UNIVERSITY OF SOUTHAMPTON

LABOUR FORCE ESTIMATION FOR SMALL AREAS IN VENEZUELA

By

Félix Leonardo Seijas-Rodríguez

Doctor of Philosophy

Department of Social Statistics

Faculty of Social Sciences

July 2002

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL SCIENCES
SOCIAL STATISTICS

Doctor of Philosophy

LABOUR FORCE ESTIMATION FOR SMALL AREAS IN VENEZUELA

By Félix Leonardo Seijas-Rodríguez

This thesis focuses on the problem of producing reliable estimates of Employment, Unemployment and Activity rates by Sex-age groups for Venezuelan States using the Population Census as auxiliary information. This is a common situation in the Latin-America region. The SPREE approach to Small Area Estimation is suited to dealing with this sort of problem. Although the use of SPREE methods in the SAE context has been treated in the literature, its use for estimation of product multinomial variables as well as a general methodology for variance estimation was largely unexplored.

There are some potential barriers to the convenient application of SPREE methods. To start, we note that SPREE involves application of the Iterative Proportional Fitting (IPF) algorithm which often requires the development of "domestic" computational algorithms. Besides this, the general computation of variance estimates for SPREE is not obvious. To address these issues, we established a link between SPREE methods, Log-linear models and Logistic models allowing the integration of complex sampling designs via the Pseudo-Likelihood approach to estimation. The main attraction of such a link is that it offers the possibility of implementing SPREE from a GLM perspective. We then show the equivalence of the Log-linear and Logistic versions of SPREE to the application of the well known "Exposure" technique from regression theory. This equivalence allows us to easily implement SPREE, computing parameter and variance estimates as well as goodness of fit measures and related diagnostic, using standard commercial statistical software. Overall, the approach to SPREE presented in this thesis makes this technique more flexible and accessible for practical application.

The "exposure" approach to SPREE was used in an empirical analysis of the Venezuelan labour force, including a simulation study to examine the properties of the estimators considered in this thesis. Superiority of the SPREE method over design-based estimators when the former is based on a good reference table was evident. However, conventional logistic model-based estimators can be regarded as favourable alternatives to SPREE-based estimators in situations when there is a reasonable doubt about the quality of the available reference information.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEGMENT

I am deeply grateful to my supervisor Prof. Ray Chambers for the quality of his supervision and support throughout my study-period. Thanks are also due to all the member of the staff of Social Statistics Department for their unconditional cooperation and assistance, particularly to Anne Owens and Sandy Mackinnon.

I would also like to express my gratitude to all my colleagues and friends from the Department of Social Statistics. Among them, Tiziana Leone, Priscilla Akwara, Juliet MacEachran, Zoe Sheppard, Nikolaos Tzavidis, Gabriele Beissel, Fatima Salgueiro, Chiweni Chimbwete, Yves Berger, Kosta Politis, Faiza Tabussum, Robert Clark and Yukka Jokinen.

I am indebted to many people in Venezuela for their support and assistance. I am especially grateful to Emiro Molina and his wife Julie for their encouragement and friendship. Thanks are also due to José Domingo Mujica, Miguel Bolívar, Isbelia Lugo, Luís Montero, Alberto Camardiel, Diana Camejo, Irina De La Rosa, Lissette D'angelo, Mónica Montero, Orángel Rivas and Guillermo Ramírez.

No words can express my gratitude to my family for all their love, support and encouragement, especially my parents Félix and Gladys and my wife Dulce María, thanks you for your love, patience and wisdom. Finally, I dedicate this thesis to the memory of one of the greatest women I have ever met, my Grandmother Mercedes.

This Ph.D. was fee supported by The British Council.

CHAPTER 1

INTRODUCTION

Small Area Estimation (SAE) is concerned with strategies and techniques to produce reliable estimates for small "groups" of units from a large universe. Different dimensions such as time (longitudinal), the characteristics of units (vertical), space (horizontal) or a combination of them commonly define these small "groups". Typically, however, SAE is related to the production of information for small geographic areas, that is, for groups defined by "space" (or the combination of space with any other dimension). This is mainly due to practical issues regarding the dynamics of the process defining information needs. At some point, it becomes more valuable to handle aggregate information for local areas than highly disaggregated information for a larger area like a country.

1.1. FROM EXHAUSTIVE ENUMERATION TO SMALL AREA ESTIMATION

Information needs have changed with the evolution of societies. Ever since organised civilisation started developing on earth, quantitative information has played an important role as an instrument for planning. Information regarding the size of organised human settlements has historically been of concern to their leaders. An early example of this can be found in the clay tables of the antique Babylon Circa

3800 BC, showing population counts to predict the future imperial income from inhabitants' tax payments. Other early censuses were registered in the Chinese, Hebrew, Egyptian, Greek and Roman civilisations, the latter being the first to carry out censuses periodically. Those were mainly undertaken with military enrolment and tax imposition in mind.

Societies have changed through the years becoming larger and consequently more complex. This has had an effect on information needs. However, the need for information at geographic levels defining the organisational structure of societies has remained a basic requirement for planning. In the modern era, democracies need population counts to establish representation in congresses or parliaments, the 1790 US Census being the first one carried out with this purpose. Distributions of national revenues, planning public services as well as economic and demographic matters have increasingly motivated governments to produce information at local levels. The UK, for instance, carried out its first census in 1801 as a simple count of population. Later in the 1821 Census, ages were registered for first time and names, addresses and occupations in the 1841 Census.

Nowadays, censuses are carried out all over the world, differing in periodicity, procedures and contents. Although factors influencing the differences between censuses can be related to a country's specific goals, the factor defining the ultimate structure of any census is the available budget. Censuses are complex operations and the available budget will dictate the sort of technology, human resources, and fieldwork complexity allowed and thus, the periodicity and content. However, censuses can not completely provide either the highly diverse information needed in a country or the periodicity with which this information is required. Even with respectable budgets, developed countries can not fully meet these needs by relying on censuses.

Other important sources of information are administrative registers (AR). This sort of information is naturally generated mainly from legal procedures related to organised societies. Land and businesses transactions, and unemployment and social security registrations are just some examples where official lists or records have to be kept.

Registers with an electoral purpose, for instance, appeared in Europe around the beginning of the 19th century. Like censuses, AR characteristics depend mainly on specific goals, technology and human resources. Usually, registers are compiled by different organisations within a country, with the process of data capture designed to meet organization specific goals. National statistics offices can rarely influence these processes. As a consequence, it is not surprising to find poor linkages between most ARs. They are also highly dependent on institutional policies which, due to the lack of linkage between organisations, make them unstable over time. Although the usefulness of AR is limited by the issues mentioned above, registers have still proved to be valuable sources of information for planning in different situations, at different geographical and demographic levels.

Sample surveys as a scientific method of information collection, originated in the 19th century with the representative method described by Kiaer in 1895. In 1925 the International Statistics Institute officially accepted sampling surveys as a tool for scientific collection of statistical data. However, it was not until the nineteen forties that governmental offices started using them. Sample surveys are able to collect more comprehensive and more complex information than censuses and AR. The low cost and ease of execution of sample surveys quickly made them the main complement to what used to be the principal, if not the only, sources of statistical information at the time, censuses and administrative registers. A number of official survey programs were initiated in many countries and by the seventies there existed little doubt about their effectiveness and usefulness. However, these sample surveys were unable to produce reliable estimates at every geographic and demographic level of interest without becoming as logistically complex as censuses. Consequently, although these surveys were used to provide information at the national level and, in some cases also at large regional levels, information at local levels was still based on censuses and AR.

However, the demands for detailed information at local areas were constantly growing. This need became more acute when it became clear that the complete coverage that censuses aim for is unfeasible, due, for example, to underenumeration. This problem becomes more evident when statistics at local levels are used for

sensitive matters such as allocation of governmental funds. Consequently, researchers were motivated to look for alternative procedures to satisfy these requirements.

The use of models in sampling theory provided a way of carrying out this task. The use of regression methods as a tool for improving estimation at local levels was first presented by Hansen, Hurwitz and Madow (1953) in their seminal text on sampling techniques. Regression methods were subsequently used by Madow (1956) and Woodruff (1966) in a report on the use of television in households and to produce monthly national estimates of retail trade respectively. Ericksen (1974) used regression methods to obtain postcensal estimates of population counts for local areas. Using 1970 census data, he showed how the proposed estimators performed better than traditional demographic procedures. Ericksen's paper significantly motivated the interest of researchers in the subject, and Fay and Herriot (1979), developed improved per capita income estimates at state and local government levels for the US Treasury Department. Their estimator was based on the Empirical Bayes method, combining direct and synthetic estimations via a weighted average. These estimates were used by the Treasury Department to allocate funds to local government units within the different states.

Since Fay and Herriot's work, much theoretical and applied research on small area estimation has been carried out using different statistical techniques, with many national statistics offices investing resources on special programs to develop procedures to meet demands for local area estimates. For instance, Statistics Canada launched an ambitious program in 1983 to develop a small area database. This program was designed to take an integrated perspective covering areas such as the production of "new small areas data sets", the organisation of "small area data in geographically oriented data bases" and the conformation of "geographic, conceptual and methodological frameworks and tools needed to support continuing small area data development, dissemination and analysis" (Brackstone 1987).

Several international conferences on the subject were also held. Statistics Canada hosted an International Symposium on Small Area Statistics in Ottawa in 1985 producing two publications one for invited papers (Platek et al. 1987) and a second

one for contributed papers (Platek and Singh 1986). A similar symposium was held in New Orleans in 1988 organised by the National Center for Health Statistics. An international conference on the topic took place in Warsaw, Poland in 1992. Invited and contributors papers presented in that conference were also published (Kalton, Kordos and Platek, 1993). Recently, the U.S. Census Bureau organised a Conference on Small Area Estimation held in Washington in 1998, the International Association of Surveys Statisticians held the International Satellite Conference on Small Area Estimation in Riga, Latvia, 1999 and the United States Postal Service in conjunction with the American Statistical Association and other organizations held the International Conference on Small Area Estimation and Related Topics, Potomac, Maryland 2001. A variety of international courses and other events have also been held in different countries, mainly in North America and Europe. For a comprehensive list of major events addressing small area estimation issues that have taken place in the last twenty years, see Gosh and Rao (1994) and Rao (1999). Some recent practical applications on Small Area Estimation can be found in Citro et al. (1997), Citro et al. (1998), Cohen (1999), Falorsi (1999), Wang et al. (1999), Knaub(1999), Wang, Fuller and Opsomer (1999) Larsen (2000), Olsen et al. (2000), Judkins and Liu (2000), Citro and Kalton (2000), Cohen (2000), You (2000).

Small area estimators are commonly known as "model estimators" or "indirect estimators". They make use of data from outside the area of interest to produce estimates for that specific area, in contrast to traditional "design estimators" or "direct estimators", which use data from the area of interest only.

Although small area estimators have been developed and their properties studied under both the randomisation and model-based approach to sample survey inference an underlying model is always present, either explicitly or implicitly. Bayesian ideas have also been used in the small area estimation context and they have proved useful particularly when estimator accuracy needs to be assessed. Many of these estimators make use of auxiliary data from censuses and/or AR. In such cases, the use of these estimators is limited by the existence of adequate auxiliary information. In other cases, estimators make use of information from just one source such as a survey, and some smoothing is carried out to decrease the variability in the estimates. In general,

however, the basic idea behind all these procedures is to borrow "strength" in some dimension (space, characteristics of units or/and time) to "aid" estimation for "groups". An overview of these developments can be found below in Section 1.5.

Most of the research carried out so far on SAE has focused of model-based approaches to estimation for local areas. However, such estimators should be used carefully. In particular, their dependency on hypothetical models does not appeal to many analysts. Model validation has to be carried out in some sensible way and estimators that are robust to model misspecification should be sought. Furthermore, stable estimates of the reliability of the small area estimators have to be computable, and when results are published, users have to be informed about the inherent characteristics of this kind of estimations. Some analyses, e.g. comparison between local areas, will be distorted when one area uses information from other areas (Schaible 1992; Kalton 1994). Singh, Gambio and Mantel (1994) argue that `a model estimator should be preferred over a design estimator only if its mean square error is estimable and it is sufficient smaller than the corresponding variance of the design estimator'.

An interesting issue that has not received much research attention so far is the investigation of sampling strategies for improved small area estimates. Key parameters in the sampling design can be modified and their impact on local area estimation measured. This can reveal considerations that should be taken into account when designing or redesigning sampling surveys with SAE in mind, thus reducing the need for model based estimators. As Kalton (1994) points out `Where possible, samples should be designed to produce direct small area estimates of adequate precision, and sample design should be fashioned with this in mind'. Singh, Gambio and Mantel (1994) report significant improvements in Canadian LFS estimates for sub-provinces by reallocating the sample in two components: a first component designed to provide reliable national and provincial level estimates (42,000 households) and a second component designed to provide improved sub-provincial level estimates (17,000 households).

1.2. SMALL AREA ESTIMATION IN LATIN AMERICA[1]

Despite the amount of attention SAE has received during the last twenty-five years, the official statistics offices in Latin America have remained largely unexposed to these ideas. Although several reasons can be advanced to explain this fact, there are three specific issues that we believe have been the main factors.

The Latin American political landscape was largely made up of dictatorship regimes and intermittent unstable democratic governments until relatively recently. This factor explains the heavily centralised governmental structures present in these countries. Planning was completely centralised and little information at local levels was required. Therefore the "demand" for local information, which has been the main factor that has led other countries to move towards SAE, was not present.

Secondly, we have to consider the depressed economies that have historically characterised many countries in Latin America. This fact has implied modest governmental allocation of funds for statistics programs. Although it is not true in general, it is possible to find national statistics offices from some countries struggling to set up a single household survey even for their capital or main cities. Countries that run continuing survey programs generally do it with surprisingly small budgets. Consequently, it is hard to imagine these countries investing money to set up programs like the one in Canada mentioned earlier.

Finally, one of the most important factors explaining the disparity between Latin American countries and North American and European countries in terms of SAE development is the lack of linkage between academic researches and governmental statistics offices. A strengthening of this linkage could motivate changes in both the way statistical offices conceive the production of statistics and the way the academic research area direct its efforts. An integrated approach to the National Statistics System in these countries, could lead to a long overdue interaction between the governmental statistics offices and the academic area.

