and 2% of the 2-4 persons engaged firms from phase 1 had 20+ persons engaged in phase 2. Because the sample design included small firms with a small probability their weights are large and because these small firms "grew" so substantially between phase 1 and phase 2 they ended up representing 12% of weighted total revenue of all phase 2 firms with more than 20+ persons engaged.

To check whether mismatches have any impact on analysis one check would be to examine variables that were asked in phase 1 and phase 2, such as ownership, legal form, the year the firm commenced operations etc. One can also identify firms that grew substantially between phase 1 and phase 2 and use a weight of 1 for these firms, instead of the design weight or our non-response adjusted weight. Alternatively they could be excluded. Most of these firms are excluded anyway in the public release data due only 10% of firms with 1-4 persons engaged in phase 1 being selected by ourselves for the public release.

[7] Taking the mean change in a categorical variable is not usually sensible. But given that similar industries have adjacent industry numbers using the mean change seems helpful in this context.

Setting the Complex Survey Design

As we have discussed above the IBES 2014 phase two was a stratified sample of phase one. In working with phase two researchers should set both the weight and stratum variable in any analysis. We recommend using the non-response adjusted weights. These are provided with the data. In Stata, the command would be:

```
svyset [pw=weight_nr], strata(stratum)
```

For some variables there are strata with only one firm and variances cannot be estimated. In this case the singleunit option can be specified in Stata:

```
svyset [pw=weight_nr], strata(stratum) singleunit(scaled)
```

Fixing the Form/Questionnaire Problem

Phase 2 data was captured using 9 different questionnaires (or "forms") depending on the sector of the firm. It is not always clear from the data which questionnaire a firm answered. In Appendix 2 we explain how we attempted to figure out which firms answered which questionnaires. Firms answering different questionnaires results in variables that are based on information collected in slightly different ways depending on the sub-sector. In some cases, the question number is all that has changed; elsewhere the wording of the question is different, sometimes even having a slightly different interpretation depending on the questionnaire. As such, it is very important that the user understand the variables in the context of the questionnaire that was answered.

The questionnaire that a phase 2 variable comes from is indicated as follows. If the variable only applies to one of the questionnaires, the variable name prefix reflects this. For example, a variable that can only be found on form 1 (agriculture) is renamed 1_*. If a variable applies to several of the forms, but not all, this is noted in the variable label. For example, revenues from contract work is labelled "Revenue from work done on contract (Forms 2,3A,3B,4,5)". The specific *question* that each variable was collected from is also noted in the label. In cases where the question wording or question number differs per questionnaire, this is indicated with an asterix * (for example 12.1*). Again, it is important that the user consult the questionnaires.

The table in appendix 1 provides further detail on all the variables in the cleaned dataset.

Description of Data

Phase two collected data on revenues, costs, profits, assets and asset purchases, input and output costs as well as other miscellaneous questions. The data GSS has provided for this public release is

revenues, costs, profits and assets. Input and Output product costs will be released once GSS releases the Supply Use Tables based on this data. Data on input and output quantities has not yet been captured and cleaned. This section provides detail on how revenues, costs and assets were collected in phase two.

Costs

Costs are captured similarly on all the questionnaires, in four places: labour (questions 3 and 4), fixed capital formation (question 6), purchase costs (question 9) and other operating costs (question 10). In this section we provide detail on how these four categories were collected and suggest how a total can be created for each. However, we do not release or recommend a formulation for overall costs, instead leaving it up to the user to combine them as they see fit.

Labour

Labour payments are collected in two questions: wages and salaries in question 3 and supplements to wages and salaries in question 4. Question 3 categorised the payments either by the level of the recipient (direct production workers vs. other employees) or by the form of the payment (cash vs. in kind). Which type of categorisation was used depends on the questionnaire type that the firm answered. In all cases there is also a total wages and salaries variable (3.3) which is the sum of both categories.

Question 4 splits supplements into the same two groups regardless of Q type: contributions for social security purposes and contributions for other reasons. There is also a total supplements variable (4.3). One can therefore easily create a total for labour payments by combining 3.3 with 4.3.

