Sizing and profiling the Small, Medium and Micro Business Market in South Africa

Galpin, Jacky
University of the Witwatersrand, School of Statistics and Actuarial Science
1 Jan Smuts Avenue
Johannesburg (2000), South Africa
jacky@galpin.co.za
Neethling, Ariane
University of the Free State and University of Stellenbosch
Ariane_Neethling@yahoo.com

Introduction

FinMark Trust is an independent Trust, established with funding from the United Kingdom Department for International Development, with the objective of "Making financial markets work for the poor" in Africa. One aim is to build a picture of the informal business sector, both as to the size and characteristics, and well as the role they can play in developing countries. Businesses range from very informal (such as vendors on street corners and hawkers) to semi-formal (such as those running a garden service or computer repair business), to more formal registered businesses, with a formal office.

The 2010 Finscope™ survey targeted Small, Micro and Medium Enterprises (SMMEs) in South Africa (SA). A nationally representative sample of business owners aged 16+, with less than 200 employees, was drawn. The objectives of the survey were to estimate the size of the small business market in SA, to quantify the number of people engaged in small business activities, and to profile the businesses.

Sample design

A stratified random sample of 1000 enumerator areas (EAs) was drawn, representative of SA at national, provincial and geo-type levels (Finscope, 2010). Probability proportional to size sampling was used, with the estimated number of households per EA in 2009 being used as the measure of size. The dominant race group (Black, Coloured, Asian, White) of the EA was used as a further stratification variable, to ensure that a representative sample of all race groups was obtained. Power rule allocation (power = 0.7) was used to ensure an adequate sample size in each of the strata, this disproportionate allocation procedure being recommended for surveys with numerous small strata where there is a need for relatively precise estimates at each stratum level (Lehtonen & Pahkinen, 1994). The distribution of the EAs is shown in Table 1.

The dwelling units (DUs) in each chosen EA were listed. A step-size was chosen to yield six segments for the EA, as the aim was to obtain six interviews with small business owners in each selected EA. From each starting point successive dwellings in the segment (as marked on a map supplied to the interviewers) were visited, with contact information being listed, as well as whether any of the household members was involved in a small business. On finding a dwelling with a SMME, a full interview was conducted, and the interviewer progressed to the next starting point. This gave the hit rate information for each EA, namely the number of dwelling with no „success", before a success was obtained. The number of successful interviews in the EA was also recorded. A total of 5676 interviews with SMMEs were obtained.

Province	Geo-area	Dominant race group				Overall total per Geo-area and province
		Blacks	Coloureds	Asians	Whites	
Western Cape	Urban	37	43	2	37	119
	Rural	0	13	0	0	13
	Total	**37**	**56**	**2**	**37**	**132**
Eastern Cape	Urban	34	15	2	18	69
	Rural	52	4	0	0	56
	Total	**86**	**19**	**2**	**18**	**125**
Northern Cape	Urban	13	15	0	7	35
	Rural	7	6	0	0	13
	Total	**20**	**21**	**0**	**7**	**48**
Free State	Urban	40	5	0	16	61
	Rural	16	1	0	0	17
	Total	**56**	**6**	**0**	**16**	**78**
Kwazulu Natal	Urban	45	7	29	27	108
	Rural	58	0	0	0	58
	Total	**103**	**7**	**29**	**27**	**166**
North West	Urban	27	3	2	16	48
	Rural	36	0	0	0	36
	Total	**63**	**3**	**2**	**16**	**84**
Gauteng	Urban	102	14	12	59	187
	Rural	10	0	0	1	11
	Total	**112**	**14**	**12**	**60**	**198**
Mpumalanga	Urban	26	2	1	14	43
	Rural	36	0	0	0	36
	Total	**62**	**2**	**1**	**14**	**79**
Limpopo	Urban	12	0	1	10	23
	Rural	67	0	0	0	67
	Total	**79**	**0**	**1**	**10**	**90**
Overall total per race:		**618**	**128**	**49**	**205**	**1000**

Table 1: Distribution of the EAs for the stratified random sample

Data were weighted to the population figures based on the EA inclusion probability, the inclusion probability of a household, and the weight of a person having one or more small businesses. The negative binomial approach was used to determine the inclusion probability of a household, by taking the number of „failures" (households with no small business owners) into account. The final weights were used to estimate the number of SMME's. It is estimated that there are just under 6 million small businesses in SA, with 5.6 million small business owners. The geographical spread of these, relative to the population, is shown in Figure 1.