[1] Although the term "Latin America" formally refers to American countries that were colonies of Spain, Portugal and France, in this work we will use this expression to denote Spanish-speaking countries only.

Nevertheless, the need for statistical information at local level in Latin America is now on the increase. Many national and international aid programs with a regional focus have started to be implemented. The "maturity" of some democratic systems is now reflected in the continuing decentralisation process that these countries have been undergoing. To learn about the effect that these phenomena have had on SAE demand as well as to establish in general terms the "state of the art" concerning the production of information for local areas in Latin America, we contacted fifteen out of eighteen of the national statistics offices in the region. We got feedback from Argentina, Chile, Ecuador, Peru, Mexico and Venezuela. It is important to point out that these are the countries with the "strongest" statistics offices in Latin America, with INEGI from Mexico being the statistics office in Latin America that has historically had the largest budget. We therefore expect that these six countries give us an upper bound regarding the situation of Latin American statistics offices. For countries from which we did not get first-hand feedback, we have relied on information published in CIENES (1995).

None of the offices contacted produces statistics at local levels from sources other than Censuses and special "ad-hoc" studies. However, all of them report pressure from users to produce information at these levels. They all run permanent businesses and labour force survey programs producing direct estimates up to levels of disaggregation consistent with the sample design. Ecuador, Mexico and Venezuela are the only countries where at least one survey covers both urban and rural areas. They all produce results for "regions" i.e. aggregations of provinces or states. Mexico produces some labour force indicators for states by combining sample from two consecutive years. The sample sizes of the Labour force surveys vary from 9,180 households (Ecuador) up to 158,960 households (Mexico) per year. AR or any other kind of auxiliary information other than censuses is rare in these countries. Educational and formal employment registers are the most common AR but they are – with the possible exception of Mexico- unreliable, usually published at the national level and not kept in any exploitable data base format. No unemployment register is available in any country. When questioned about the indicators that are the most demanded at local levels by users, the common answer was the set of basic rates describing labour force structure by sex and age groups. This is the main reason why

this thesis will focus on estimation of those rates at local levels. In particular, we develop an approach for the specific case of Venezuela.

1.3. THE VENEZUELAN CASE

Although we focus on the Venezuelan case from now on, it should be noted that due to the similarity among Latin American countries this development can be regarded as generally applicable to the problem in Latin America.

As a consequence of a long history of centralised government, official information sources in Venezuela are national level indicators. In the social context, several ARs are available. These are designed to produce figures at the national level in most cases. ARs of acceptable quality are compiled for births, deaths and educational enrolment. Other ARs are rare and their quality is poor for local areas. Thus, the use of statistics techniques based upon auxiliary information other than the Census'90 and its population projections does not seem feasible at least in the short term.

The Labour Force Survey[2] (LFS) is the instrument officially used in Venezuela to estimate Labour Force indicators related to the supply side of the Labour Market. This survey has been carried out by the Statistics and Informatics Central Office of Venezuela (OCEI)[3] since 1967.

The LFS produces six-monthly estimates at the national level based on a 12.000 household sample. In 1977, this sample size was increased to 75.000 households to produce six-monthly estimates for nine regions. These regions consist of a group of neighbouring states with similar characteristics such as weather, geography, and predominant economic activity. In 1994 the LFS sample size was cut down to the original 12.000 households due to budget restrictions.

[2] "Encuesta de Hogares por Muestreo" (EHM)
[3] It became the National Statistic Institute (INE) in the year 2000.

However, as a consequence of political reform, requirements for information at finer geographic levels have arisen in the country since the beginning of the nineties. In 1989 the first democratic election for State Governors and County Mayors was held in Venezuela. This started a process within the country, which has progressively changed the levels at which social and economic planning is carried out in Venezuela. Several economic and social strategies have been designed and executed at state and county levels. Therefore, providing information for each of the 23 states and 353 counties[4] has become increasingly important.

The current LFS sample size only allows the production of reliable estimates at the national level. Nevertheless, it is important to point out that through special agreements reached between governors and the OCEI, the LFS sample size has been increased within seven states and the Metropolitan Area of Caracas (AMC)[5] (see table 1) in order to obtain reliable estimates at those levels (though not with the same level of disaggregation as the national estimates).

1.4. AIMS OF THE STUDY

We base our study on the basic set of labour force indicators identified as essential by Latin American users, i.e. rates describing the labour force structure by sex-age groups for states. Although SAE is usually associated with spatial problems, it can be related to any other dimension of data disaggregation. In our specific case, we shall focus on sub-populations that are defined by both spatial (states) and demographic (sex-age groups) dimensions. In what follows, we refer to such groups as "sub-populations". We do not use the term "domains" to avoid conflict with the definition given by the United Nations and popularised by Kish (1965), which relates "domains" to sub-populations *'about which the enquiry is planned to supply numerical information of known precision'* (U.N., 1950).

[4] This is the total of Venezuelan counties for 1997. There have been some changes since that year.
[5] The AMC concern the Federal District (DF) and some counties from Miranda State.

We are therefore basically concerned with three indicators: Employment Rate ($_eR$), Unemployment Rate ($_uR$) and Non-active Rate ($_nR$). The LFS conceptual definitions of Employee, Unemployed and Non-active agree with the International Labour Organization (ILO) definitions. The LFS operational definitions of these concepts (variables and rules used to classify people into these groups) are adjusted so that the particular socio-economic characteristics of Venezuela reflect the ILO conceptual definitions as closely as possible. Making use of the variables contained in the Summary Code, the LFS operational definitions of Employees, Unemployed and Non-actives are as follows:

Employees: *people over 14 years old who have received or will receive money due to any kind of work carried out during the survey reference period (the week previous to the interview). This category includes self-employed.*

Unemployed: *people over 14 years old who have actively looked for a job and who have not received money due to any kind of work during the survey reference period (the week previous to the interview).*

Non-actives: *people over 14 year old who have not actively looked for a job and who have not received money due to any kind of work during the survey reference period (the week previous to the interview).*

Notice that these categories are mutually exclusive and exhaustive for people over 14 years old. We further define *Actives* as the combination of both groups Employees and Unemployed, this is, the complement of Non-actives category.

We now define the basic LF rates $_eR$, $_uR$ and $_nR$:

$$_eR = \frac{Employees}{Actives} \qquad _uR = \frac{Unemployed}{Actives}$$

$$_nR = \frac{Non-actives}{(Actives)+(Non-actives)}$$

In the same way we can also define an *Activity Rate* ($_AR$) as $_AR = (1 - {}_nR)$.

Table 1.1
Venezuela. Activity Rate and Unemployment Rate by States

State	Activity			Unemployment			n
	Rate	se %	cv %	Rate	se %	cv %	
Venezuela	65.4	0.25	0.38	11.0	0.25	2.23	52,614
Dtto. Federal	68.0	0.66	0.97	7.8	0.57	7.33	5,263
Anzoategui	64.6	1.30	2.02	11.8	0.95	8.03	1,589
Apure	67.6	3.16	4.67	10.5	2.52	23.96	226
Aragua	69.4	1.23	1.77	7.9	0.84	10.72	1,739
Barinas	64.5	2.10	3.26	7.8	1.05	13.56	687
Bolivar	61.4	0.62	1.02	12.3	0.70	5.68	5,699
Carabobo	66.5	0.95	1.42	14.0	1.32	9.42	2,317
Cojedes	63.1	3.60	5.71	13.0	4.61	35.62	246
Falcon	64.4	0.84	1.30	16.6	0.88	5.33	4.754
Guarico	62.7	1.19	1.90	18.8	1.79	9.54	798
Lara	63.0	0.69	1.10	9.0	0.63	6.97	5,153
Merida	64.1	1.32	2.05	5.2	1.04	19.77	1,168
Miranda	65.2	0.89	1.36	7.6	0.65	8.54	3,567
Monagas	64.4	1.56	2.42	18.4	1.79	9.69	1,142
Nva. Esparta	59.5	2.68	4.51	4.1	1.21	29.72	453
Portuguesa	61.4	0.82	1.34	14.1	0.95	6.74	3,411
Sucre	58.9	1.54	2.61	8.4	1.19	14.17	1,429
Tachira	64.6	1.03	1.60	7.7	1.05	13.55	1,677
Trujillo	63.7	1.43	2.25	13.2	2.65	20.08	894
Yaracuy	66.4	1.57	2.36	18.6	2.27	12.20	593
Zulia	68.5	0.58	0.85	14.7	0.55	3.72	8,531
Amazonas	68.0	2.32	3.41	5.4	1.47	27.05	483
Amacuro	67.1	1.48	2.21	5.6	1.35	24.08	795

Surce: Labour Force Survey, 1998 , OCEI

Estimation of these rates disaggregation by sex and age for each of the 23 Venezuelan states will be our ultimate target. As ages are usually grouped into four groups (see table 2), we will have a total of 184 sub-populations for which estimates of these rates are required (736 estimates). These can be calculated by using the ratio estimators commonly used by the LFS (section 2.2.4). However, as the LFS was designed to produce reliable estimates at national level, the sample size within states is too small to provide estimates with adequate precision for sex-age groups within states. As an example, Table 1.1 shows these estimates along with sample sizes and precision indicators for the 23 Venezuelan states. In the same way, Table 1.2 shows examples of estimates for sub-populations defined by large sample size, medium sample size and small sample size states.

Table 1.2
Venezuela. Activity Rate and Unemployment Rate by Sex and Age Groups

Sex	Age	Activity			Unemployment			n
		Rate	se %	cv %	Rate	se %	cv %	
		Dtto.Federal						
Male	15 - 24	60.0	2.11	3.5	15.0	1.89	12.6	676
	25 - 34	96.0	0.90	0.9	5.9	1.02	17.3	601
	35 - 44	97.5	0.72	0.7	3.4	0.82	23.9	504
	45 - +	76.5	1.66	2.2	4.5	0.89	19.7	691
Female	15 - 24	43.3	2.11	4.9	21.9	2.78	12.7	696
	25 - 34	71.6	1.98	2.8	6.1	1.14	18.7	600
	35 - 44	73.0	1.91	2.6	5.8	1.20	20.8	575
	45 - +	41.9	1.65	3.9	5.5	1.24	22.8	920
		Anzoategul						
Male	15 - 24	63.7	3.34	5.2	20.6	2.88	14.0	252
	25 - 34	96.3	1.28	1.3	15.9	2.51	15.8	199
	35 - 44	95.0	2.39	2.5	4.5	1.82	40.5	144
	45 - +	80.0	2.85	3.6	7.7	2.06	26.9	190
Female	15 - 24	34.6	3.63	10.5	21.5	6.18	28.8	223
	25 - 34	56.7	4.42	7.8	13.1	3.23	24.7	200
	35 - 44	66.2	4.05	6.1	4.8	2.10	43.9	144
	45 - +	40.1	2.95	7.4	3.4	1.85	55.0	237
		Apure						
Male	15 - 24	76.2	8.32	10.9	17.0	7.78	45.8	39
	25 - 34	95.5	3.28	3.4	6.4	5.86	91.8	29
	35 - 44	95.9	4.33	4.5	20.5	12.71	62.0	18
	45 - +	74.1	8.43	11.4	3.6	3.34	92.5	29
Female	15 - 24	42.6	8.84	20.7	25.4	10.49	41.2	35
	25 - 34	68.3	10.24	15.0	4.7	4.70	100.2	27
	35 - 44	82.0	10.32	12.6	0.0	0.00	*	20
	45 - +	27.8	14.70	52.9	0.0	0.00	*	29

Surce: Labour Force Survey, 1998 , OCEI

It is important to point out that the Unemployment Rate is regarded as the most critical from the political point of view. Public opinion and analysis is extremely sensitive to the size of this indicator. Therefore, estimates of high precision are required before these indicators can be released. For instance, a Coefficient of Variation (CV) below 3% is required for the national Unemployment Rate estimate. For states, a CV of 6% is usually required.

As can be seen from Table 1.1, the LFS estimation technique is not able to produce reliable estimates even for states as population groups. Fifteen states in this table have $_{u}R$ estimates with CV estimates larger than 9%. From these fifteen states, six exceed 20% and one (Cojedes) is above 30%. CV estimates are calculated as the ratio of the

estimated standard error (se) to the estimated rate; for instance, the estimated CV for the ${}_uR$ estimate is designed as:

$$\hat{CV}\left({}_u\hat{R}\right) = \frac{s\hat{e}\left({}_u\hat{R}\right)}{{}_u\hat{R}}$$

From Table 1.2 it is clear that estimates for sub-groups related to states with medium sample sizes and small sample sizes are far from adequate in terms of precision. Even for states with a larger sample size, the CV values are not encouraging.

It is worth noting that for some subgroups (e.g. the last two shown in the Table 1.2), the direct estimate is zero because no sample cases were observed for that category. This is more likely due to the small sample size in these sub-populations rather than to the non-existence of people in these categories.

Achieving high levels of precision for the estimates we are concerned with in this study is a difficult task. We aim to produce improved estimates relative to the precision that design based estimators can currently offer for those sub-groups. We aim to do this by using simple procedures that can be accepted and applied with relative ease by OCEI.

We approach the problem from an estimation point of view. As the main source (and the only one in many cases) of auxiliary information in Latin America countries is the census, any techniques employed in this study will be based only on the availability of population censuses and their population projections as auxiliary information.

As mentioned in section 1.1, the basic idea behind SAE techniques is to "borrow strength" in some dimension (space, characteristics of units or/and time) to "aid" estimation for "small groups". This is achieved by using models either explicitly or implicitly. If we identify the structure of a model that adequately explains the distribution of the counts of interest, we can then estimate its parameters to obtain model estimates of these counts. These model estimates should be more precise than

direct estimations due to the fact that they are depend on the estimation of fewer parameters, thus, each of those parameters is estimated using a larger effective sample size. However, models are approximations to reality so they can lead to bias in inference. Consequently, the model in use should sufficiently explain reality for the reduction in variance due to the estimation of fewer parameters to offset the reduction in accuracy due to bias.

