Estimating the number of garment factories in Bangladesh

Boldi, Marc-Olivier

UNIVERSITÉ
DE GENÈVE

FACULTÉ D'ÉCONOMIE
ET DE MANAGEMENT

Estimating the number of garment factories in Bangladesh

10.08.2015

Author
Marc-Olivier Boldi, collaborateur scientifique
Faculté GSEM - Institut de Recherche en Statistique
Groupe de consultation en statistique
Bureau 3239
Tel (022 37) 98837
UNI MAIL, Bd du Pont-d'Arve 40, CH-1211 GENEVE 4

Table of Content

I. Context

For a study about the size of the garment industry in Bangladesh, a data base about the garment factories in this country have been built by The Center for Business and Human Rights (NYU – STERN). The data base contains information about 7169 factories. Although, the database contains information about the structure, locations, size (etc.) of the factories, the aim of this report is a statistical analysis of the number of factories in the country.

Indeed, these factories are observed from 5 professional lists: BGMEA, DIFE, BKMEA, ACCORD and ALLIANCE. It is clear that only a part of the factories have been observed and that a potentially very large number remains hidden.

The methodology of the estimation, namely the Multiple Systems Estimation, and main definitions are presented in Section II. Section III presents descriptive statistics and the results from MSE database. Section IV is a short discussion of the results. Further technical results can be found in the Appendix and in the Reference section.

II. Methodology

A. Multiple Systems Estimation (MSE)

The problem is to estimate the unseen part of a population the garment factories. These factories are partially observed in lists but a part is still hidden. The total population size is called the abundance. Actually, this kind of problems is fairly common in ecology where the abundance estimation is of interest. The problem can be illustrated with the well-known problem of fish abundance estimation.

Assume that N, the number of fishes in a lake, is to be estimated. A first catch is made and n1 fishes are being caught, marked, and released back in the lake. At that moment, the proportion of marked fishes in the lack is n1/N. Another catch is made: n2 fishes are caught, among them n12 are marked. Assuming that the chance of capturing a given fish is the same between the two catches, then the proportion of marked fishes in the second catch should be the same as the one in the whole lake, that is n1/N = n12/n2. One concludes that N=n1*n2/n12, which is the estimate of the fish abundance in the lake. This is called a capture-recapture experience.

This simple theoretical experience has been extended in complexity and applied to the field of survey on human beings. The idea is not to compare men to fishes, but to recognize that the problem is mathematically the same. If two independent but equivalent lists count the number of addict people in a city, not seeing them all, then we may expect that the proportion of addict people in the second list also seen in the first list (n12/n2) is the same as the proportion of addict people in the population seen in the first list (n1/N). Thus N is estimated by N=n1*n2/n12, the same way as for capture-recapture. In that context, the methodology is called Multiple Systems Estimation. A good review can be found in (Lum, Price, & Banks, 2013).

The formula extends to more lists in more complex mathematical way. However the complexity here is not in the math. It is in the fact that, even if we have a formula to estimate N, we have a huge uncertainty on that estimate. This has nothing to do with the math and its complexity. This has to do with the fact that the experiment, in fact, contains very few information about the abundance. This difficulty cannot be solved by math or stats whose role is to reveal an existing information from data, not to create new information.

This can be illustrated by the formula itself N=n1*n2/n12. If N is very large, then it is unlikely that someone is seen on the two independent lists. On the other hand, a large n12 (relative to N) is the sign that N is not large and that almost all the individuals have been seen in n1 and n2. Of course, all this is true when the two lists are independent. If the two lists are very dependent[1] then it is very likely that the two lists contains almost the same individuals, even when N is very large. In such a case, the formula is not valid anymore, and a correction term should be included in the numerator n12 to take into account for the dependence between the two lists. Ultimately, when the two lists are the same, then we have only one list from which no meaningful estimate of N can be obtained.

We see that the initial assumption that the two lists are independent or not conveys a lot of information about the abundance estimate. The fact is that, either we know N and can estimate the dependence between the two lists, or we know the dependence and then we can estimate N. No mathematical model in the world can overcome this without incorporating a prior information about N or about the dependence between the lists explicitly or implicitly[2].