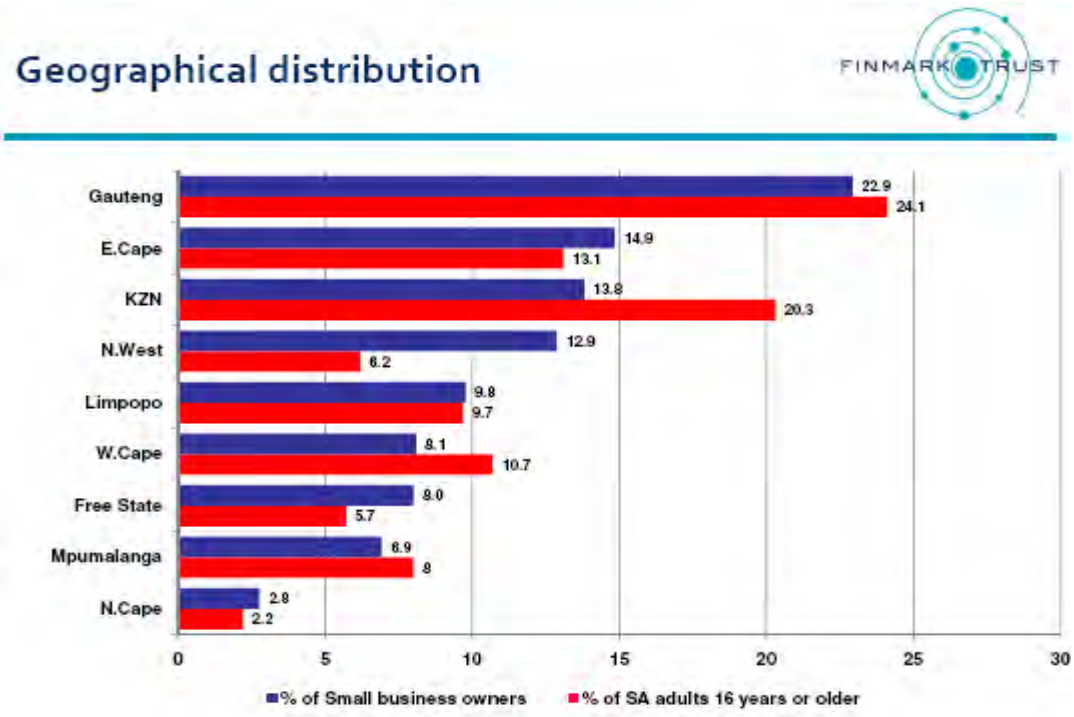

Figure 1: Geographical distribution of small businesses and population (Finscope, 2010)

Profiling the small businesses

In order to profile the businesses, a Business Sophistication Measures (BSM) was created**.** Questions used concerned „hard facts", such as where the business operates from (e.g. street corner, no fixed address, house, office block) as well as access to and use of a number of services such as water, electricity, computers, banking and insurance. The responses were coded as 1=yes or 0=no, which places all 181 questions on the same scale, making principal component analysis (PCA) the appropriate technique for creating an index. The data were weighted up to the estimated population size.

The first principal component explained 14.3% of the variance, and formed the BSM index. The sensitivity of the analyses was investigated by omitting questions with very few informants in either the yes or no categories. Two scenarios were investigated, omitting variables with fewer than 10, and fewer than 30 informants in one of the categories. A k-means cluster analysis was used to group informants into similar groups. Groups with small numbers of informants (i.e. informants who differ from the others) were omitted from the PCA, in order to check the sensitivity of index. The results of these investigations showed little sensitivity of the major principal components to the combinations of variables and informants.

The resulting principal component scores were divided into equi-sized groups, giving and initial ranking of the businesses. Initially, 20 groups were formed, which would allow for merging of similar groups, to obtain a smaller number of distinct groups. The choice of 20 initial groups resulted in approximately 284 equivalent informants per group. (Equivalent informants are obtained by rescaling the sum of the weighted informants from the population size to the sample size.) The groups could then be examined as to their characteristics, and „similar" groups could be merged, and anomalous groups split.

The stability of these groups was investigated using discriminant analysis (DA), using the groups and the variables used to create the scores. The success in recovering the original 20 groups is shown in Table 2. The highlighting indicates the number of equivalent informants correctly assigned by the DA, while the last column gives the percentage correctly classified for each 20 individual group. A total of 54.1% were correctly classified.