The rates we want to estimate are computed from the three mutually exclusive and exhaustive categories for people over 14 years old specified above. The literature about SAE for labour force characteristics is mainly concerned with estimation of the total or the rate of the unemployed population (e.g. Gonzalez and Hoza 1978, Cassel et. al. 1987, Cronkhite 1987, Feeney 1987, Roberts et. al. 1987, Falorsi et. al. 1994, Harter 2000, You et al. 2000). This literature, with few exceptions, assumes the availability of auxiliary information highly correlated with unemployment figures such as unemployment administrative registers. In this study we deal with the task of obtaining simultaneous estimates for the four rates that fully define the basic structure of the labour force, which are based on the three labour force counts for each sub-group. The aggregate of these counts at state levels as well as at the national sex-age groups have to agree with the LFS direct estimates. Finally, these estimates have to be calculated using only censuses as auxiliary information. These conditions constrain the spectrum of techniques that can be applied and the flexibility with which they can be handled. Multivariate techniques that make use of auxiliary information are not suitable in this case.

The method of Structure Preserving Estimation (SPREE) for categorical variables (Purcell and Kish 1980) offers a possible answer to this situation. In this document, we concentrate our attention on the general definition of that method, the development of some variants and their application to our specific situation. Our intention is to explore the SPREE method as a potential method for addressing the problem described above, generalising it to make it more flexible in its application and practical implementation.

We start by describing the Venezuelan Census and Labour Force Survey (LFS) programs and data, focusing in particular on the characteristics of those programs and their resulting data that might have implication for the procedures to be considered in this study. In particular we emphasise the theoretical description of the LFS parameter and variance estimators since there is no document currently available with such a detailed description.

We then give a theoretical definition of the SPREE method and specify some natural variants, describing how they can be used in our specific case. An issue that we pay particular attention to is the practical complexity associated with the implementation of these methods. The SPREE method requires the application of iterative procedures to compute estimates (see Chapter 3). In our case, it also involves the use of complex design-based estimation techniques. These procedures require the development of computational algorithms that are not found in the standard statistical software. Additionally, the computation of variance estimates for the SPREE method is essentially an unexplored area, reflecting the fact that SPREE has not been investigated to any great extent in the SAE literature.

A primary result developed in this thesis is the formal specification of the SPREE method as a particular application of Generalized Linear Model (GLM) theory (McCullagh and Nelder 1983), specifically constrained Log-linear and Logistic models. This link is expanded further in this thesis to allow for complex samples. We then propose a new procedure to compute SPREE estimates that makes use of standard statistical tools found in commonly used statistical software. This approach allows the computation not only of the traditional SPREE estimates but also of all the variants proposed in this document taking into account the sample design. An obvious consequence is that the SPREE method then has all the advantages that flow from working under the GLM framework, so that the estimation of variances is possible as well as the computation of different goodness of fit measures and measures relating to the identification of outlying cells and influential points. A critical evaluation of the SPREE estimation procedure is then possible. Equally important is the fact that the practical implementation of this procedure is then a simple and straightforward application of commonly used statistical software.

Following on from this theoretical development, we use the Venezuelan 1990 Census to explore different model structures relevant to our specific LFS problem. In particular, we design and carry out a simulation study replicating the LFS sample design in order to examine the properties of a number of the estimators proposed for this problem. In doing so we study the gains that SPREE methods can offer over the traditional LFS direct estimates as well as compare the impact of applying SPREE methods given census data from different periods. These data define the reference information used by SPREE, and an investigation of the sensitivity of these methods to the quality of this information is of some interest.

The final aim of this work is to extend the SPREE method to incorporate time as an extra dimension within the estimation process. This is particularly useful when recent reference information is not available but previous runs of the survey are available. Although data limitations preclude any empirical investigation of the behaviour of this extension, theoretical considerations indicates that this "borrowing of strength over time" should substantially improve the performance of SPREE within the Venezuelan context. Further research using this idea looks promising.

1.5. OVERVIEW OF SYNTHETIC ESTIMATORS AND OTHER SMALL AREA ESTIMATION TECHNIQUES

The SPREE methods belong to the class of synthetic estimators. In an attempt to put the SPREE method in context, we therefore describe in this section the ideas behind this class of estimators. Before doing this, however, we feel that a brief discussion of survey design issues concerning SAE as well as an overview of the demographic methods that pioneered the work on SAE might be useful. Also, since the ideas that are developed in this thesis can be expanded to cover another class of SAE, i.e. those based on the inclusion of specific-area level random effects, we briefly discuss this approach at the end of the section.

Relevant references are given throughout this section. Comprehensive overviews of SAE can be found in Gosh and Rao (1994), Marker (1999), Rao(1999), Pfeffermann (1999) and Rao (2000).

1.5.1. Survey Design Issues

Apart from institutional related issues such as data development, infrastructure, policy and management (for details on these issues see Brackstone 1987), the SAE problem can be approached from two main technical points of views, survey design and estimation techniques. Although surveys specifically designed to produce reliable direct estimates for small areas are in general infeasible, different issues regarding survey design can be considered in order to minimize the need for indirect estimators. However, since the estimation approach to SAE has received the most attention in the literature, research addressing survey design issues in the context of SAE is scarce. The two main references addressing these issues are Singh et al. (1994) and Marker (1999).

Singh et al. (1994) describes different ways in which sampling designs can be adapted to increase the reliability of direct estimates for small areas without a significant impact on the estimates for larger areas. They stress the need for an overall strategy at planning, sampling design and estimation stages.

At the planning stage, they point out the importance of anticipating small areas for which estimates might be required. Such anticipation allows survey designers to consider different strategies whose feasibility can be studied in term of budget and operational capabilities.

Once the data needs have been defined, they discuss how sampling designers should ponder those requirements so that the sampling design reflects a compromise between the requirements for small areas and the requirements for larger areas. They identify two ways of working out that compromise, based on sample allocation and clustering. They argue that disproportionate allocation of the sample in favour of small areas can

have an appreciable impact on the precision of small areas estimates without significantly affecting the reliability of estimates for larger areas. Regarding clustering, they suggest that attempts to reduce the level of clustering in the sampling design should be made in order to increase the chances of having sample in as many small areas as possible.

Finally they acknowledge that, no matter what the anticipation at the planning stage and the compromise reflected in the sampling design, there will always be small area information requirements for which direct estimates will not be satisfactory. It is in such cases that the work on small area estimation plays an important role.

Marker (1999) also argues that one should concentrate efforts at the survey design stage in order to increase the possibilities of producing direct estimates for as many domains as possible. He suggests stratification and oversampling as strategies worth considering. For small areas for which reliable direct estimates are not attainable after stratification and oversampling have been considered, he suggests other strategies such as the use of dual-frame estimation to "combine the national survey with supplements in specific areas to produce direct estimates". Finally Marker also acknowledges the importance of special indirect estimation techniques to deal with small areas for which suitable direct estimates are not possible.

1.5.2. Estimation Techniques

The first techniques developed to deal with estimation for small areas have their roots in demographic projection and synthetic methods. They all are based on implicit models that are assumed to be valid, producing indirect estimators with low variability thanks to the "borrowing of strength" across areas. That is, these methods assume that all the areas of interest behave similarly with respect to the variable of interest and consequently one can "borrow strength" for any one small area by capitalising on the similar behaviour of many small areas. This similar behaviour is usually represented in terms of a common model for the distribution of the variable of interest. By definition, these methods do not take into account area specific

variability. Therefore, we can find situations in which the validity of the assumed model fails leading to biased estimators. This fact has motivated researchers to develop techniques based on models that include area specific effects; although these methods are commonly referred in the SAE literature as model-based methods, we shall refer to them in this chapter as "*specific area-level random effect model estimators*".

1.5.2.a. Demographic Methods

Demographic methods that use "symptomatic" data to produce demographic projections for inter-censal years are commonly used in national statistics offices and international organizations. Sieguel et. al. (1954), for instance, gives a comprehensive overview of methods used in the 1940s and early 1950s for making estimates of population below the State level. Some of those methods are still being used with minor adjustments. Development has focussed on the improvement of mechanisms and techniques to obtain the symptomatic data used in those demographic methods. The principal demographic methods found in the literature are the Vital Rates, Arithmetic, Geometric, and Component methods. We now briefly describe the Vital Rates and Components methods.

The ***Vital Rates (VR)*** method was first described by Bogue (1950). It uses Administrative Registers (AR) of births and deaths for the period t for small areas, say b_t and d_t and for larger areas containing the small areas, say B_t, D_t as well as AR or any other reliable estimate for the larger area population counts $\hat{P}_t$. We also assume that corresponding figures for the last population census, say b_0, d_0, B_0, D_0 and P_0, are available. If we define $r_{bt} = b_t / p_t$ and $r_{dt} = d_t / p_t$ we can write the population count of interest for time t for the target small area as,

$$p_t = \frac{1}{2} \cdot \left(\frac{b_t}{r_{bt}} + \frac{d_t}{r_{dt}} \right)$$

The method consists of estimating the unknown rates r_{bt} and r_{dt} assuming the large area ratios $\hat{R}_{bt}/R_{b0}$ and $\hat{R}_{dt}/R_{d0}$, where $\hat{R}_{bt} = B_t/\hat{P}_t$, $R_{b0} = B_0/P_0$, $\hat{R}_{dt} = D_t/\hat{P}_t$ and $R_{d0} = D_0/P_0$, are the same as the small area ratios r_{bt}/r_{b0} and r_{dt}/r_{d0}, so that,

$$\hat{r}_{bt} = \frac{\hat{R}_{bt}}{R_{b0}} r_{b0} \quad \text{and} \quad \hat{r}_{dt} = \frac{\hat{R}_{dt}}{R_{d0}} r_{d0}$$

The VR estimate for the population count of interest for time t for the target small area is then,

$$\hat{p}_t = \frac{1}{2} \cdot \left(\frac{b_t}{\hat{r}_{bt}} + \frac{d_t}{\hat{r}_{dt}} \right)$$

The ***Component*** method uses AR of immigration, emigration and net interstate migration to compute net migration. It basically "updates" the census population count for a small area by adding to it the net migration as well as the difference in number of births and number of deaths since the census year. That is,

$$p_t = p_0 + b_{ot} - d_{ot} + m_{ot}$$

where b_{ot}, d_{ot} and m_{ot} are respectively the number of births, deaths and net migration since the census year.

1.5.2.b. Synthetic Estimators

Synthetic estimators use direct estimates for larger areas to produce indirect estimates for small areas. The implicit model of synthetic estimators assumes that the characteristics of a large area are similar to the local characteristics of its smaller areas. Several estimators that fall into this category have been developed in the last 45 years.

The earliest formal publication recording the use of synthetic estimators in the context of SAE is due to the U.S. National Center for Health Statistics (NCHS) (1968). They used the National Health Interview Survey (NHIS) to obtain direct estimates of National rates of disability $\hat{R}_{\cdot j} = \hat{N}_{\cdot j}^{dis} / \hat{N}_{\cdot j}$ for $J = 78$ population subgroups defined in terms of socio-economic characteristics. Then, they use the 1960 census population counts for the same population subgroups N_{aj} with $N_{a\cdot} = \sum_{j=1}^{J} N_{aj}$ to compute synthetic State estimates of disability $R_a = Y_a / N_{a\cdot}$, where $Y_a = \sum_{k=1}^{N_{a\cdot}} \sum_{j=1}^{J} y_{ajk}$, using the following expression,

$$\tilde{R}_a^{syn} = \sum_{j=1}^{J} \frac{N_{aj}}{N_{a\cdot}} \hat{R}_{\cdot j} \tag{1.1}$$

This estimator has been used in different applications. The assumption here is that $\hat{R}_{\cdot j} = \hat{R}_{aj}$. Note that an equivalent expression for totals is obtained eliminating $N_{a\cdot}$ from (1.1).

The variance of (1.1) is small because that estimator depends on reliable direct estimates for national subgroups. On the other hand, the bias of this estimator might be important if the assumption $\hat{R}_{\cdot j} = \hat{R}_{aj}$ does not hold. However, this kind of estimator is simple to implement and can perform better than direct estimators when sample sizes for small areas are small, as Gonzalez et al. (1996) reports. Other examples of the use of estimators like (1.1) can be found in different US. Bureau of the Census methodological reports as well as in Gonzalez and Hoza (1978) and Haskey (1991).

Holt et al. (1979) formulate explicit analysis of variance models for different population structures implicitly assumed in different synthetic estimators, so that those assumptions can be tested using the available data. They obtain Best Linear Unbiased (BLU) estimators for small area totals and derive the bias under alternative

models. For instance, an appropriate model for the synthetic estimator (1.1) assuming Simple Random Sampling with sample size $n = \sum_{a=1}^{A} \sum_{j=1}^{J} n_{aj}$ is,

$$y_{ajk} = \beta_j + \varepsilon_{ajk} \tag{1.2}$$

The BLU estimator of β_j is $\hat{R}_{\cdot j}$ and the BLU estimator of Y_a is given by,

$$\tilde{Y}_a^{BLU} = \sum_{j=1}^{J} n_{aj}(\hat{R}_{aj} - \hat{R}_{\cdot j}) + \sum_{j=1}^{J} N_{aj}\hat{R}_{\cdot j} \tag{1.3}$$

Estimator like (1.1) can be seen as special cases of the following general synthetic estimator or ***Individual-level Synthetic Regression Estimator*** (Skinner 1991),

$$\tilde{\bar{Y}}_a^{synr} = \sum_{j=1}^{J} \bar{X}_{aj}\tilde{\beta}_j \tag{1.4}$$

Here we assume we have individual level information about J auxiliary variables for the entire population, so that we can compute their population means $\bar{X}_{aj}$. Applying traditional regression methods to the sample data we estimate the coefficients $\tilde{\beta}_j$. The implicit assumption behind (1.4) is that the following model is a reasonable one for every individual $i=1,\ldots,N$ in the population,

$$y_{ai} = \left(\sum_{j=1}^{J} x_{aji}\beta_j \right) + \varepsilon_{ai} \tag{1.5}$$

The random error ε_{ai} is assumed to have mean zero for all individuals belonging to area a and uncorrelated with the X variables. Skinner (1991) points out two possible departures from this model, misspecification of the systematic component and misspecifications of the random (error) term. The former can be dealt with applying traditional diagnostic techniques. The latter is a more complex issue and it is the key focus for SAE. The error structure in the implicit model might not have mean zero for

area a but u_a, that is, $\varepsilon_{ai} = u_a + v_{ai}$ with v_{ai} having mean zero across individuals in the same area. In this case, (1.4) will have a bias equal to u_a but a lower variance than the direct estimates.