Purchase Costs

Purchases costs are collected in the same way on all the questionnaires: Table 9 collects the purchase costs of several items, being: raw materials, fuel, electricity, water, goods for resale, and other purchases. Variables corresponding neatly to these exist in the raw data, excluding the case of the *purchase costs of raw materials* variable[8]. The user can easily create a total purchase costs figure by summing the variables that correspond to line items 9.1 through 9.6.

Other Operating Costs

The final question for costs is question 10: other operating costs. This section is highly disaggregated in the Q, but the raw data only provides the total of these costs, which is row 10.28.

[8] See “difficulties with the purchase costs of raw materials” in the second appendix.

Revenues by Questionnaire

Unlike costs, the revenue data[9] collected vary by questionnaire. This complicates things, but fortunately the situation can be simplified by loosely grouping the questionnairess into four approaches: Agriculture (form 1), Industrial (forms 2,3A, 4,5), Services (forms 6 & 7) and Wholesale (form 8).

Agriculture (1)

Agricultural revenues are captured in two questions, being 11.1 (total revenues from agricultural products) and 11.2[10] (total revenues from other sources). In the public release we provide the total revenues from both questions, as well as some of the line items from 11.2. A "grand" total revenue for agricultural establishments can be created by simply summing the two totals.

Industrial (2, 3A, 4 and 5)

Industrial revenues are collected in the same way on forms 2, 3A, 4 and 5. Like agriculture, there is a section for industrial revenues (question 12) and then another for "non-industrial" revenues (question 13). The public release includes the line items in table 12. These depend on the questionnaire type, although the first item (12.1) is always revenues from sales of the "main" good (for example, a water company's sale of water). Note that the line-item of goods produced for own use (line 12.5) is missing from the raw data. Note also that for construction firms (form 5) this table is not number 12 but number 13. On the other hand, non-industrial revenues are included as a single total variable. Creating a "grand" total can thus be done by summing the industrial revenue items and adding this to non-industrial revenues.

Services (6 and 7)

Revenues collected for service companies are straightforward. There is a table in question 11 for revenues the main service provided (the number of which depends on the type of activity), and then another table for "other revenues". The user can create a grand total by adding the revenue from service income to the revenues from other income[11].

[9] The terms "revenues" and "receipts" are used somewhat ambiguously by GSS in the questionnaires and the data provided to Datafirst; we simply use "revenues" throughout.

[10] There is a numbering mistake in the agricultural questionnaire. Question 11 is split into two parts, which should be 11.1 and 11.2. 11.2 and its sub-items (for example 11.2.1, 11.2.2 etc) are correct, but the 11.1 is missing a digit. So 11.1.1 is labelled 11.1 erroneously in the form

[11] For forms 6-8 it appears that GSS had forgotten to add the subsidies and resale amounts to the "other income" amounts, so this was done by Datafirst.

Wholesale (8)

The wholesale establishment revenues are collected in a slightly different way to the rest. The questionnaire has three questions; question 11.1 (revenues from maintenance and repairs if applicable), 11.2 (revenues from sales as a wholesalers) and 12 (other incomes). These questions have been captured in the services, resale and other revenues variables respectively. This makes the information in these variables slightly different to the other questionnaires. For example, the resale variable is now not just a sub-item but really the primary source of income for the establishment. Creating a "grand" total of revenues for wholesalers requires summing these three totals.

Small Manufacturers (3B)

Form 3B differs to the rest in the case of revenues. In the data we have the total revenues from the main industrial activity (question 9.8), as well as revenues from contract work (question 9.13). A grand total for small manufacturers can be created by summing these two figures.

Asset Data

Data on assets as per question 6 of the questionnaires was received in a separate file from the rest of the phase 2 data and merged into the main data, with a perfect match (each firm receiving a record of its assets). Variables from the asset data were named and labelled using the same principles as the rest of the data.

Some entries were missing from the data: the first column of the question 6 table in the questionnaire (book value as at the beginning of financial year 2013) is missing. There is also a missing cell – we received no depreciation for buildings (column 4).

References

Ghana Statistical Service (2015). Integrated Business Establishment Survey Summary Report.

Ghana Statistical Service (2017). Integrated Business Establishment Survey Phase II Summary Report.