Fortunately, when more than two lists are available, the number of observations increases, bringing more information. It should be noted that the number of parameters to estimate also increases. For example, with three lists the dependence between each pair of lists should be estimated (p12, p13, and p23)[3] but also the triple dependence factor (p123). Thus the problem again cannot be completely solved. However, it is intuitively easier to assume some form of independence between the three lists altogether than between any two lists[4].

This overall problematic is in link with the model choice that is treated below. It should also be noted that other problems are central to this application:

- The open/closed population assumption: all along we have assumed that the population is closed. This is often an approximation of the reality, sometimes quite crude. Capture-Recapture models for open populations aim at estimating the parameters of the birth-end-death process rather than the abundance (which is not defined). Due to the short amount of time between the various list surveys, in this study, it is assumed that the population is closed although in future researches this could be relaxed.
- The inhomogeneity of record probabilities: it is possible that a given unit has not the same chance of being caught by one list rather than another. For example, when one list is specialized in some garment style production then the corresponding factories have higher chance of being caught by that list. Although this is for sure the case in the present study, it should be noted that available models for tackling with this are so complex that they would be inappropriate. In addition, in average, the inhomogeneity effect is in general lower than the between list dependence effect[5].
- The credibility of the record: maybe some of the records should not have been counted. The recognition of a given factory cannot be solely based on its name. First because it may not have a name, second because when it has one, this name may not be clearly defined, or recognizable from a western perspective. This has been tackled at the data level, with a huge work of human recognition. The data are assumed clean enough that the mistakes, due to this imperfect though very serious work, are not influencing the final result in a significant way (compare to the precision which is aimed at).

[1] A classic case of dependence is when one list is national and another is regional.
[2] The incorporation of prior information is not in itself a solution, above all when no such information is available.
[3] The sign p12 refers to the proportion of people seen in list 1 and 2 simultaneously.
[4] In general, this is of the form p123=p12*p3 or p123=p1*p23, etc.
[5] It is also sometimes confounded with this effect, meaning that tackling the dependence partially solves the inhomogeneity.

B. The Poisson model family and Bayesian Model Averaging

The analysis of five lists of counts can be done following several principles that can be found in (Hoeting, Madigan, Raftery, & Volinsky, 1999). The association Human Rights Data Analysis Group (HRDAG) has made a huge work for bringing these complex methods to an easy access. This section is freely inspired from the very good text in (HRDAG, 2015) (see Q14 therein).

An appropriate family of models the Poisson generalized linear model family[6]. In example for 3 lists, a Poisson model fits the observed cell counts to a formula like

log(m100) = a + b1

log(m010)=a+b2

log(m101)=a+b1+b3+b13 (etc.)

where m100 is the count of items only seen in list 1, m010 is the count of items in list 2 only, m101 is the count of items seen in both list 1 and 3 (not in list 2), etc. From such a model, the estimate of the unseen item count m000 is

log(m000)=a

that is m000 is estimated by the exponential[7] of a. The abundance is thus estimated by the sum of the seen items (in the data) and exp(a).

In the equation, the term b13 is an interaction term. That is, it estimates the dependence between lists 1 and 3. Assuming b13=0 means that they are independent. In theory, there could be a triple interaction term b123. However, as the count m000 is not observed, this term cannot be estimated in the model. Thus, one often assumes b123=0. Note that one could also have chosen b12=0, and incorporate b123 into the model, or assume b1=0 also. This is the whole problem of model choice.

From a pure statistical perspective, several model choices are compared using a numerical quality criteria[8]. However, in practice, the model choice is guided by the user considering for example that a 3-way interactions (terms like b123) should not be included if the two-way interactions are absent (terms like b12 and b13).

Any model carries an uncertainty on the estimate. The uncertainty on the abundance can be represented by a confidence interval. A confidence interval covers the true abundance value with a given confidence. The higher the confidence, the larger the interval. In practice, the confidence level is set to 95%. The construction of the confidence interval is obtained from the statistical properties of the model. When fitted to the data, the model brings also an estimate of the uncertainty "sa" on the parameter of interest here, a. Using the statistical properties of the estimate a, it is known that a confidence interval for a at 95% is given by the formula

[a – 1.96*sa; a + 1.96*sa]

The confidence interval for the abundance N is thus

Nobs + [exp(a – 1.96*sa); exp(a + 1.96*sa)]

---

[6] No link with the fishes.
[7] Parameter a is the intercept of the model.
[8] The criterion is often the BIC, a tradeoff between a goodness-of-fit measure and a complexity measure of the model.