Assuming a direct estimator $\hat{\bar{Y}}_a$ can be considered unbiased and its covariance with $\tilde{\bar{Y}}_a^{synr}$ is approximately zero, the Mean Squared Error (MSE) of (1.4) is,

$$\begin{aligned} MSE(\tilde{\bar{Y}}_a^{synr}) &= E(\tilde{\bar{Y}}_a^{synr} - \bar{Y}_a)^2 \\ &= E\left[(\tilde{\bar{Y}}_a^{synr} - \hat{\bar{Y}}_a) + (\hat{\bar{Y}}_a - \bar{Y}_a) \right]^2 \\ &= E(\tilde{\bar{Y}}_a^{synr} - \hat{\bar{Y}}_a)^2 - E(\hat{\bar{Y}}_a - \bar{Y}_a)^2 \end{aligned}$$

Therefore, an estimator of $MSE(\tilde{\bar{Y}}_a^{synr})$ is given by,

$$\hat{MSE}\left(\tilde{\bar{Y}}_a^{synr}\right) = \left(\tilde{\bar{Y}}_a^{synr} - \hat{\bar{Y}}_a\right)^2 - \hat{V}\left(\hat{\bar{Y}}_a\right) \qquad (1.6)$$

The direct estimators and its variance for areas with small sample size can be unstable and consequently so can be (1.6). Gonzalez (1973) suggests to use the average of the $\hat{MSE}(\tilde{\bar{Y}}_a^{synr})$ over a as a global stable indicator of the MSE. However, this indicator might be misleading as it does not represent an area-level specific measure of the MSE.

If individual-level information is not available, area-level variables can be used instead to produce ***Area-level Synthetic Regression Estimates*** (Skinner 1991), described by Ericksen (1974). Let $\bar{Z}_{ak}$, $k=1,...K$, be a set of K area-level auxiliary variables. The Area-level Synthetic Regression Estimator is given by,

$$\tilde{\bar{Y}}_a^{synra} = \sum_{k=1}^{K} \bar{Z}_{ak} \tilde{\beta}_k \qquad (1.7)$$

where $\tilde{\beta}_k$ is the estimated regression coefficient for the model underlying (1.7), that is,

$$\overline{Y}_a = \left(\sum_{k=1}^{K} \overline{Z}_{ak} \beta_k \right) + \varepsilon_a$$

with the mean of ε_a equal to zero.

Various improvements to (1.4) have been proposed. For instance, Nichol (1977) suggested including the synthetic estimator as an extra independent variable in a conventional regression estimator. Another approach was suggested by Battese et al. (1988) in which the sample area-mean of the random component ε_{ai}, $\overline{e}_a$ is used as an estimate of the area specific effect u_a. That estimate is added to (1.4) multiplied by a suitable proportion p_a; that is,

$$\tilde{\overline{Y}}_a^{com} = \tilde{\overline{Y}}_a^{synr} + p_a \overline{e}_a \tag{1.8}$$

If we define the conventional regression estimator (Skinner 1991) as,

$$\hat{\overline{Y}}_a^{cr} = \hat{\overline{Y}}_a + \sum_{j=1}^{J} (\overline{X}_{aj} - \hat{\overline{X}}_{aj}) \hat{\beta}_j$$

and knowing that $\overline{e}_a = \hat{\overline{Y}}_a - \left(\left(\sum_{j=1}^{J} \hat{\overline{X}}_{aj} \beta_j \right) \right)$, we just have to add $\left(\sum_{j=1}^{J} \overline{X}_{aj} \beta_j - \sum_{j=1}^{J} \overline{X}_{aj} \beta_j \right)$ to $\overline{e}_a$ to rewrite (1.8) as,

$$\tilde{\overline{Y}}_a^{com} = p_a \hat{\overline{Y}}_a^{cr} + (1 - p_a) \tilde{\overline{Y}}_a^{synr} \tag{1.9}$$

These estimators are commonly referred to as ***Composite Estimators (CE)***. In general, different combinations of indirect and direct estimators can be used in a CE. Purcell et al. (1980) suggest the use of the National mean direct estimator $\hat{\overline{Y}}$ or some

other larger area mean $\hat{\bar{Y}}_l$ instead of the indirect estimator in expressions like (1.9), as a convenient alternative in situations where auxiliary data is lacking or not sufficiently reliable. They also suggest the use of a multifactor classification as an alternative in specific situations, leading to an expression of the form,

$$\tilde{\bar{Y}}_a^{commf} = p_a \hat{\bar{Y}}_a + \sum_{i=1}^{I} p_{ai} \tilde{\bar{Y}}_{ai} \tag{1.10}$$

where $\hat{\bar{Y}}_a$ represents a direct estimator, $\tilde{\bar{Y}}_{ai}$ $i=1,...,I$, represents I different predictors for $\bar{Y}_a$ and $\left(p_a + \sum_{i=1}^{I} p_{ai} \right) = 1$.

Different methods of computing the weights p_a have been proposed in the literature based on design-based arguments. Let $\hat{\bar{Y}}_a$ be the direct estimator used in (1.9); assuming $Cov(\tilde{\bar{Y}}_a^{synr}, \hat{\bar{Y}}_a) = 0$, we obtain optimal weights minimising $MSE(\tilde{\bar{Y}}_a^{com})$,

$$MSE(\tilde{\bar{Y}}_a^{com}) = p_a^2 V(\hat{\bar{Y}}_a) + (1-p_a)^2 MSE(\tilde{\bar{Y}}_a^{synr})$$

with respect to p_a, so that,

$$p_a^{opt} = \frac{MSE(\tilde{\bar{Y}}_a^{synr})}{MSE(\tilde{\bar{Y}}_a^{synr}) + V(\hat{\bar{Y}}_a)} \tag{1.11}$$

These optimal weights can be estimated using (1.6) but they have the same instability problem as (1.6). However, empirical studies have suggested that composite estimators tend to be insensitive to deviations from the optimal weights (see e.g. Lundström 1987).

Purcell and Kish (1979) propose the use of a single weight $p_a = p \ \forall a$. An optimal estimator for the common weight p is obtained by minimising $\sum_{a=1}^{A} MSE(\tilde{\bar{Y}}_a^{com})/A$ with respect to p, resulting in the following expression,

$$p^{opt} = \frac{\sum_{a=1}^{A} MSE(\tilde{\bar{Y}}_a^{synr})}{\sum_{a=1}^{A} MSE(\tilde{\bar{Y}}_a^{synr}) + \sum_{a=1}^{A} V(\hat{\bar{Y}}_a)}$$

Thompson (1968) considers the case of shrinking a direct estimator $\hat{\theta}$ towards a "natural origin" θ_0 using a shrinking factor c. Let $\hat{\theta} = \hat{\bar{Y}}_a$, $\theta_0 = \bar{Y}_0$ and $c = p_a$. This estimator, called a ***"Shrinkage" estimator*** (Thompson, 1968), also has the form of a composite estimator,

$$\tilde{\bar{Y}}_a^{shr} = p_a \hat{\bar{Y}}_a + (1 - p_a)\bar{Y}_0 \tag{1.12}$$

Following Thompson (1968), an optimal estimator for p_a is,

$$p_a^{opt} = \frac{(\hat{\bar{Y}}_a - \bar{Y}_0)^2}{(\hat{\bar{Y}}_a - \bar{Y}_0)^2 + v(\hat{\bar{Y}}_a)} \tag{1.13}$$

Note that (1.13) is a special case of (1.11) where $\bar{Y}_0 = \tilde{\bar{Y}}_a^{synr}$.

Another approach that uses the idea of shrinking a direct estimator towards a conjecture is the ***James-Stein estimator*** (James and Stein, 1961),

$$\tilde{\theta}_a^{JS} = p_a^{JS}\hat{\theta}_a + (1 - p_a^{JS})\theta_a^0 \tag{1.14}$$

where $\hat{\theta}_a$ is the direct estimator of the function $\theta_a = f(\bar{Y}_a)$, with values that are independently $N(\theta_a, V)$ distributed and $p_a^{JS} = 1 - [(A-2)V/S]$ with $S = \sum_{a=1}^{A}(\hat{\theta}_a - \theta_a^0)^2$. Here θ_a^0 is our guess. Finally, $\tilde{\bar{Y}}_a^{JS} = f^{-1}(\tilde{\theta}_a^{JS})$.

Efron and Morris (1972) noted that $\tilde{\bar{Y}}_a^{JS}$ may perform poorly when the guessed values θ_a^0 are not close to θ_a. They proposed a "restricting" rule to avoid estimates that differ from $\hat{\theta}_a$ more than $\pm c$ times the value of $\sqrt{V}$. That is, the final estimate is $\tilde{\theta}_a^{JS}$ if that value is between $\hat{\theta}_a \pm c\sqrt{V}$ and either $\hat{\theta}_a + c\sqrt{V}$ or $\hat{\theta}_a - c\sqrt{V}$ depending whether $\tilde{\theta}_a^{JS}$ exceeds the upper or lower bound of $\hat{\theta}_a \pm c\sqrt{V}$. Denoting the resulting estimator as $\tilde{\theta}_a^{RJS}$, we have the ***Restricted JS estimator*** $\tilde{\bar{Y}}_a^{RJS} = f^{-1}(\tilde{\theta}_a^{RJS})$.

Other approaches to defining the p_a weights lead to another kind of composite estimator called a ***Sample-size dependent estimator.*** In this case the weight depends on the estimator of the size of the small area population $\hat{N}_a$. Drew et al. (1982) suggest using just the direct estimator, i.e. $p_a^{ssd} = 1$, if $\hat{N}_a$ is higher or equal to δN_a and $p_a^{ssd} = (\hat{N}_a / \delta N_a)$ otherwise. Here, δ is a suitable factor that needs to be chosen.

Hidiroglou et al. (1985) propose the use of $p_a^{MRE} = (N_a / \hat{N}_a)$ ("modified regression" factor). Later, Särndal et al. (1989) proposed that p_a^{MRE} be modified by imposing a "dampening" factor $(\hat{N}_a / N_a)^H$ in order to avoid "extreme" estimates for very small areas. They suggest the following rule,

$$P_a^{DMRE} = \begin{cases} (N_a / \hat{N}_a) & \text{if } \hat{N}_a \geq N_a \\ (\hat{N}_a / N_a)^{H-1} & \text{if otherwise} \end{cases}$$

Holt and Holmes (1994) propose an estimator for unequal probability designs in situations in which borrowing information across small domains is not feasible. Let the target of inference Y_k, $k=1,...,K$ be the unknown population total of units with a specific characteristic k. Let $y_{kh} = \sum_{i=1}^{n_h} y_{khi}$ be the sample count of units with characteristic k in sampling design stratum h, $h=1,...,H$, where y_{khi} is a Bernoulli variable with $\Pr(y = 1/h) = P_{k/h}$. Also let N_h and n_h be the population total and the

sample size for the design stratum h so that $\bar{y}_{kh} = y_{kh} / n_h$. The maximum likelihood (ML) estimator for $Y_k = \sum_{h=1}^{H} N_h P_{k/h}$ is the stratified direct estimator $\hat{Y}_h = \sum_{h=1}^{H} N_h \bar{y}_{kh}$, which is unreliable if the sample sizes within strata are small.

Now suppose we can define a new stratification f, $f=1,...,F$, so that f cuts the population across the original strata h and so that $\Pr(y_{fkhi} = 1) = P_{k/hf}$ can be considered constant across h, i.e. $P_{k/hf} = P_{k/f}$. The expected value for the population total Y_k can now be rewritten as $E(Y_k) = \sum_{h=1}^{H} N_h Q_{f/h} P_{k/f}$, where $Q_{f/h}$ is the probability of belonging to stratum f given the unit belongs to the original stratum h. The ML estimator for this expected value is finally,

$$\tilde{Y}_k = \sum_{h=1}^{H} \sum_{f=1}^{F} N_h \frac{y_{hf}}{n_h} \frac{y_{kf}}{y_f} = \sum_{h=1}^{H} \hat{Y}_f \frac{y_{kf}}{y_f} \qquad (1.15)$$

where y_{hf} is the sample total of units belonging to stratum h and stratum f at the same time, y_f is the sample total of units in stratum f and $\hat{Y}_f$ is the stratified direct estimator of the population total of units in stratum f. The resulting estimator (1.15) borrows strength across the design strata h and Holt and Holmes (1994) shows that it is more accurate than the traditional stratified estimator provided the assumption $P_{k/hf} = P_{k/f}$ holds.

The ***Structure Preserving Esimation (SPREE)*** method also belongs to the class of synthetic methods of estimation. Purcell and Kish (1980) describe this method which is based on application of the Iterative Proportional Fitting algorithm (IPF) (Deming and Stephan., 1940). SPREE requires the specification of an association structure linking the distribution of the target variable and some covariates at small area level and an allocation structure characterising the current relationship between these variables at a larger area level. The information necessary for specifying the association structure is typically obtained from recent censuses or administrative records. The current information required for the allocation structure is typically

obtained from national surveys or any other reliable source. The basic idea behind the SPREE method is: "(i) to conform to the current information in the allocation structure; and to (ii) preserve somehow the earlier relationships present in the association structure without interfering with aim (i) above." (Purcell and Kish, 1980).