Kerr, A. and McDougall B., 2020. What is a Firm Census in a Developing Country? An Answer from Ghana. CSAE Working Paper WPS 2020-11.

Lohr, S. (2010). Sampling: Design and Analysis. 2nd Edition. Brookes/Cole.

Tables

Table 1: Firm and Persons Engaged

	Phase 1	Phase 2- GSS weight	Phase 2: NR adjusted weight
Total firms	638234	532707	631674
95% Confidence Interval		(528241 to 537173)	621906 641442
Total Persons Engaged	3383206	2564441	3091608
95% Confidence Interval		(2524776 to 2604105)	(3010182 to 3173034)

Table 2: Revenues, Costs and Gross Profits by Region

	GSS Weights			Non-response adjusted Weights		
Region	Revenue	Cost of Goods Sold	Gross profit	Revenue	Cost of Goods Sold	Gross profit
Western	25362	8480	16882	100149	14284	85865
Central	8518	3095	5424	10754	3752	7002
Greater Accra	304306	94212	210094	473840	176315	297525
Volta	5421	2640	2781	7318	3629	3689
Eastern	11268	5018	6250	14570	6214	8356
Ashanti	51536	21418	30118	52690	19410	33280
Brong Ahafo	11953	5530	6423	13718	5761	7957
Northern	10505	6496	4009	15130	9287	5843
Upper East	4870	1804	3066	12218	3532	8686
Upper West	2495	1145	1350	3281	1341	1941
All Regions	456942	156067	300875	714711	247905	466806

Table 3: Revenues, Costs and Gross Profits by Sector and Size Group

		GSS Weights			Non-response adjusted Weights		
Sector	Size Group	Revenue	Cost of Goods Sold	Gross profit	Revenue	Cost of Goods Sold	Gross profit
Agric	Large	513	119	394	3469	1153	2316
	Medium	1544	1045	499	2897	2023	874
	Small	2486	436	2050	1999	336	1662
	Micro	937	504	433	1219	604	615
	All Agric	5479	2104	3375	9584	4116	5468
Industry	Large	57904	18314	39589	202559	48676	153883
	Medium	27167	7851	19316	57341	14866	42475
	Small	45062	19747	25314	80078	33729	46349
	Micro	3010	1170	1841	3523	1399	2124
	All Industry	133143	47082	86060	343501	98670	244830

Services	Large	73984	11956	62028	68764	19377	49387
	Medium	68422	25148	43274	105536	46465	59071
	Small	99005	44467	54539	105536	46986	58550
	Micro	76909	25310	51599	81792	32291	49501
	All Services	318321	106881	211440	361627	145119	216508

Notes: this table replicates table 3.4 in the GSS IBES 2014 Summary report.

Table 4: Persons engaged Size categories for firms in both phase 1 and 2.

		Phase 1								
		1 PE	2-4 PE	4-9 PE	10-19 PE	20-49 PE	50-99 PE	100-499 PE	500+ PE	Total
Phase 2	1 PE	1,286	882	138	83	129	60	29	3	2,610
	2-4 PE	1,636	4,419	971	329	388	127	71	8	7,949
	4-9 PE	379	1,494	1,640	780	506	164	75	8	5,046
	10-19 PE	93	313	550	1,331	981	162	66	5	3,501
	20-49 PE	36	114	143	419	1,716	512	102	3	3,045
	50-99 PE	13	24	25	48	219	521	201	6	1,057
	100-499 PE	5	9	20	20	82	126	405	28	695
	500+ PE	0	2	4	2	19	9	23	67	126
	Total	3,448	7,257	3,491	3,012	4,040	1,681	972	128	24,029

Note: Own calculations from Phase 1 and 2 IBES data. PE is persons engaged.