In addition to the model uncertainty, the choice of the model itself conveys an uncertainty. To take this into account, Bayesian Model Averaging (BMA) can be used. The method consists in weighting the possible models according to a compromise between a prior probability and its appropriateness to the data. The final estimate is then a compromise between all the possible rather than being the result of a single model (even if it is the best, it may be close to the second model).

In addition, variations along the various models can be integrated into the uncertainty. Then the fact that choosing the model itself increase the final uncertainty is taken into account. The complexity of these methods exceeds that report. We refer to (Hoeting, Madigan, Raftery, & Volinsky, 1999) for full details.

C. Synthesis

The Poisson model family is used. The fact that some interaction terms expressing the dependence between the series are included or not into the model provides a family of model in which a selection should be done. The selection is performed using a selection criterion (the BIC). The selected model provides an estimate of the abundance and a confidence interval around it. The estimation is refined using a BMA to incorporate the model choice uncertainty and improve the final estimation.

D. Tools

The analysis is performed using the computer program R and the package BMA (R Core Team, 2015). The code for the analysis is reported in the Appendix.

III. Application and results

The count of factories per list is reported in the data table below.

Count	BGMEA	DIFE	BKMEA	ALLIANCE	ACCORD
1569	1	0	0	0	0
620	0	1	0	0	0
1314	1	1	0	0	0
1170	0	0	1	0	0
74	1	0	1	0	0
490	0	1	1	0	0
87	1	1	1	0	0
98	0	0	0	1	0
47	1	0	0	1	0
6	0	1	0	1	0
110	1	1	0	1	0
4	0	0	1	1	0
2	1	0	1	1	0
14	0	1	1	1	0
5	1	1	1	1	0
227	0	0	0	0	1
187	1	0	0	0	1
25	0	1	0	0	1
494	1	1	0	0	1
58	0	0	1	0	1
25	1	0	1	0	1

97	0	1	1	0	1
95	1	1	1	0	1
34	0	0	0	1	1
53	1	0	0	1	1
6	0	1	0	1	1
222	1	1	0	1	1
2	0	0	1	1	1
7	1	0	1	1	1
15	0	1	1	1	1
22	1	1	1	1	1

For example, 1569 factories have been registered in list BGMEA alone, and 87 have been observed in the three lists BGMEA, DIFE and BKMEA (and not in the two others). The statistical analysis aims at estimating the missing row

Count	BGMEA	DIFE	BKMEA	ALLIANCE	ACCORD
???	0	0	0	0	0

The following table reports some descriptive statistics.

Seen in	**Counts**
Total	7169
BGMEA	4313
DIFE	3622
BKMEA	2167
ALLIANCE	647
ACCORD	1569
exactly 1 list	3684
exactly 2 lists	2239
exactly 3 lists	890
exactly 4 lists	344
exactly 5 lists	22

As explained in Section II, the model selection is performed using the R package BMA functions, applying a selection based on the BIC criterion. For this, a set of prior probabilities has to be given reflecting our prior information about the fact that a given parameter (e.g. interaction) is in the model or not. This is a difficult task due to the lack of prior information, the number of parameters, and the potential influence such specification may have on the final result. The results below are presented for a choice made to force the inclusion of the list effects, and a dependence term between lists BGMEA and BKMEA. Higher order interactions are considered less likely. Further details can be found in the code in the Appendix.

The final model selection (with the highest posterior probability) is given in the Appendix. For this model, the abundance estimation is

Model parameter estimate "a"	9.635
Unobserved factories "m00000 = exp(a)"	15'286
Abundance "N = m00000 + 7169"	22'465

Using the standard deviation of the estimate for this model (se = 0.123), the confidence interval at 95% for the abundance is

	Lower bound	Upper bound
Unobserved factories "m00000"	12'020	19'465
Abundance "N"	19'189	26'634

As discussed in the methodology, the uncertainty due to model choice may be quite high. Incorporating this uncertainty using BMA gives another set of (approximate) estimates as follow

	Posterior expectation	Lower bound	Upper bound
Unobserved factories "m00000"	17'050	11'557	27'829
Abundance "N"	24'219	18'726	34'999

IV. Discussion

The total number of observed factories in the survey is 7'169, out of 5 lists.