These authors describe different situations that might arise, depending on the available information, for the specific case of a categorical target variable required for certain small domains and one categorical auxiliary variable. Let N_{aij} be a set of counts from a cross-tabulation defining the association structure from, say, the last census. Here a and j refer to area and category of auxiliary variable whilst i, $i=1,...,I$, represents the category of the target variable. Let m_{aij} be the set of "unknown" counts of which we only know or have reliable estimates of certain aggregates corresponding to the allocation structure. They consider situations where the information available is: (a) N_{aij}, $m_{\cdot ij}$, (b) N_{aij}, $m_{\cdot ij}$, $m_{a\cdot\cdot}$, (c) N_{aij}, $m_{\cdot ij}$, $m_{a\cdot j}$, (d) $N_{a\cdot j}$, $m_{\cdot ij}$, (e) $N_{a\cdot j}$, $m_{\cdot ij}$, $m_{a\cdot\cdot}$ and (f) $N_{a\cdot j}$, $m_{\cdot ij}$, $m_{a\cdot j}$. Cases d, e and f represent special situations in which a full association structure is not available. They use a weighted least squares approach and minimize the loss function,

$$f(\tilde{m}_{aij}) = \sum_{a=1}^{A}\sum_{i=1}^{I}\sum_{j=1}^{J} N_{aij}^{-1}(\tilde{m}_{aij} - N_{aij})^2 - \sum_{c=1}^{C}\lambda_c\left(\phi_c(\tilde{m}_{aij})\right)$$

where $\phi_c(\tilde{m}_{aij})$, $c=1,...,C$, denote the c*th* constraining equation in the Lagrangian $f(m_{aij})$, corresponding to the marginal constraints specified by the allocation structure. For case (a), for instance, there is only one constraint, i.e. $\phi_c(\tilde{m}_{aij}) = \sum_{a=1}^{A}\tilde{m}_{aij} - m_{\cdot ij}$, and the estimate of the target count $m_{ai\cdot}$ is given by

$$\tilde{m}_{ai\cdot} = \sum_{j=1}^{J}\frac{N_{aij}}{N_{\cdot ij}}m_{\cdot ij} \qquad (1.16)$$

For cases (b) and (c) they apply the IPF algorithm to obtain maximum likelihood estimates of $m_{ai\cdot}$. For cases (d), (e) and (f), where a full association structure N_{aij} is

not available, they define a "dummy association structure" by assuming proportionality across the i categories, so that a full set of estimated counts $\tilde{N}_{aij}$ is used in the analysis.

The assumption behind the use of SPREE in (*a*) and (*b*) is that the interactions from the association structure of higher order than the one defined by the allocation structure is the same as the corresponding interactions of same order for the target counts. Using again case (*a*) as example, the assumption behind (1.16) is $N_{aij}/N_{\cdot ij} = m_{aij}/m_{\cdot ij}$. For SPREE estimates in cases (*d*), (*e*) and (*f*), an additional assumption is that the missing interactions from the association structure are not relevant i.e. the association structure (odd ratio) for that specific missing component equals one, for the set N_{aij} or for the set m_{aij}, so that in case (*d*), for instance, $N_{aij}/N_{\cdot ij} = N_{a\cdot j}/N_{\cdot\cdot j} = m_{aij}/m_{\cdot ij}$.

Note that the gains in precision from using SPREE are due to the fact that, thanks to the information from the past used in the SPREE process, we only use present data to estimate the marginal given by the allocation structure. That is, if the information regarding the allocation structure comes from a national survey that leads to poor direct estimates for m_{aij} and $m_{ai\cdot}$, then it might be possible that reliable direct estimates of an aggregate like $m_{\cdot ij}$ can be identified so that (1.16) produces estimates with a lower variance than direct estimators. Here again, the superiority of the SPREE estimator over the direct estimator depends on the trade off between gains in precision and increases in bias due to departures from the assumptions behind the method.

A more detailed description of the SPREE method and some of its variants is set out in Chapter 3, where it is also put into the context of the specific LFS problem of interest in this thesis.

Feeney (1987) carried out an evaluation study of the use of the SPREE method to produce estimates of total unemployed for Local Government Areas in four states of Australia. He used information from the Department of Social Security (DSS) related

to unemployment benefits to construct the association structure and the 1981 Population Census to obtain the allocation structure. He compared his results with the "real" values from the Census and argued that, should this evaluation show SPREE methods work well using the Census data, it should also work well using the Monthly Labour Force Survey (MLFS). Two variables, Sex-marital status and Age group, were used in the analysis. The target set of counts were $m_{a..} = \sum_{ij} m_{aij}$, where m_{aij} represents the total of unemployed for the small area *a,* sex-marital status *i* and age group *j*. The allocation structure was defined by the marginals $m_{.i.}$ and $m_{..j}$ taken from the Census. To compare the SPREE estimates with MLFS direct estimates, Feeney derived estimates of the standard error of the SPREE estimates by assuming these estimates had the same error structure as the MLFS estimates. The results showed that in general, the SPREE methods provide improved estimates compared with MLFS direct estimates.

Lundström (1987) carried out an evaluation study comparing two direct estimators, three SPREE estimators and a three composite estimator using as indirect estimators each of the SPREE estimators to estimate the number of nonmarried cohabiting persons in Swedish municipalities. He found that SPREE methods were superior to the direct estimators. Apparent differences were found between SPREE estimators and composite estimator.

Griffiths (1996) used SPREE estimates within a composite estimator to produce employment and household income estimates for Congressional Districts in Iowa State, U.S.A. He used the estimator (1.16) to compute the SPREE estimates. The allocation structure was obtained from the 1994 Current Population Survey (CPS) and the association structure was produced using the 1990 US Decennial Census.

1.5.2.c. Specific Area-level Random Effects Model Estimators

Synthetic estimators have the advantage of being simple to implement. They commonly lead to estimates that are more precise than comparable direct estimators

for small areas. If good auxiliary information is available and an appropriate model can be formulated so that the error specification in the implicit (or explicit) model behind the estimator, i.e. $E(\varepsilon_{ai}) = 0$ in (1.5), is reasonable, synthetic estimators are highly accurate; in these cases they are certainly the appropriate choice for producing estimates for small areas. However, in the presence of an important misspecification of the error structure, e.g. $\varepsilon_{ai} = u_a + v_{ai}$, $E(\varepsilon_{ai}) = u_a$ with u_a large, synthetic estimators lose accuracy; if the gains in precision do not offset the bias caused by u_a synthetic estimators can be even less accurate than traditional direct estimators.

In those situations, specific area-level random effect model methods provide a better approach to small area estimation; in particular, since they take into account local variation they can be more efficient than synthetic estimators in situations like the one described above. Another advantage of those methods is that they offer the possibility of obtaining "stable area specific measures of variability associated with the estimates" (Rao, 2000).

Specific area-level random effect models also offer the flexibility to formulate and handle complex cases. In this sense, traditional random effect models has been extended to more complex situations such as multivariate models (see e.g. Datta et al. 1999, Datta et al. 1996, Cressie 1992, Freedman et al 1992), time series models (see e.g. Datta et al. 1999, Pfeffermann et al. 1998, Tiller 1992, Pfeffermann et al. 1990), multilevel models (see e.g. Moura et al. 1999, You et al. 1999) and logistic models (see e.g. Jian et al. 1999, Malec et al. 1999, Booth et al. 1998, Malec et al. 1997, Farrel et al. 1997a, Farrel et al. 1997b).

For comprehensive reviews and appraisals of specific area-level random effects model-based methods of SAE see Gosh and Rao (1994), Rao (1999), Pfeffermann (1999) and Rao (2000).

CHAPTER 2

DESCRIPTION OF THE DATA

Two main sources of data are used in this study: the last Venezuelan Population Census (1990) and the Labour Force Survey. In the following sections, we describe these sources in more detail. We focus particularly on the LFS parameter and variance estimation methodology since a detailed description of this does not currently exist.

2.1. THE VENEZUELAN 1990 CENSUS

The Venezuelan Population Census carried out in 1990 (Census'90)[1], is the most important source of local level information in the country. Its information is not only used for descriptive and analytical purposes but also as the base for sampling designs and field operations and as auxiliary information for survey estimation. The LFS sampling design is based on the cartographic and demographic database of the Census'90. Sex-age population projections based on this census are used by the LFS as auxiliary information for Post-stratification estimators. It is also the basis for the simulation analysis reported later in this thesis. Consequently it is appropriate that we now describe some aspects of this census that are relevant.

[1] XII Censo de Poblacion y Viviendas.

The Census'90 was carried out via face to face household level interviews. On the 21st of October 1990 about 80% of the country was interviewed. The remaining 20% were interviewed within the next 30 days, keeping the same reference date i.e. 20-10-1990.

2.1.1. General Format, Field Organisation and Sampling Design

The Census'90 was a "mixed" operation in which a set of basic demographic variables was obtained for the whole population (Basic Questionnaire) whilst a supplementary set of variables was additionally obtained from a sample (Expanded Questionnaire). The sample consisted of the whole population in the rural areas and approximately 20% of the population in the urban areas as explained below.

For this Census, private addresses (PA) in Venezuela were partitioned into physical groups called "*Segments*" (urban areas) of approximately 200 PA each and "*Sectors*" (rural areas) of sizes around 100 PA. These Segments and Sectors are mutually exclusive and exhaustive for PA. In the urban areas, each Segment was partitioned into approximately 10 "*Sections*" of 20 PA. Finally, these Sections were partitioned into 2 "*Subsections*" of 10 PA each. This hierarchical structure reflects the logistics of the execution of the census. A Segment contains the quantity of PA that one supervisor was able to handle with ease in one day. In the same way, the amount of work assigned to one interviewer consisted of either one Section if he/she was applying the Basic Questionnaire or one Subsection if he/she was applying the Expanded Questionnaire. In the rural areas each Sector corresponded to a population settlement (towns). Each Segment and Sector was identified in the official maps produced by "National Cartography"[2] whilst sketches were drawn for each Section and Subsection; a written description for each of those units (Segments/Sectors, Sections and Subsections) was recorded. A code system was used to link the information registered for each address in the Census database and the respective

[2] Cartografia Nacional. This is the official body of the Venezuelan government responsible for building and maintaining the cartography of the country.

units to which it belonged. This cartographic and demographic database is now used for the design and implementation of the household surveys carried out by OCEI.

The expanded questionnaire sample in the urban areas was selected using a stratified cluster sampling design. The strata and the clusters were Segments and Sections respectively. Around 20% of the Sections were selected within each Segment with equal probability, i.e. 2 out of 10 Sections in most of the cases. Finally, the Expanded Questionnaire was applied to all households within the selected Sections.

2.1.2. The 1990 Census Variables and Their Correspondence with The LFS Variables

The variables needed to construct the basic labour force indicators were collected in the Basic Questionnaire. Two databases were built from these questionnaires, one containing the first set of variables for the whole population (Basic Database) and the other containing the second set of variables obtained using the expanded questionnaire (Expanded Database). Since it was not possible to obtain the Basic Database from OCEI, the one used in this thesis is the Expanded Database.

An important aspect of the Census'90 is the fact that all the labour force variables collected in the census have the same conceptual definitions as the ones collected in the LFS. Therefore, any difference between these sources is mainly related to practical issues such as non-sampling errors, which are expected to be larger in an exhaustive enumeration like the census than in a sampling survey situation like the LFS.

Another important practical difference between these sources is found in the process used to classify people over 15 years old into activity/non-activity and employee/unemployed groups. The LFS uses an algorithm based on a set of additional questions to carry out this classification process (Summary Code). The Census'90 in contrast used a single question to classify people. Thus, although both sources follow the same conceptual definitions, the LFS is expected to get a better measurement. The

Tabla 2.1
Venezuela. Clasification of the Labour Force with and without Summary Code (SC) and Employment, Unemployment and Activity Rates
(Cells are percentages calculated for rows)

Cateory without SC	Cateory with SC				Rates
	Employee	Unemployed	Non-Active	Total	
Employee	99.91	0.05	0.04	48.08	92.04
Unemployed	4.88	93.90	1.22	4.16	7.96
Non-Active	4.01	5.11	90.88	47.76	47.76
Total	50.15	6.37	43.47	100	
Rates	88.73	11.27	43.47		

impact of the summary code on the classification outcomes can be seen from the difference in the LFS estimates with and without using the code. Table 2.1 shows examples of these differences for the second semester of the 1998 LFS. In this table, the columns and the rows represent the LFS classification taking and without taking into account the Summary Code.

As it can be seen from the table, use of the Summary Code increases the percentage of people classified as Unemployed and Employed, due mainly to a reclassification of people initially classified as Non-Active (4.01% reclassified as Employed and 5.11% reclassified as Unemployed). The category Unemployed rises from 4.16 to 6.37, that is, a relative increment of 53%, whilst Employed increases from 48.08 to 50.15, a relative increment of 4.3%. Consequently, the Employment Rate and the Unemployment Rate decrease (from 92.04 to 88.73) and increase (from 7.96 to 11.27) 3.31 units respectively, whilst the Non-Activity Rate decreases 4.19 units, from 47.76% to 43.47%. Therefore, if we assume the classification produced by using the Summary Code is the "correct" one, we can expect both the Unemployment Rates and the Activity Rates produced by the Census to underestimate the real figures.

This information can be used to adjust Census values is an attempt to make them comparable to LFS outcomes. This adjustment can be carried out as a simple multiplication of the Census rates by the variation adjustment estimated from the LFS. Given the characteristics of our study, such adjustment would require a set of tables like table 2.1 at sub-population levels, preferably for the Census year. These sub-populations, as described in Chapter 1, consist of combination of both spatial (states) and demographic (sex-age groups) dimensions. Although the estimates set out in these tables would then be based on small sample sizes, the adjustments we are interested in are the differences between two highly correlated variables and therefore should be reliable enough for our purpose.

The adjustment described above is useful when applying methods that involve the use of labour force indicators from both the Census and the LFS. Obviously, it is unnecessary when the information required from the Census does not consist of labour force indicators, for instance sex-age population totals. In this thesis we study SPREE methods for producing reliable sub-groups estimates. SPREE methods are explained in the next chapter and they involve the use of labour force indicators from both databases. However, our study of SPREE is mainly based on simulations from the Census data and does not involve the use of the LFS database. Therefore, no adjustments were necessary in our study. Nevertheless, we emphasize that such adjustment would be necessary if this technique was apply in a real life situation.

2.2. THE LABOUR FORCE SURVEY

The LFS is the oldest continuous survey run in Venezuela. This survey has supplied the country with valuable information related to households since 1967. Despite the fact that the LFS mainly aims to produce information regarding the Labour Force, this survey has always been an important general source of social information related to households.