Appendix 1: Table of Variables

PHASES 1 AND 2				
Variable Name	**Question**	**Description**	**Min**	**Max**
beid	00 (N/A)	**Unique Identification Number** for each establishment. Created by concatenating the region, district, sub-metro and environmental zone codes, and then adding a unique 3-digit establishment code for each unique firm at this level to distinguish them	double	
stratum	00 (N/A)	**Stratum** used to select sample for both phase 1 public release sample and phase 2 sample from survey design. Stratification was done by 6 firm size categories, 10 regions and 101 industrial sub-sectors.	1	1252
rural	00 (N/A)	**Rural/urban** identifier. NOT collected in IBES but derived from a dataset of EAs in the 2010 population census. Missing for 1.25% of the observations where matching was not possible.	categorical	

PHASE 2					
Variable Name	**Form**	**Q**	**Description**	**Min**	**Max**
weight	N/A		**Design weight** for phase 2 sample with adjustment for 40% sample by DataFirst	1	4112
weight_nr	N/A		**Design weight** for phase 2 sample with adjustment for 40% sample by DataFirst AND *adjusted for non-response*	1	5827
form	N/A		**Phase 2 questionnaire (form) answered by firm.** Note that the physical questionnaires “form 3A”and “Form 3B” are recoded to 9 and 10 in this variable. Variable takes values: 1) Agriculture (Form 1) 2) Mining & Quarrying (Form 2) 4) Electricity and Water (Form 4) 5) Construction (Form 5)	Categorical	

			6) Services 1 (Form 6) 7) Services 2 (Form 7) 8) Wholesale & Retail 9) Manufacturing – Large Firm (>=30 persons engaged, form 3A) 10) Manufacturing – Small Firm (<30 persons engaged, form 3B)		
district	all forms	1.1.6	**District Code** with text label (districts based on 2012 census)	101	1011
region	all forms	1.1.7	**Region Code** with text label	1	10
orgform	all forms	1.2.4	**Organisational Form** of the establishment. Variable takes values: 1) Head office 2) Single Establishment 3) Subsidiary	categorical	
ownership_type	all forms	1.2.5	Captures **Company Ownership Type**. Variable takes values: 1) State-owned 2) Privately owned 3) Public-Private Partnership	categorical	
owership_nation_type	all forms	1.2.6	**Nationality of Ownership**. Variable takes values: 1) Ghanaian 2) Non-Ghanaian 3) Mixed (Ghanaian and Non-Ghanaian)	categorical	
legal_type	all forms	1.2.7	**Type of Legal Organisation**. Variable takes values: 1) Sole Proprietorship 2) Partnership 3) Private Limited 4) Public Limited 5) Statutory 6) Other Governmental Institution 7) Quasi-government	categorical	

			8) Parastatal 9) NGO 10) Cooperative 11) Association/Group		
startyear	all forms	1.2.9	**Start year.** The year of commencement of 'business'	1670	2014
months	all forms	1.3.1	**Months of operation over the past year**	0	12
accounts	all forms	1.5	Existence of **formal accounts** in some form	dummy	
isic_4	all forms	1.6	**Four-digit ISIC revision 4 activity code (Class)** of "principal activity", defined as the activity that is the "main purpose of the establishment" or that which "accounts for the largest part of the value of output". The industry codes used are GSS codes which are based on ISIC Revision 4	111	9609
isic_3	all forms	1.6	**Three-digit ISIC revision 4 activity code (Group)** for the principal activity	11	960
isic_2	all forms	1.6	**Two-digit ISIC revision 4 activity code (Division)** for the principal activity	11	96
isic_section	all forms	1.6	The **ISIC Section** that that principal activity falls under. This is the highest level of aggregation from ISIC. These sections have been assigned numeric values (sequentially) which is something ISIC itself does not do. It was according to these ISIC Sections that the appropriate questionnaire types were inferred by DataFirst.	1	18
pe_employees	all forms	2.1a	**Total employed persons.** Employed persons are those working for pay. This includes the sum of operatives (directly involved in production) and other employees	0	10813
pe_unpaid	all forms	2.2a	**Total unpaid workers.** Includes proprietors, learners and family members	0	1179
pe_total	all forms	2.4a	**Total persons engaged.** All persons engaged by the establishment, including employees, unpaid workers and national service persons	0	10813