Gp	#infs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	%
1	320	315	6																			98
2	284	81	132	31	21	18																47
3	255	48	40	126	22	17	1															49
4	261	27	37	15	140	26	4	12														54
5	296	22	18	16	50	123	27	30	4	4		1										42
6	285	21	7	5	9	21	122	64	24	11												43
7	291	6	12	4	9	27	32	142	17	35		6										49
8	281	8	1	6	8	7	13	76	81	47	28	6	1									29
9	281	11	3		3	9	7	22	18	131	41	17	18	1								47
10	293	3	2	5	11	3	6	24	14	46	93	22	44	16	3							32
11	280	3		2		1	6	3	9	9	47	99	56	33	8	4						35
12	277	2		4		1	4	17	1	3	22	23	152	25	9	13	1					55
13	286						1	2	2	1	15	30	36	117	58	23	3					41
14	210				4	3		8	6	4	2	8	21	23	80	34	18					38
15	352	1				1	1	1	1	1	5	11	27	20	19	193	69	2				55
16	289							1	1		5	1	9	8	15	36	197	15				68
17	286					1				2		2	4	1	6	11	55	163	42			57
18	282												1	2	1		17	39	208	12		74
19	285											1				1	2	5	42	230	4	81
20	282																		2	21	259	92

Table 2: Comparison of the classification of BSM groups (rows) and DA (columns)

BSM group	A (8)	B (7)	C (7)	D (6)
1 -3	84.8	89.4	93.3	93.3
4	70.3			
5		68.3		
6			76.8	76.8
7 -8	69.7			
9-10		78		
11-13	71.5		76.6	76.6
14	73	73		
15-16			72	85.4
17 -18	80.6	80.6	80.6	
19	83.4	83.4	83.4	83.4
20	91.5	91.5	91.5	91.5
% correct	**75.6**	**79.8**	**80.6**	**82.9**

Table 3: Combinations of the 20 groups, showing the percentages correctly classified by the DA

Possible combinations of these groups were investigated, in order to determine „similar" BSM categories, namely those for which the reconstruction by the DA was satisfactory (at least 70% for all groups). Table 3 shows the results for 4 possible combinations, resulting in 6-8 final groups, together with the overall percentage of correct classification. These groups were then profiled in terms of the variables, to allow the Finscope experts

to assess the coherence of the groups with respect to interpretation in business terms. The 8 group solution (scenario A) was chosen on the grounds of providing reasonable differentiation and business usefulness.

Investigation of variables discriminating between the BSM groups, using Chi-squared automatic interaction detection (CHAID)

Chi-squared automatic interaction detection (CHAID) was used to profile of the BSM groups (Kass, 1980, Hawkins and Kass, 1994). At the first step, CHAID looks at which of the predictors differentiate between the 8 BSM categories. The most discriminating variable was „do not have any insurance"=1, and 0="have some insurance" (p=1.6e-809 for the contingency table). At the second step, CHAID examines each of the new nodes to determine the most significant predictor for each node. The stronger discriminator between the BSM groups with some insurance, was „own, lease or hire: internet" (p=3.6e-96). For the BSM groups who did not have any insurance, the strongest predictor was „do not use a bank for the business" p=1.9e-588). Table 4 shows the details of the CHAID, against the 8 BSM groups, showing increase in sophistication for the BSM groups. Percentages above 10 are highlighted.

				BSM1	BSM2	BSM3	BSM4	BSM5	BSM6	BSM7	BSM8	Infs
No insurance	No bank used for business	No hot running water inside		54.8	35.4	8.2	1.1	0.2	0.2			2010
		No inside toilet		4.2	34.0	42.5	10.7	6.0	2.6	0.1		740
	Bank used for business	Business registered	No up-to-date financial records	0.7	23.4	50.2	15.3	6.6	3.8			687
	No bank used for business	Hot running water inside			6.3	63.8	19.4	8.8	1.9			160
	Bank used for business	Business registered	Up-to-date financial records		0.7	17.6	28.0	37.6	14.2	2.0		917
	No bank used for business	Toilet inside			0.7	24.5	26.6	21.7	22.4	4.2		142
	Bank used for business	Not registered	No running water inside			4.5	17.4	29.2	39.9	8.4	0.6	177
		Not registered	Running water inside				0.4	4.0	61.2	26.3	8.3	278
Some insurance	No internet	Registered				0.6	1.3	13.2	48.4	30.2	6.3	159
	No internet	Not Registered							15.8	45.8	38.3	241
	Lease/own internet									4.9	95.1	165