2.2.1. General Format of the LFS

This survey is run twice a year, 22 weeks during the first six months of the year and 22 weeks during the last six months. The sample of the survey is distributed randomly across the 22 weeks in such a way that each week can be considered as a sub-sample of the country. The information is collected using direct interviews in which a "*interviewer*" fills out a questionnaire with the information provided by an "*interviewed*" on behalf of every member of the household. The person interviewed can be any member of the household older than eighteen years. The current questionnaire contains 13 questions about the characteristics of the house, 9 questions about the household and 62 about the inhabitants of the households, including the10 questions that comprise the Summary Code referred to in the previous section.

2.2.2. LFS Sampling Design

Although the sampling design has changed three times since 1967, its main features have remained the same, i.e. a stratified three-stage sampling design. We now describe this design pointing out the differences, where relevant, between the 1985-1993 design and the current design.

From 1985 to 1993 the strata were the nine regions described in section 1.3. Currently, the strata consist of spatial areas within states (see table 2.2). These areas are made up of neighbouring counties with similar characteristics. In particular, each area is expected to be internally homogeneous with respect to economic activities and services.

The first stage of selection is a random selection of segments and sectors (Primary Sampling Units, PSU) from the census database. In the 1985-1993 design, selection was made with probabilities proportional to the number of private properties registered within each PSU following the 1981 census. A geographically ordered systematic procedure was used for this selection. In the current design, this selection was carried out in two phases. In the first phase, a group of PSUs were independently

Table 2.2 1/2

Venezuela. Labour Force Survey (LFS)

Sample Size (PSU) by Selection Strata, Current Design, 98-II

STATES (s)	SELEC. STRATA (h)	NAMES	SEGMENTS (PSU)		
			Census	Master Sample	LFS
		VENEZUELA	**21854**	**6078**	**1833**
1		**DTTO. FEDERAL**	**2734**	**400**	**261**
	ST.1	A.M. DE CARACAS	2319	250	221
	ST.2	MUNICIPIO VARGAS	415	150	40
2		**ANZOATEGUI**	**1156**	**450**	**66**
	ST.3	A.M. DE BARCELONA-PTO. LA CRUZ	484	200	28
	ST.4	TRS	672	250	38
3		**APURE**	**386**	**83**	**24**
	ST.5	A.M. DE SAN FERNANDO	151	28	12
	ST.6	MUNICIPIO PAEZ	86	42	6
	ST.7	TRS	149	13	6
4		**ARAGUA**	**1335**	**310**	**74**
	ST.8	A.M. DE MARACAY	901	160	56
	ST.9	TRS	434	150	18
5		**BARINAS**	**598**	**220**	**28**
	ST.10	A.M. DE BARINAS	191	90	9
	ST.11	TRS	407	130	19
6		**BOLIVAR (a)**	**1149**	**330**	**115**
	ST.46	MUNICIPIO CARONI		167	61
	ST.47	MUNICIPIO HERES		83	26
	ST.48	LOCALIDAD UPATA		15	4
	ST.49	MUNICIPIO CEDEÑO, EL CALLAO, GRAN SABANA, PIAR, MUNICIPIO PIAR, RAUL LEONI		65	24
7		**CARABOBO**	**1713**	**500**	**81**
	ST.12	A.M. DE VALENCIA	1327	300	63
	ST.13	TRS	386	200	18
8	ST.14	**COJEDES**	**240**	**120**	**11**
9		**FALCON (a)**	**920**	**350**	**50**
	ST.15	MCP : ZAMORA, COLINA, MIRANDA	195	94	11
	ST.16	MCP : CARIRUBANA, LOS TAQUE, FALCON	191	100	10
	ST.17	TRS	534	156	29
10		**GUARICO**	**704**	**300**	**34**
	ST.18	A.M. DE CALABOZO	124	60	6
	ST.19	TRS	580	240	28
11		**LARA (a)**	**1602**	**300**	**231**
	ST.20	A.M. DE BARQUISIMETO	869	150	126
	ST.21	TRS	733	150	105
12		**MERIDA**	**767**	**310**	**44**
	ST.22	A.M. DE MERIDA	265	140	15
	ST.23	TRS	502	170	29

Table 2.2 2/2

Venezuela. Labour Force Survey (LFS)

Sample Size (PSU) by Selection Strata, Current Design, 98-II

SELEC. STRATA (h)	SELEC. STRATA (h)	NAMES	SEGMENTS (PSU)		
			Census	Master Sample	LFS
13		**MIRANDA**	**2629**	**565**	**212**
	ST.24	REGION LOS TEQUES	347	120	28
	ST.25	REGION BARLOVENTO	313	65	25
	ST.26	VALLES DEL TUY	467	80	38
	ST.27	EJE GUARENAS GUATIRE	274	100	22
	ST.28	METROPOLITANA	1228	200	99
14		**MONAGAS**	**646**	**200**	**34**
	ST.29	A.M. DE MATURIN	315	80	16
	ST.30	TRS	331	120	18
15	ST.31	**NVA. ESPARTA**	**357**	**170**	**16**
		A.M. DE PORLAMAR - PAMPATAR	166	80	7
		TRS	191	90	9
16		**PORTUGUESA (a)**	**704**	**250**	**150**
	ST.32	A.M. DE ANAGUA -ARAURE	222	60	38
	ST.33	TRS	482	190	112
17		**SUCRE**	**862**	**330**	**52**
	ST.34	A.M. DE CUMANA	241	110	14
	ST.35	TRS	621	220	38
18		**TACHIRA**	**447**	**185**	**58**
	ST.36	A.M. DE SAN CRISTOBAL	376	150	25
	ST.37	SAN ANTONIO DEL TACHIRA - PEDRO	71	35	11
	ST.38	TRS	615	100	22
19		**TRUJILLO**	**731**	**200**	**31**
	ST.39	A.M. DE VALERA	193	80	8
	ST.40	TRS	538	120	23
20		**YARACUY**	**513**	**175**	**26**
	ST.41	A.M. DE SAN FELIPE	112	55	6
	ST.42	TRS	401	120	20
21		**ZULIA (a)**	**2581**	**590**	**315**
	ST.43	A.M. MARACAIBO	1501	300	195
	ST.44	ZONA ORIENTAL DEL LAGO	577	170	68
	ST.45	TRS	503	120	52
22	ST.50	**T.F. AMAZONAS**	**95**	**30**	**15**
23	ST.51	**DELTA AMACURO**	**134**	**40**	**20**

* TRS = The rest of the state

* *A.M. = Metropolitan Areas*

* *(a) = Sample size extended on OCEI-Governors agreements*

Table 2.3

VENEZUELA. LABOUR FORCE SURVEY (LFS)
SAMPLE SIZE – DESIGN 1985-1993

Regions	LFS Sample Size (PSU) 1985-1993
VENEZUELA	3,794
Capital - Metropolitan Area of Caracas - Rest	613 371 242
Central	414
Centro-Occidental	485
Guayana	734
Los Llanos	100
Los Andes	471
Nor-Oriental	489
Zulia	488

selected within each stratum for the MMSV[3]. A total of 6,078 PSUs were selected in this phase with probabilities proportional to the number of PA registered within each PSU following the 1990 census. The systematic procedure used in this case followed a geographical order in the rural areas and a combined socio-economic order and a geographical order in the urban areas. The socio-economic order was defined by a classification of segments made by OCEI according to a special method. For the second phase, a sub-sample of PSUs from the first phase was selected using the same procedure explained above, but with equal probabilities. This sub-sample is the set of PSUs used by the LFS

The second stage of selection was a random selection of sub-segments (Secondary Sampling Units, SSU). These sub-segments are spatial divisions of the segments made for sampling purposes. Each segment was divided into four sub-segments of

[3] Master Sample. This is a sample of primary units used as the base for the design of any survey in which the observation units are houses, households and/or persons.

Table 2.4

VENEZUELA. LABOUR FORCE SURVEY (LFS)
STATES AND PERIODS FOR WHICH AGREEMENTS WERE SIGNED

90-I – 98-II

State	I 90	II 90	I 91	II 91	I 92	II 92	I 93	II 93	I 94	II 94	I 95	II 95	I 96	II 96	I 97	II 97	I 98	II 98
Metropolitan Area of Caracas 1																		
Aragua																		
Falcon																		
Bolivar-D.A.-T.F.A.																		
Lara																		
Portuguesa																		
Tachira																		
Zulia																		

1 It is comprised by Dtto.Federal and 4 Miranda State counties (Metopolitan Area)

approximately 50 private properties. Two sub-segments were independently selected within each segment. The selection was made with probabilities proportional to the number of PA within each area determined by a "reference count" carried out after the sub-segments had been created. A systematic procedure was used for this selection. In rural sectors only one sub-segment was selected.

The third stage of selection was a random selection of PA (Tertiary Sampling Units, TSU). Approximately five PA were independently selected within each sub-segment with equal probabilities. A systematic procedure was used for this selection from an updated list of PA constructed for each sub-segment. In rural sectors approximately ten PA per sub-segment were selected.

In the 1985-1993 design, the 3,794 PSU roughly yield a sample size of 190,000 people of which approximately 120,000 are over fourteen years old. In the current design, for the second 1998 LFS run (98-II), the 1,833 PSU yield a sample size of approximately 110,000 people of which about 70,000 are over fourteen years old. The

sample sizes for both designs are shown in the tables 2.2 and 2.3. It is important to point out that agreements between OCEI and the governors of same States have been signed in order to increase the sampling size for these states. The states involved in such agreements as well as the periods for which these agreements have been signed are shown in table 2.4. The sample sizes comprising these agreements for the 98-II LFS run are reflected in table 2.2.

The survey has a rotation system in which each PA selected remains in the sample for six runs (three years) and then is permanently dropped from the sample. Each Segment in the sample belongs to one of six "*rotation panels*", each panel being a sub-sample of the country. At every new run of the survey, one panel is "rotated"; i.e. the PA in the sample within the Segments belonging to those panels are replaced with other PA from the same segment.

2.2.3. Selection Probabilities

The LFS technical reports (OCEI 1987 and OCEI 1997) contain general details concerning the main characteristics of the survey. We now describe the estimation methodology used in the LFS.

The selection probability of any PA –and consequently any person who lives in this PA- is given by:

$$\mathrm{p(PA_{hijk})} = n_{oh} \cdot \frac{T_{hi}}{T_h} \cdot n_h \cdot \frac{1}{n_{oh}} \cdot b \cdot \frac{T'_{hij}}{T'_{hi}} \cdot \frac{c_{hij}}{T''_{hij}} \qquad (2.1)$$

where T_h is the total of PA (Census) in the hth stratum (h=1,...,H). T_{hi} is the total (Census) of PA in the ith PSU, in the hth stratum (i=1,..., N_h). T'_{hi} is the total of PA after the reference count, in the ith PSU, in the hth stratum. T'_{hij} is the total of PA after the reference count, in the jth SSU, in the ith PSU, in the hth stratum (j=1,..., B_{hi}). T''_{hij} is the total of PA after the exhaustive list, in the jth SSU, in the ith PSU, in

the hth stratum. n_{oh} is total of PSU in the MMSV in the hth stratum. n_h is the total of PSU in the LFS sample in the hth stratum. c_{hij} is the total of PA in the LFS sample in the jth SSU, in the ith PSU, in the hth stratum. b is the number of SSU in the LFS, which equals 2 for segments and 1 for sectors.

The sampling design is self-weighting within each stratum:

$$c_{hij} = \frac{l \cdot T'_{hi} \cdot T''_{hij}}{T_{hi} \cdot T'_{hij}}$$

in which case (3.1) reduces to:

$$p(PA_{hijk}) = \frac{n_h \cdot b \cdot l}{T_h}$$

where l=5 and b=2 for urban areas and l=10 and b=1 for rural areas, so, $b \cdot l = 10$ in both cases.

The design weight attached to the kth PA in the sample in the jth SSU, in the ith PSU, in the hth stratum (k=1,..., T_{hij}) is then given by:

$$w_{hijk} = p(PA_{hijk})^{-1} = \frac{T_h}{n_h \cdot 10} \qquad (2.2)$$

Noting that (2.2) only depends on h, we have that $\forall \; i, j, k$:

$$w_{hijk} = w_h \qquad (2.3)$$

2.2.4. Parameter Estimators

The LFS estimator of total is the standard Horwitz-Thompson estimator but adjusted for non-response and for post-stratification.

The non-response adjustment is made for E socio-economic groups. The segments within a stratum are grouped into socio-economic classes using the information on the census database. However, the original census classification for a specific segment is sometimes updated when significant changes of the standard of living in the area are observed. The weight of each PA is then "inflated" by the factor c_{he} / c'_{he} $(e=1,...,E)$ where the numerator is the total of PA in the sample in the eth socio-economic group hth stratum and the denominator is the total of PA successfully interviewed in the same group in the same stratum. In the LFS technical documentation there is no formal justification for this adjustment. However, it is clear that the implicit model assisting this procedure is a group mean model as in Sarndal et al. (1992, p.264), in which it is assumed that a common mean $E_{\xi}(y_{heijk}) = \beta_{he}$ and variance $V_{\xi}(y_{heijk}) = \sigma^2_{ohe}$ is shared by all PA within the same socio-economic group. The weight attached to the kth PA in the sample in the jth SSU, in the ith PSU, in the hth stratum after non-response adjustment then:

$$w'_{heijk} = w'_{he} = w_h \cdot \frac{c_{he}}{c'_{he}} \quad ; \ \forall \ i,j,k \tag{2.4}$$

The post-stratification adjustment is made using the projected state level census sex-age distribution at the reference time of the survey.