pe_X	all forms	2.1b/c 2.2b/c 2.4b/c	Six variables capturing a breakdown of **Persons engaged**. Specifically, total persons engaged by the establishment is broken down according to male vs. female and employed vs. unpaid	0	9904
paybyclass_oper	2, 3A, 5	3.1	**Cash and in-kind payments to operatives** (direct production workers)	0	7.1b
paybyclass_other	2, 3A, 5	3.2	**Cash and in-kind payments to other employees** (including directors)	0	167m
paybyclass_total	2, 3A, 5	3.3	**Total cash and in-kind payments to employees**	0	7.26b
paybyform_cash	1, 3B, 4, 6, 7, 8	3.1	**Cash payments to employees**	0	91.2m
paybyform_kind	1, 3B, 4, 6, 7, 8	3.2	**In kind payments to employees**	0	64.9m
paybyform_total	1, 3B, 4, 6, 7, 8	3.3	**Total payments to employees**	0	91.2m
supp_social_security	all forms	4.1	**Supplements to wage & salary payments for social security**	0	12.6m
supp_other	all forms	4.2	**Supplements to wage & salary payments for other reasons**	0	43.0m
supp_total	all forms	4.3	**Total supplements** to wages & salaries	0	33.3m
stock_other_op	all forms	5*	**Total value of other stock** at opening of financial year	0	279m
stock_other_cl	all forms	5*	**Total value of other stock** at close of financial year	0	565m
stock_total_op	all forms	5*	**Total value of stock** at opening of financial year	0	1.31b
stock_total_cl	all forms	5*	**Total value of stock** at close of financial year	0	1.38b
Assets_X	All forms	6.X	31 variables capturing **value of fixed assets**		
materials_purchased_total	1,2,3A,4,5, 3B,6,7,8	7.11 7.2.11	**Total purchase costs of raw materials** (purchasers' prices). This is total over several line items of input costs. See note in appendix 3.		
costs_rawmat	All forms	**9.1**	**Purchase cost of raw materials, supplies, etc. purchased.** Should in theory be equal to *materials_purchased_total*. See note in appendix 3.	0	1.68b
costs_fuel_total	all except 3B 3B	**9.2** 7.3.5	**Total purchase costs of fuel** for transport and operation (excludes fuels in final product and fuel produced and consumed in the establishment)	0	822m
costs_elec	all except 3B 3B	**9.3** 7.3.2	**Purchase costs of electricity**	0	122m

costs_wate	all except 3B 3B	**9.4** 7.3.3	**Purchase costs of water**	0	18.3m
costs_good	all except 3B 3B	9.5 7.4	**Purchase costs of goods for resale**	0	810m
costs_other	all except 3B 3B	9.6 7.3.12	**Purchase costs of other purchases**	0	189m
costs_other_oper	all except 3B 3B	10.28 7.3.13	**Total other operating costs.** Note that for form 3B this is total indirect costs which is qualitatively different	0	616m
rev_agri	1	11.1.11	**Total revenue from agricultural production.** Note that there is a numbering mistake in the agricultural questionnaire. Question 11 is split into two parts, which should be 11.1 and 11.2. 11.2 and its sub-items (for example 11.2.1, 11.2.2 etc) are correct, but the 11.1 is missing a digit. So 11.1.1 is labelled 11.1 erroneously in the form	0	753m
rev_nonagri	1	11.2.13	**Total Revenue from non-agricultural production** ("other revenue")	0	18.3m
rev_resale	1	11.2.3	**Revenue from resale** of livestock	0	13750
ev_ownconc	1	11.2.4	**Value of livestock produced and consumed on the farm**	0	12000
rev_subs	1	11.2.5	**Revenue from agricultural subsidies**	0	336000
rev_subs	6 & 7 8	11* 12.1.5	**Revenue from subsidies and grants.** Note that the question number in the services forms depends on the type of activity the firm does	0	40m
rev_main	2, 3A, 4 3B 5	**12.1** 8.1 13.1	**Revenue from main industrial sales/activity.** This is revenue from the main output of the establishment. For example, for mining or electricity it is sales income from sale of electricity; for construction firms it is revenue from construction activities. This question was not asked for service firms. Note also that for wholesale firms GSS considers their income from resale a "service". So main revenues for wholesalers is also not captured here but rather in rev_services below	0	2.08b