Using an MSE approach, an estimate of the unseen number of factories can be obtained. This estimate being around 15'300, or equivalently that the total number of factories is around 22'500. Similar estimates are obtained using various methods. Therefore, this method suggests that the survey would have counted about one third of the factories (between one half and one fourth when incorporating the uncertainty).

In addition to these estimates, several other elements could be extracted from the statistical analysis, like the quantification of the dependence between the lists. For example, BGMEA and DIFE have a positive dependence (more common counts are observed than expected if they were independent). This analysis is not pursued here as it is not the aim of this study.

V. References

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayeisan Model Averaging: A Tutorial. *Statistical Science*, 14(4):382-417.

HRDAG. (2015, 08 01). Retrieved from https://hrdag.org/2013/03/20/mse-stratification-estimation/

Lum, K., Price, M. E., & Banks, D. (2013, November 18). Applications of Multiple Systems Estimation in Human Rights Research. *The American Statistician*, 67(4):191-200.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

VI. Appendix

A. The R code

The code applies to the R object tab.grid containing the survey results.

```
> tab.grid
   count bgmea dife bkmea alliance accord
1      0     0    0     0        0      0
2   1569     1    0     0        0      0
3    620     0    1     0        0      0
4   1314     1    1     0        0      0
5   1170     0    0     1        0      0
6     74     1    0     1        0      0
7    490     0    1     1        0      0
8     87     1    1     1        0      0
9     98     0    0     0        1      0
10    47     1    0     0        1      0
11     6     0    1     0        1      0
12   110     1    1     0        1      0
13     4     0    0     1        1      0
14     2     1    0     1        1      0
15    14     0    1     1        1      0
16     5     1    1     1        1      0
17   227     0    0     0        0      1
18   187     1    0     0        0      1
19    25     0    1     0        0      1
20   494     1    1     0        0      1
21    58     0    0     1        0      1
22    25     1    0     1        0      1
23    97     0    1     1        0      1
24    95     1    1     1        0      1
25    34     0    0     0        1      1
26    53     1    0     0        1      1
27     6     0    1     0        1      1
28   222     1    1     0        1      1
29     2     0    0     1        1      1
30     7     1    0     1        1      1
31    15     0    1     1        1      1
32    22     1    1     1        1      1
```

```
require(BMA) # package containing bic.glm

## ############################
## Applying the selection and BMA
## Prior probabilities are set such that main effect are kept
## and 2-way interaction between BGMEA and BKMEA is likely
## Higher order interaction terms are less likely
model.bic <- bic.glm(f=count~bgmea*dife*bkmea*alliance*accord,
data=tab.grid[-1,],
glm.family=poisson(link="log"), strict=TRUE, maxCol=31,
prior.param=c(rep(1,6),0.5, 0.75, rep(0.5,8),rep(0.25,14),0.1))
summary(model.bic)

## ############################
## Computation using the selected model (with the highest posterior
probability
a.hat <- model.bic$mle[1,1] # estimate of the intercept (a)
m00000.hat <- exp(a.hat) # estimate of the missing count (m00000)
N.hat <- m00000.hat + sum(tab.grid[,1]) # estimate of the abundance (N)
## Confidence interval at 95% around the abundance N
CI.inf <- sum(tab.grid[,1]) + exp(a.hat + model.bic$se[1]*qnorm(0.025))
CI.sup <- sum(tab.grid[,1]) + exp(a.hat + model.bic$se[1]*qnorm(0.975))

## ############################
## Computation using Bayesian Model Averaging (all selected models)
m00000.exp <- sum(exp(model.bic$mle[,1])*model.bic$postprob) # posterior
expectation of the missing count (m00000)
```

```
N.exp <- m00000.exp + sum(tab.grid[,1]) # posterior expectation of the
abundance (N)
N.Bayes <- sum(exp(model.bic$mle[,1])^(-
1)*model.bic$postprob)/sum(exp(model.bic$mle[,1])^(-2)*model.bic$postprob)
# lower risk Bayes estimate E(N^(-1))/E(N^(-2))

## Estimation of the posterior standard deviation of a.hat
## Var(a.hat) = E(Var(a.hat|Model)) + Var(E(a.hat|Model))
Var <- sum(model.bic$mle[,1]^2*model.bic$postprob)-
sum(model.bic$mle[,1]*model.bic$postprob)^2+sum(model.bic$se[,1]^2*model.bi
c$postprob)
SD <- sqrt(Var)

## Approximate interval for N incorporating the model uncertainty
a.exp <- sum(model.bic$mle[,1]*model.bic$postprob)
Int.inf <- sum(tab.grid[,1]) + exp(a.exp + qnorm(0.025)*SD)
Int.sup <- sum(tab.grid[,1]) + exp(a.exp + qnorm(0.975)*SD)
```