Let $a=1,2,3,...,A$ denote the $A=22$ sex-age sub-strata. Let ${}_aX_h$ be the projected census population of the hth stratum in the a sex-age group and let ${}_aX_s = \sum_{h \in s} {}_aX_h$ be the aggregate projected census population of the a sex-age group for the strata

comprising the sth state. The estimator of the total for a specific variable y is then given by:

$$\hat{Y}_s^R = \sum_{a=1}^{A}\left({}_a\hat{R}_s^y \cdot {}_aX_s\right) = \sum_{a=1}^{A}\left(\frac{{}_a\hat{Y}_s}{{}_a\hat{X}_s} \cdot {}_aX_s\right) = \sum_{a=1}^{A}\left(\frac{\sum_{h\in s}\sum_{e=1}^{E}\sum_{v=1}^{m_{he}}(w'_{he} \cdot {}_ay_{hev})}{\sum_{h\in s}\sum_{e=1}^{E}\sum_{v=1}^{m_{he}}(w'_{he} \cdot {}_ax_{hev})} \cdot {}_aX_s\right) \quad (2.5)$$

where ${}_ax_{hev} = 1$ if the vth element belongs to the sex-age group a and zero otherwise. This is a form of combined Post-stratification estimator. Note that $\sum_{h\in s}\sum_{e=1}^{E}\sum_{v=1}^{m_{he}}(w'_{he} \cdot {}_ay_{hev}) = {}_a\hat{Y}_s$ and $\sum_{h\in s}\sum_{e=1}^{E}\sum_{v=1}^{m_{he}}(w'_{he} \cdot {}_ax_{hev}) = {}_a\hat{X}_s$ are the Horwitz-Thompson estimators of ${}_aY_s$ and ${}_aX_s$ respectively.

Since ${}_ax_{hev} = 1$ for $x_{hev} \in a$, the implicit model assisting this procedure can also be regarded as a group mean model, in which it is assumed that a common mean $E_\xi({}_ay_{hev}) = {}_a\beta_s$ and variance $V_\xi({}_ay_{hev}) = {}_a\sigma_s^2$ is shared by all the individuals within the same sex-age group within a specific stratum.

The final weight attached to the vth person in the sample within the hth stratum and the eth soci-economic is:

$${}_aw''_{he} = w'_{he} \cdot \frac{{}_aX_s}{{}_a\hat{X}_s} \quad (2.6)$$

Therefore, (2.5) can be written as:

$$\hat{Y}_s^R = \sum_{a=1}^{A}\sum_{h\in s}\sum_{e=1}^{E}\sum_{v=1}^{m_{he}}({}_aw''_{he} \cdot {}_ay_{hev}) \quad (2.7)$$

Finally, the national level estimate is obtained as the sum of $\hat{Y}_s^R$ for all the Venezuelan States: $\sum_{s=1}^{S} \hat{Y}_s^R$.

Let y be a dichotomous variable (0 and 1) indicating the membership to a specific group defined by a characteristic of interest. The proportion of people that belong to that group is estimated by:

$$\hat{P}_s^R = \frac{\hat{Y}_s^R}{\hat{X}_s^R} = \frac{\hat{Y}_s^R}{X_s} \tag{2.8}$$

Indicators such as the unemployment rate are estimated as the ratio of two Post-stratification estimators. For instance, let ${}_a y_{hev}$ be equal to one if the vth person in (*ahe*) is over 14 year old and unemployed and zero otherwise. Let also ${}_a z_{hev}$ be equal to one if the vth person in (*ahe*) is over 14 year old and active (employed or unemployed) and zero otherwise. The Unemployment Rate as defined in section 1.4 is then estimated at state level as follows:

$$\hat{R}_s^R = \frac{\hat{Y}_s^R}{\hat{Z}_s^R} = \frac{\sum_{a=1}^{A}\sum_{h\in s}\sum_{e=1}^{E}\sum_{v=1}^{m_{he}}({}_a w_{he}^{*} \cdot {}_a y_{hev})}{\sum_{a=1}^{A}\sum_{h\in s}\sum_{e=1}^{E}\sum_{v=1}^{m_{he}}({}_a w_{he}^{*} \cdot {}_a z_{hev})} \tag{2.9}$$

At the national level this rate is estimated as:

$$\hat{R}^R = \frac{\sum_{s=1}^{S} \hat{Y}_s^R}{\sum_{s=1}^{S} \hat{Z}_s^R}$$

Estimates of a specific R_s^R for any sex-age group can be obtained by taking into account just the people who belong to this specific group in (2.9).

2.2.5. *Variance Estimators*

The variance of the LFS estimators is complicated by the fact that without replacement sampling is used to select PSUs. However, assuming that the PSUs were sampled with replacement within each stratum, we can estimate the variance and covariance of the Horwitz-Thompson estimators by using the ultimate cluster technique (see e.g. Kish 1965, Wolter 1985, Skinner et al. 1989, Sarndal et al. 1992):

$$\operatorname{cov}\left(\hat{Y}_h, \hat{Z}_h\right) = \frac{n_h}{(n_h - 1)} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \hat{\bar{Y}}_h\right)\left(\hat{Z}_{hi} - \hat{\bar{Z}}_h\right)$$

$$\operatorname{cov}\left(\hat{Y}_s, \hat{Z}_s\right) = \sum_{h \in s} \frac{n_h}{(n_h - 1)} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \hat{\bar{Y}}_h\right)\left(\hat{Z}_{hi} - \hat{\bar{Z}}_h\right) \qquad (2.10)$$

where $\hat{Y}_{hi} = \sum_{j=1}^{b} \sum_{v=1}^{m_{hij}} w'_{he} y_{hijv}$ is the Horwitz-Thompson estimators of Y_{hi} and, $\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi}$, $\hat{\bar{Y}}_h = \hat{Y}_h / n_h$, $\hat{Y}_s = \sum_{h \in s} \hat{Y}_h$. Similar definitions apply to $\hat{Z}_{hi}$, $\hat{\bar{Z}}_h$, and $\hat{Z}_s$. The assumption that the PSUs were sampled with replacement leads to a conservative variance estimator (e.g. Skinner et al. 1989, p.49).

To get the variance and covariance estimators of a total such as $\hat{Y}_s^R$ in (2.7), we replace $\hat{Y}_{hi}$ and $\hat{Z}_{hi}$ with $\hat{D}_{hi}^y$ and $\hat{D}_{hi}^z$ respectively in (2.10) where $\hat{D}_{hi}^y$ (and similarly for $\hat{D}_{hi}^z$) is the estimated sum of the residuals ${}_a\hat{e}_{hijv} = {}_a y_{hijv} - {}_a\hat{R}_s^y \cdot {}_a x_{hijv}$ weighted by w'_{he} for the ith segment in the hth stratum:

$$\hat{D}_{hi}^y = \sum_{a=1}^{A} \sum_{j=1}^{b} \sum_{v=1}^{m_{hij}} w'_{he}\, {}_a\hat{e}_{hijv} \qquad (2.11)$$

where ${}_a\hat{R}_s^y = \dfrac{{}_a\hat{Y}_s}{{}_a\hat{X}_s}$ and noting that:

$$\hat{\bar{D}}_h^y = \sum_{i=1}^{n_h} \hat{D}_{hi}^y / n_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \sum_{a=1}^{A} \sum_{j=1}^{b} \sum_{v=1}^{m_{hij}} w'_{he} \left({}_a y_{hijv} - {}_a \hat{R}_s^y \, {}_a x_{hijv} \right)$$

$$= \sum_{a=1}^{A} \left[{}_a \hat{\bar{Y}}_h - \frac{{}_a \hat{Y}_s}{{}_a \hat{X}_s} \, {}_a \hat{\bar{X}}_h \right]$$

Thus $\left(\hat{D}_{hi}^y - \hat{\bar{D}}_h^y \right)$ takes the form:

$$\begin{aligned} \left(\hat{D}_{hi}^y - \hat{\bar{D}}_h^y \right) &= \sum_{a=1}^{A} \left({}_a \hat{Y}_{hi} - \frac{{}_a \hat{Y}_s}{{}_a \hat{X}_s} \, {}_a \hat{X}_{hi} \right) - \sum_{a=1}^{A} \left({}_a \hat{\bar{Y}}_h - \frac{{}_a \hat{Y}_s}{{}_a \hat{X}_s} \, {}_a \hat{\bar{X}}_h \right) \\ &= \sum_{a=1}^{A} \left(\left({}_a \hat{Y}_{hi} - {}_a \hat{\bar{Y}}_h \right) - \frac{{}_a \hat{Y}_s}{{}_a \hat{X}_s} \left({}_a \hat{X}_{hi} - {}_a \hat{\bar{X}}_h \right) \right) \end{aligned} \tag{2.12}$$

Substituting (2.12) (and its corresponding expression for z) in (2.10) we obtain the expression for the covariance of two estimators of totals:

$$\operatorname{cov}\left(\hat{Y}_s^R, \hat{Z}_s^R \right) = \sum_{h \in s} \frac{n_h}{(n_h - 1)} \sum_{i=1}^{n_h} \left(\hat{D}_{hi}^y - \hat{\bar{D}}_h^y \right) \left(\hat{D}_{hi}^z - \hat{\bar{D}}_h^z \right) \tag{2.13}$$

The estimator of the covariance of two proportions like the one in (2.8) is given by:

$$\operatorname{cov}\left(\hat{P}_s^{R,y}, \hat{P}_s^{R,z} \right) = \frac{1}{X_s^2} \operatorname{cov} ar \left(\hat{Y}_s^R, \hat{Z}_s^R \right)$$

in order to estimate the covariance of two estimators of ratios like (2.9) we use in (2.11):

$$\begin{aligned} {}_a \hat{e}_{hijv} &= \left({}_a y_{hijv} - {}_a \hat{R}_s^y \cdot {}_a x_{hijv} \right) - \hat{R}_s^R \left({}_a z_{hijv} - {}_a \hat{R}_s^z \cdot {}_a x_{hijv} \right) \\ &= \left({}_a y_{hijv} - \hat{R}_s^R \cdot {}_a z_{hijv} \right) \end{aligned}$$

It is important to point out that the LFS database does not contain the weights w'_{he} but the weights after sex-age adjustment ${}_a w''_{he}$ for each person in the sample.

If we use ${}_a w''_{he}$ in (2.11), and noting that:

$$\frac{{}_a\hat{Y}_h^R}{{}_a\hat{X}_h^R} = \frac{\sum_i\sum_j\sum_v w'_{he}\cdot\frac{{}_aX_h}{{}_a\hat{X}_h}\cdot y_{haijv}}{\sum_i\sum_j\sum_v w'_{he}\cdot\frac{{}_aX_h}{{}_a\hat{X}_h}\cdot x_{haijv}} = \frac{\sum_i\sum_j\sum_v w'_{he}\cdot y_{haijv}}{\sum_i\sum_j\sum_v w'_{he}\cdot x_{haijv}} = \frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h}$$

we obtain:

$$\begin{aligned}\hat{D}_{hi}^y &= \sum_{a=1}^{A}\sum_{j=1}^{b}\sum_{v=1}^{m_{hij}} {}_a w''_{he}\left({}_a y_{hijv} - \frac{{}_a\hat{Y}_s^R}{{}_a\hat{X}_s^R}\,{}_a x_{hijv}\right)\\ &= \sum_{a=1}^{A}\sum_{j=1}^{b}\sum_{v=1}^{m_{hij}} w'_{he}\cdot\frac{{}_aX_s}{{}_a\hat{X}_s}\left({}_a y_{hijv} - \frac{{}_a\hat{Y}_s^R}{{}_a\hat{X}_s^R}\,{}_a x_{hijv}\right)\\ &= \sum_{a=1}^{A}\frac{{}_aX_s}{{}_a\hat{X}_s}\left[{}_a\hat{Y}_{hi} - \frac{{}_a\hat{Y}_s}{{}_a\hat{X}_s}\,{}_a\hat{X}_{hi}\right]\end{aligned} \qquad (2.14)$$

And $\hat{\bar{D}}_h^y$ in this case is:

$$\hat{\bar{D}}_h^y = \sum_{a=1}^{A}\frac{{}_aX_s}{{}_a\hat{X}_s}\left({}_a\hat{\bar{Y}}_h - \frac{{}_a\hat{Y}_s}{{}_a\hat{X}_s}\,{}_a\hat{\bar{X}}_h\right)$$

Thus $\left(\hat{D}_{hi}^y - \hat{\bar{D}}_h^y\right)$ takes the following form:

$$\begin{aligned}\left(\hat{D}_{hi}^y - \hat{\bar{D}}_h^y\right) &= \sum_{a=1}^{A}\frac{{}_aX_s}{{}_a\hat{X}_s}\left(\left({}_a\hat{Y}_{hi} - {}_a\hat{\bar{Y}}_h\right) - \frac{{}_a\hat{Y}_s}{{}_a\hat{X}_s}\left({}_a\hat{X}_{hi} - {}_a\hat{\bar{X}}_h\right)\right)\\ &= \left(\hat{Y}_{hi}^R - \hat{\bar{Y}}_h^R\right) - \sum_{a=1}^{A}\frac{{}_a\hat{Y}_s^R}{{}_a\hat{X}_s}\left({}_a\hat{X}_{hi} - {}_a\hat{\bar{X}}_h\right)\end{aligned} \qquad (2.15)$$

Note that (2.12) and (2.15) differ just in the term ${}_aX_s / {}_a\hat{X}_s$. Our experience is that this term tends to vary closely around one. Thus, we can expect that the use of ${}_a w''_{he}$ in (2.11) should work as an approximation to (2.14). However, if we substitute (2.15) in (2.10) to obtain the estimator of the variance of $\hat{Y}_s^R$ we obtain:

$$\begin{aligned}\operatorname{cov}\left(\hat{Y}_s^R,\hat{Y}_s^R\right) &= \sum_{h\in s}\frac{n_h}{(n_h-1)}\sum_{i=1}^{n_h}\left[\left(\hat{Y}_{hi}^R-\hat{\bar{Y}}_h^R\right)-\sum_{a=1}^{A}\frac{{}_a\hat{Y}_s^R}{{}_a\hat{X}_s}\left({}_a\hat{X}_{hi}-{}_a\hat{\bar{X}}_h\right)\right]^2 \\ &= \sum_{h\in s}\frac{n_h}{(n_h-1)}\sum_{i=1}^{n_h}\left[\left(\hat{Y}_{hi}^R-\hat{\bar{Y}}_h^R\right)^2\right. \\ &\quad -\left(\hat{Y}_{hi}^R-\hat{\bar{Y}}_h^R\right)\sum_{a=1}^{A}\frac{{}_a\hat{Y}_s^R}{{}_a\hat{X}_s}\left({}_a\hat{X}_{hi}-{}_a\hat{\bar{X}}_h\right) \\ &\quad \left.+\left(\sum_{a=1}^{A}\frac{{}_a\hat{Y}_s^R}{{}_a\hat{X}_s}\left({}_a\hat{X}_{hi}-{}_a\hat{\bar{X}}_h\right)\right)^2\right]\end{aligned} \tag{2.16}$$