rev_contract	2, 3A, 4 3B 5	**12.2** 9.13 12.4	**Revenue from contract work done for other companies** (not the same establishment)	0	107m
rev_repair	2, 3A, 4 5	**12.3** 13.3	**Revenue from repair and installation services**	0	57m
rev_otherind	2, 3A, 4 5	12.6 **12.4** 13.6	**Revenue from other industrial services.** These are revenues from industrial services/activities other than the primary activity of the firm as in rev_main		339m
rev_resale	2 3A, 4 5 6, 7 8	12.4 **12.5** 13.4 12.1.2 11.2	**Revenue from resale** of goods sold *in the same condition as purchased*. Note that for wholesale firms GSS considers their main sales income (from the 11.2 tables) revenues from resale (see questionnaire)	0	1.01b
rev_nonind	2, 3A, 4 5	**13.8** 14.7	**Total revenue from non-industrial services / activity.** For construction firms this is revenue from non-construction activities	0	2.51b
rev_services	6 & 7 8	11* 11.1*	**Revenue from service income.** The question number in the services forms depends on the type of activity the firm does. Note that for wholesale firms service revenues are being captured by table 11.1.	0	5.69b
rev_otherinc	6 - 8	12.1	**Revenue** from other income	0	238m
cogs_final			A variable that seems to measure **cost of goods sold** was provided by GSS. It is not clear what exactly it is. See the note in appendix 3.	-154k	1.6b
totrev			A **Total Revenue** variable released by GSS that was used in the GSS phase 2 summary report. Because some of the variables used to create the variable were not released, DataFirst could not infer exactly how GSS created this variable. See note in appendix 3.	0	5.67b
totgrossprofit			A **Gross Profit** variable released by GSS. This variable suffers the same problems as *GSS_totrev* above. It is not clear how it was constructed, and it was not replicable by DataFirst. Unfortunately, it is also not	-48	5.57b

			possible to tell how it was constructed by using a simple accounting identity (profit=revenues-costs), as we do not know how the total revenues were calculated. A few manual checks confirmed there is no straightforward relationship between the costs and revenues variables and these “GSS” revenue and profit variables. Again, it is up to the user to decide how to proceed.		

Appendix 2: Detail on Fixing the Qtype/Form Problem

The most reliable way to tell which questionnaire a firm answered is to look at the costs and revenues variables. This is because each questionnaire has a different structure which results in a different set of variables being answered. For example, a firm answering the agriculture questionnaire should have values for the agricultural revenue variable but missing values for the revenues from main industrial activity.

The raw data provided to Datafirst included a variable called "qtype" which was meant to identify the questionnaire that a given observation (establishment) had answered. Unfortunately, it appears this variable was erroneous in around 5% of cases. For example, we have industrial revenues and no agricultural revenues for a firm that the qtype variable said was agricultural.

Fortunately, there is another variable called *isi2min* which seems to correctly identify the qtype as it corresponds perfectly with the revenue variables observed. We proceed on the basis that the original qtype variable provided to Datafirst was wrong but can be corrected by using the *isic2min* variable. Doing so fixes the qtype and provides a neat correspondence between the variables for which answers for and the (new) qtype identifier. For example, firms recorded as having answered the agriculture Q now have revenues recorded in the agricultural variables but not the others, and *vice-versa* for the other industries.

The last piece of this puzzle is the manufacturing firms. There were two manufacturing questionnaires used depending on establishment size. Unfortunately, the *isic2min* variable does not make clear which manufacturing firms were answering the "small" and which the "large" questionnaire. Our approach is to once again turn to the revenue variables. Form 3A (manufacturing large) has a question for revenues from repairs, and from "non-industrial services", whereas 3B does not. Fortunately, it appears that the old qtype variable was highly accurate for these firms, as only 26/2615 firms required adjustment for the qtype to match the variables observed.

One concern with these firms is that the firm size variable does not correspond neatly to the new or old qtype variable. For example, we have firms that have more than 30 persons engaged recorded as having answered the "small" (<30 persons engaged) questionnaire. However, we proceed on the basis that the revenues variables are still the most important indicator of the form that was answered as opposed to the firm size recorded. That the firm size might be different to what the questionnaire requests is not impossible – a firm might be given the large questionnaire, but then when looking more closely discover engages fewer persons than expected. This is more likely if the

questionnaires were handed out based on p1 firm sizes. Another explanation is that enumerators were simple using the wrong form.