B. The result of the selection

The selection provides 116 models. The best five are (result of `summary(model.bic)`)

```
  116  models were selected
 Best  5  models (cumulative posterior probability =  0.2502 ):
```

	p!=0	EV	SD	model 1	model 2	model 3	model 4	model 5
Intercept	100	9.793841	0.22432	9.6347	9.6763	10.0317	10.0143	9.6185
bgmea	100.0	-2.431625	0.22404	-2.2706	-2.3151	-2.6717	-2.6512	-2.2560
dife	100.0	-3.375764	0.21751	-3.2227	-3.2657	-3.6028	-3.5827	-3.2146
bkmea	100.0	-2.737098	0.22553	-2.5769	-2.6170	-2.9774	-2.9627	-2.5604
alliance	100.0	-5.201847	0.22567	-5.0790	-5.1032	-5.4069	-5.4016	-5.0016
accord	100.0	-4.344086	0.22389	-4.1973	-4.2248	-4.5774	-4.5738	-4.1558
bgmea:dife	100.0	3.193256	0.21452	3.0404	3.0898	3.4189	3.3911	3.0407
bgmea:bkmea	65.7	-0.303150	0.26607	-0.4986	-0.4611	.	.	-0.5052
dife:bkmea	100.0	2.544163	0.21921	2.3974	2.4370	2.7685	2.7538	2.3793
bgmea:alliance	100.0	1.655033	0.23717	1.5534	1.6139	1.8497	1.7956	1.3642
dife:alliance	100.0	1.071825	0.18145	1.0350	0.9808	1.0799	1.1508	1.1013
bkmea:alliance	18.2	-0.030220	0.14480	.	.	.	.	.
bgmea:accord	100.0	2.167704	0.22349	2.0203	2.0498	2.3987	2.3916	1.9561
dife:accord	100.0	1.218070	0.09518	1.1976	1.1880	1.2168	1.2305	1.2073
bkmea:accord	100.0	1.380500	0.23430	1.2324	1.2514	1.6426	1.6483	1.2053
alliance:accord	100.0	3.129028	0.33695	3.1682	3.0214	3.3163	3.5013	2.6336
bgmea:dife:bkmea	100.0	-2.279005	0.25965	-2.1532	-2.1885	-2.5335	-2.5331	-2.1221
bgmea:dife:alliance	4.5	-0.012754	0.09330	.	.	.	.	.
bgmea:bkmea:alliance	5.7	-0.022348	0.11482	.	.	.	.	.
dife:bkmea:alliance	5.4	0.027042	0.14449	.	.	.	.	.
bgmea:dife:accord	5.4	-0.013084	0.06761	.	.	.	.	.
bgmea:bkmea:accord	43.1	-0.226042	0.28969	.	.	-0.5480	-0.5843	.
bgmea:alliance:accord	87.4	-0.773429	0.40725	-0.8718	-0.7027	-0.9196	-1.1300	.
dife:alliance:accord	56.6	-0.381883	0.37608	-0.6064	.	.	-0.7254	.
bkmea:alliance:accord	61.4	-0.401476	0.37880	-0.5746	-0.6738	-0.6904	-0.5531	-0.6115
bgmea:dife:bkmea:alliance	15.8	-0.100582	0.27499	.	.	.	.	.
bgmea:dife:bkmea:accord	4.6	-0.013206	0.07621	.	.	.	.	.
bgmea:dife:alliance:accord	41.5	-0.294104	0.38229	.	-0.6113	-0.7286	.	-0.9002
bgmea:bkmea:alliance:accord	6.6	-0.007955	0.19944	.	.	.	.	.
dife:bkmea:alliance:accord	9.0	-0.046836	0.17529	.	.	.	.	.
bgmea:dife:bkmea:alliance:accord	1.8	-0.011903	0.09466	.	.	.	.	.
nVar				18	18	18	18	17
BIC				-23.6341	-23.6097	-25.8050	-25.3239	-20.3759
post prob				0.058	0.057	0.057	0.045	0.034