Table 4: Characterisation of the 8 BSM groups

It should be noted that, since CHAID is aimed at determining the most predictive variables and splits at each stage, the p-values can be interpreted as giving a ranking of the usefulness of the variables. So although all variables used are significant at the 1% level, the ratio of the p-values gives an indication of the relative predictive power of the variables. The no insurance variable is approximately 10^7 times more predictive than the next strongest predictor: „have a credit or debit card".

Looking at the number of employees in the businesses in the different BSM groups, shown in Figure 2, it can be seen that the lower BSM groups essentially have less than 10 employees.

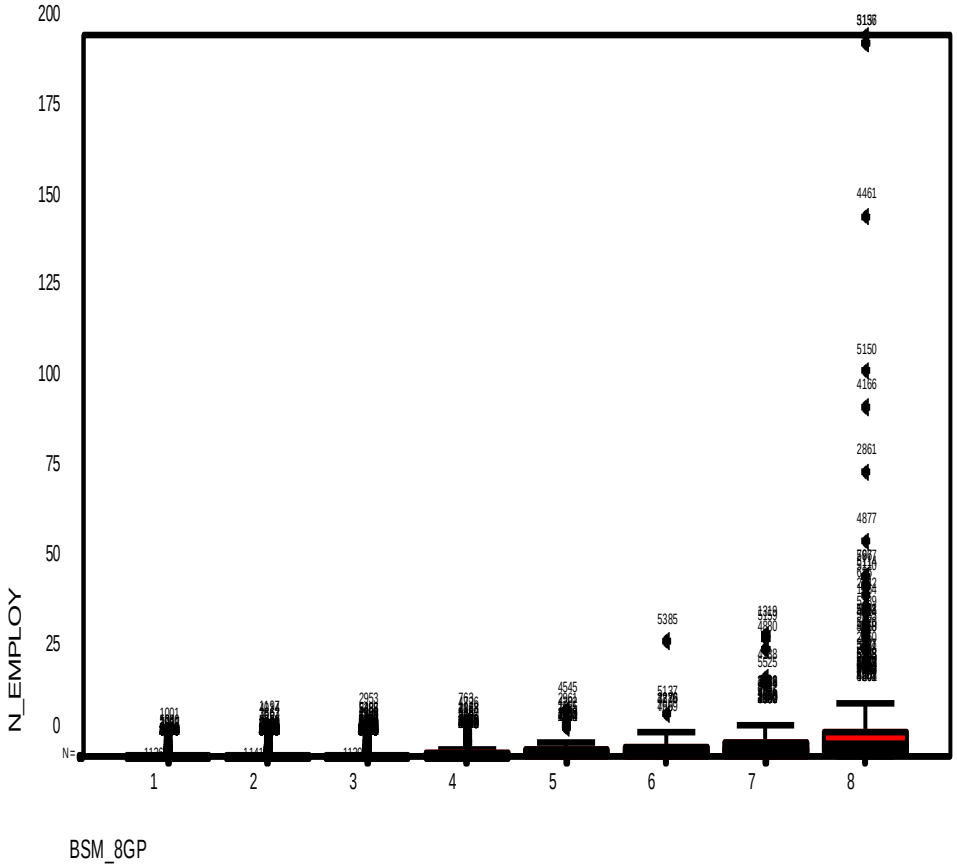

Figure 2: distribution of the number of employees by BSM group

Conclusions

The BSM index has been interrogated by users, and has been accepted as a useful categorization.

References

Finscope (2010). *Finscope South Africa Small Business Survey 2010*. Finmark Trust, South Africa.

Kass, GV (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society, Series C (Applied Statistics),* Vol. 29, No. 2, pp 119-127.

Hawkins, DM and Kass, GV. (1982) Automatic Interaction Detection. In: Hawkins, DM. (Ed) *Topics in Applied Multivariate Analysis.* Cambridge University Press: Cambridge.

Lehtonen, R. and Pahkinen, E.J. (1994) *Practical Methods for Design and Analysis of Complex Surveys*. John Wiley & Sons, New York.