In summary, to fix the qtype variable we create a new variable called form based on the isic2min and revenue variables, which corrects the original one and provides the user with a questionnaire variable that correctly corresponds to the questions answered in the data. The table on the following page summarizes the process above, showing the new form variable (col 1) vs the old qtype variable (col 2), the other variables used to identify the correct form (cols 3-6), and the number of firms that are affected by the adjustment (col 7).

Table: Qtype Adjustments

Variables that supposedly identify the Questionnaire		**Variables we can use to see which Q was really answered. Read as follows: If qtype is really (col 1), this variable should be:**				**n wrong in qtype/total in qtype (note: these numbers from full p2 sample)**
form new (our creation)	qtype old (raw)	Mainrec	repairec	nonindrec	isic2min	firms moved out/raw number
1 Agriculture	T1	missing	Not Used		1. Agriculture, forestry & fishing	11/574
2. Mining	T2	has values			2. Mining and quarrying	none
4. Elec & Water	T4	has values			4. Electricity, gas, steam and 5. Water supply; sewerage, waste	25/223
5. Construction	T5	has values			6. Construction	29/611
6. Services 1	T6	missing			8. Transportation and storage 9. Accommodation and food 10. Information and comm 12. Real estate 14. Administrative and support 17. Arts, entertainment and rec 18. Other service activities	563/7512
7. Services 2	T7	missing			11. Financial and insurance 13. Professional, scientific 15. Education 16. Human health and social	241/4194
8. Wholesale	T8	missing			7. Wholesale and retail, repair	197/5942
9. Manufacturing L	T3	has values	values	values	3. Manufacturing	2/693
10. Manufacturing S	t9	has values	missing	missing	3. Manufacturing	104/5091

Appendix 3: Miscellaneous Notes

A note on the GSS Variables

Above we have suggested ways that users can combine costs and revenues variables to produce totals for analysis. In the data provided to Datafirst there was three variables that seem to be pre-calculated totals: total revenue (*totalrev*), gross profit (*grossprofit_final)* and the total cost of goods sold (*cogs_final)*. Unfortunately, how these were constructed was not well documented and looking at the data doesn't provide an obvious answer. Further, it seems that GSS sometimes made mistakes in these totals, for example excluding non-agricultural revenue from total revenues for agricultural firms. We opt to leave these "GSS" variables in the data should the user wish to replicate the results produced by GSS or try figure out how these variables came to be. However, we advise that users treat the variables with caution, and also that they create their own totals using this guide, their own judgements and the questionnaires.

A Note on Form 3B

Form 3B is used to capture data from small manufacturing firms. It differs somewhat from all the other forms although there is some overlap. Importantly, the question numbers are shuffled around for 3B even if the variables are measuring the same thing. Rather than repeatedly make exceptions throughout the discussion above, we have preferred to ignore this issue in the discussion. A detailed record of the how the question number differs in 3B compared to the other forms is included in the table of variables in appendix 1.

Fixing the Purchase Costs of Raw Materials problem

Unlike the other cost variables, how to use the information that was related to the purchase costs of raw materials was not straightforward. In this case some assumptions were necessary, which are documented here.

In the raw data there are two variables which seem to be measuring total purchase costs of raw materials. In the questionnaires there are also two places that the total purchase costs of raw materials are asked – once as a total at the bottom of table 7, and later again as line item 9.1. Based on the variable labels[12], which correspond loosely to the text in the Qs, I assume that the total from table 7 is being captured in the *purchases* variable and that line item 9.1 is being captured by *totrawmat* from the raw data. These have been renamed *materials_purchased_total* and costs_rawmat to reflect these assumptions.

[12] The *purchases* variable was labelled "*total* raw material purchased" whereas the label of *totrawmat* read "Purchase of raw materials, supplies" which matches 9.1 on the Q.

These assumptions are straightforward and likely correct. An outstanding problem is the two variables are not always equal, which they should be according to the questionnaires . It is not clear which is correct (if either), why they differ or how to choose one over the other. A reasonable approach might be to assume that *purchased* is more accurate as it comes after the manual calculation in table 7 and probably leads to better recall. Rather than make any further assumptions we opt to leave it up to the user to decide how to handle this situation.