Assuming a positive correlation between Y and X, the second (negative) term within the squared brackets will be larger than the third one. Thus, taking into account just the first term of (2.16) yields a conservative estimator of the variance, that is:

$$\operatorname{var}\left(\hat{Y}_s^R\right)\cong\sum_{h\in s}\frac{n_h}{(n_h-1)}\sum_{i=1}^{n_h}\left(\hat{Y}_{hi}^R-\hat{\bar{Y}}_{hi}^R\right)^2 \tag{2.17}$$

$$\operatorname{cov}\left(\hat{Y}_s^R,\hat{Z}_s^R\right)\cong\sum_{h\in s}\frac{n_h}{(n_h-1)}\sum_{i=1}^{n_h}\left(\hat{Y}_{hi}^R-\hat{\bar{Y}}_{hi}^R\right)\left(\hat{Z}_{hi}^R-\hat{\bar{Z}}_{hi}^R\right) \tag{2.18}$$

The OCEI uses this approximation, collapsing pairs of consecutive PSUs. That is, pairs of consecutive PSUs are considered as being selected from a single stratum. Let $\hat{Y}_{hl1}^R$ and $\hat{Y}_{hl2}^R$ denote the estimates for the two PSUs collapsed into the *l*th "assumed" stratum, in the *h*th selection stratum. We have then that (2.18) reduces to:

$$\operatorname{cov}\left(\hat{Y}_s^R,\hat{Z}_s^R\right)\cong\sum_{h\in s}\sum_{l=1}^{n_h/2}\left(\hat{Y}_{hl1}^R-\hat{Y}_{hl2}^R\right)\left(\hat{Z}_{hl1}^R-\hat{Z}_{hl2}^R\right) \tag{2.19}$$

CHAPTER 3

LOG-LINEAR MODELLING FOR SPREE ESTIMATION

In sections 1.3 and 1.4 we discussed the necessity for improved estimates for labour force rates for sex-age groups, at state levels in Venezuela. We also noted the lack of auxiliary information available with the minimum requirements to be used for statistics estimation techniques. This led us to seek estimators that require, as auxiliary information, only censuses or their population projections. Although this applies to many countries in Latin America (see section 1.2), we refer only to Venezuelan states in this work.

We begin this chapter defining the necessary notation as well as introducing the Pseudo-Likelihood technique which will be used later in this chapter. We then explore the SPREE method and its relationship to log-linear and multinomial logistic models used to obtain synthetic estimators for small areas represented as cross-tabulations. Here we shall show the theoretical details establishing the equivalence between SPREE and the model-based approaches. The practical procedure to implement those approaches, including variance and covariance estimations, as well as its formal theoretical developments will be addressed in the next chapter.

Throughout this chapter, we illustrate each development using examples from the LFS. For multinomial logistic models, we use the variables (sex, age, state) defining the sub-populations as the independent variables or effects in the model to obtain

smoothed estimates of the proportion of people in each of the three categories (Employed, Unemployed and Non-active) needed for the calculation of the rates within each sub-population.

3.1. NOTATION AND THE PL APPROACH

3.1.1. Notation

We now define some basic notation that will be used from now on in this thesis. This basic notation is defined with the purpose of simplifying later explanations. Additional notation will be defined elsewhere in this document when required.
Let us consider each sub-population as the cross-classification of people over 14 years old by sex, age groups and states. Let the subscript *i=1,2* denote the *i*th sex category, *j=1,2,3,4* denote the *j*th age group and *k=1,2,...,23* denote the *k*th state (see section 1.4). Thus, we have $C = 2 \cdot 4 \cdot 23 = 184$ cells or sub-groups. Similarly, let the subscript *q=1,2,3* denote the groups Employed, Unemployed and Non-Actives respectively. In some cases and for the sake of simplicity, we will refer to the subgroups ijk ordered lexicographically as $s = 1, \ldots, S$.

Regarding counts, proportions and rates, let us define:

M_{ijkq} = The "true" finite population count of the number of people over 14 years old in the *i*th sex category, *j*th age group, *k*th state, *q*th labour force group, for any specified time. We use the term "true count" to indicate the "assumed" true count, i.e. the Census counts for a census year. For any off-census year, we consider this count as "unknown".

$M_{ijk\bullet} = \sum_{q=1}^{3} M_{ijkq}$ = The true finite population count of the number of people over 14 years old in the *i*th sex category, *j*th age group, *k*th state, for any specified time. Likewise, any other subscript that is replaced with a point (•) will indicate that the figure is a total for that subscript. For example, $M_{\bullet jk\bullet}$ will

indicate $\sum_{i=1}^{2}\sum_{q=1}^{3} M_{ijkq}$. Here we also use the term "true count" to indicate the "assumed" true count. In this case, the assumed true counts will still be the Census counts for a census year, but now, providing that the labour force groups (q) are aggregated, the "Census population projections" will be the "assumed" true count for any off-census year; otherwise the count will be regarded as "unknown".

$P_{q/ijk}$ = M_{ijkq}/M_{ijk} = True finite population proportion of the number of people over 14 years old in the qth labour force group, belonging to the ith sex category, jth age group, kth state for any specified time. For the term "true", the same conditions as the ones described for M_{ijkq} apply.

${}_{q}R_{ijk}$ = True rate associated with the qth group for the ith sex category, jth age group, kth state. That is, ${}_{e}R_{ijk}$, ${}_{u}R_{ijk}$ and ${}_{n}R_{ijk}$ would be the Employment Rate, Unemployment Rate and Non-activity Rate for the ith sex category, jth age group, kth state, as they were defined in section 1.4. Additionally, we define ${}_{A}R_{ijk}$ as the Activity Rate (section 1.4). For the term "true", the same conditions as the ones described for M_{ijkq} apply.

When we use the small letters m and r instead of the capital letters M, and R used above, we will refer to sample figures. For instance, m_{ijkq} will denote the sample count of people over 14 years old in the ith sex category, jth age group, kth state, qth labour force group. To denote sample proportions we will use p instead of π .

Likewise, when we use the symbols "^" and "~" over any of these letters (including the symbol π), we will refer to a design-based estimator and to a model-based estimator of the letter or symbol in use. For instance, $\hat{M}_{ijkq}$ and $\tilde{M}_{ijkq}$ will refer to the design based estimator and to the model based estimator of the true count of people over 14 years old in the ith sex category, jth age group, kth state, qth labour force group, for a specified time.

3.1.2. Pseudo-Maximum Likelihood (PL)

We now give a brief description of the Pseudo-Maximum Likelihood method, which we shall use as the key tool to link the "standard" analytic statistics procedures to the case of complex samples.

A common standard approach when estimating parameters for regression models is the use of Maximum Likelihood Estimation (MLE). Let y_i $(i=1,...,n)$ be n independent variables with known probability density function $f_i(y_i;\boldsymbol{\theta})$. Suppose we have observed one realisation for each variable y_i. The joint distribution of y_i $(i=1,...,n)$ is given by $f_1(y_1;\boldsymbol{\theta})\cdot f_2(y_2;\boldsymbol{\theta})\cdots f_n(y_n;\boldsymbol{\theta})$ which can be regarded as a function of $\boldsymbol{\theta}$ and is called the Likelihood Function:

$$l(\boldsymbol{\theta})=\prod_{i=1}^{n} f_i(y_i;\boldsymbol{\theta})$$

The MLE of $\boldsymbol{\theta}$ is given by the vector $\tilde{\boldsymbol{\theta}}$ that maximises the Likelihood Function, which is the same as maximising the logarithm of the Likelihood Function $L(\boldsymbol{\theta})=Log\left(l(\boldsymbol{\theta})\right)$.

The vector $\tilde{\boldsymbol{\theta}}$ can often be obtained by solving for $\boldsymbol{\theta}$ the set of Likelihood Equations:

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}=\sum_{i=1}^{n}\frac{\partial Log\left(f_i(y_i;\boldsymbol{\theta})\right)}{\partial \boldsymbol{\theta}}=\sum_{i=1}^{n}U_i(\boldsymbol{\theta})=\mathbf{0} \qquad (3.1)$$

We have then that the set of Maximum Likelihood Statistics $\tilde{\boldsymbol{\theta}}$ is a function of the variables y_i, that is, $\tilde{\boldsymbol{\theta}}=\mathbf{g}(y_1,y_2,...,y_n)$.

When taking a sample using Simple Random Sampling (SRS) the assumption of independence is usually made and the vector of parameters $\boldsymbol{\theta}$ is obtained by MLE. When the sampling design is complex, the distribution functions $f_i(y_i;\boldsymbol{\theta})$ are

affected, being now the conditional distributions of the population given the sampling design. If we want to apply MLE with complex samples we have to define first the structure of those conditional distributions to be used in the likelihood function. This process may be highly complicated requiring the modelling of the relation between y_i and the design variables (Skinner 1989).

Pseudo-Maximum Likelihood Estimation (PL) is an approach to the estimation of the vector of parameters $\boldsymbol{\theta}$ avoiding the complexity of defining the conditional distribution of y_i. Suppose the sample y_i , $i=(1,..,n)$ has been taken from a finite population Ω using a complex sampling design. Suppose Ω consist of a realisation of N independent random variables Y_i , $i=(1,..,N)$ with distribution function $f_i(Y_i;\boldsymbol{\theta})$ known.

Suppose for the moment that we have observed the whole finite population Ω, that is, we have a census. Using this information in (3.1) and solving for $\boldsymbol{\theta}$, we would get a MLE of the vector of parameter $\boldsymbol{\theta}$ based on the finite population Ω; that is, the MLE for $\boldsymbol{\theta}$ would be a vector of population parameters $\boldsymbol{\theta}^{\Omega}$ defined by the values in Ω. We shall call $\boldsymbol{\theta}^{\Omega}$ the census vector (Binder 1983), which is usually the target parameter in many common analysis where the estimation of $\boldsymbol{\theta}$ can be seen as a necessary step to estimate $\boldsymbol{\theta}^{\Omega}$ itself.

We have now that $n=N$ in the equation system (3.1) and thus $\sum_{i=1}^{N}\mathbf{U}_i(\boldsymbol{\theta})=\mathbf{A}(\boldsymbol{\theta})$ is a vector of finite population totals that are function of $\boldsymbol{\theta}$.

In practice we do not know the vector of finite population totals $\mathbf{A}(\boldsymbol{\theta})$. The PL approach consists of replacing the vector of totals $\sum_{i=1}^{N}\mathbf{U}_i(\boldsymbol{\theta})=\mathbf{A}(\boldsymbol{\theta})$ in (3.1) by a design consistent estimator $\hat{\mathbf{A}}(\boldsymbol{\theta})$ and solving the equation system to obtain the pseudo-likelihood estimate $\tilde{\boldsymbol{\theta}}_{PL}$ which therefore estimates the census vector $\boldsymbol{\theta}^{\Omega}$, $\tilde{\boldsymbol{\theta}}_{PL}=\tilde{\boldsymbol{\theta}}^{\Omega}$.

Here again, $\tilde{\boldsymbol{\theta}}_{PL}$ is a function of the variables y_i, that is, $\tilde{\boldsymbol{\theta}}_{PL} = \mathbf{g}(y_1, y_2, ..., y_n)$ and the structure and properties of its variance-covariance matrix estimator will depend on the variance of that function under the complex sampling design.

The estimator $\tilde{\boldsymbol{\theta}}_{PL}$ might coincide in some cases with the MLE $\tilde{\boldsymbol{\theta}}$, for instance, if we use a self-weighted sampling design while using the Horwitz-Thompson estimator. However, it is clear that $\tilde{\boldsymbol{\theta}}_{PL}$ is not unique since it depends on the structure of the estimator $\hat{\mathbf{A}}(\boldsymbol{\theta})$ used.

3.2. THE SPREE METHOD

As it has already been discussed, the method of Structure Preserving Estimation (SPREE) for categorical variables (Purcell and Kish 1980) offers a possible answer to the situation of main concern in this document, i.e. to obtain Labour Force estimates making use of auxiliary information derived from population censuses.

3.2.1. General description of the SPREE Method

The SPREE method consists in obtaining a cross-tabulation with estimated counts for the required period (often the present moment) using as a starting point a cross-tabulation of the same dimension whose internal structure (marginals and counts) is believed to be highly correlated with the structure of the required cross-tabulation. The cross-tabulation used as starting point is often –but not necessarily- a table with the same variables as the one required by the researcher but for a previous reference period, i.e. a previous Census.

The basic idea is, knowing that a cross-tabulation is fully defined by its internal association or interaction structure, to "update" some elements of that structure in the starting table by using "present" reliable information related to some of the margin of the cross-tabulation. We use the term "present" here to denote the period of time for

which estimations are required. Potential sources for present information are, for instance, large sample surveys, census updates and administrative registers. By using the Iterative Proportional Fitting (IPF) algorithm (see Deming and Stephan 1940, Purcell and Kish 1980, Agresti 1990, Chambers 1999), the chosen marginals in the starting table are forced to agree with the present marginals. This process updates the structural elements associated with the chosen marginals whilst leaving the remaining elements unchanged. The table counts resulting from this process are the estimated counts for the required period.

The IPF algorithm consists of two basic steps that are repeated until a convergence criterion is achieved. Let us suppose we want to obtain updated count estimates for an x-dimensional cross-tabulation using as "initial" or "starting" point the same but outdated table whose counts we shall denote by $\mathbf{M}_c^o$, (c=1,...,C) where c denote the lexicographic order of the table cells. Let us also suppose that we have current or updated reliable information about R marginals of that cross-tabulation and let $\mathbf{M}_\bullet^{curr} = (M_1^{curr},...,M_r^{curr},...,M_R^{curr})'$ be a R-vector containing those reliable marginal counts. The IPF algorithm proceeds as follows:

a) Adjust the initial counts $\mathbf{M}_c^o$ by an appropriate scaling factor to make them agree with one of the "present" marginals. Let us denote the adjusted counts by $\mathbf{M}_c^A$.

b) Go back to step a), using the adjusted counts as initial values i.e. $\mathbf{M}_c^o = \mathbf{M}_c^A$, and using a different "present" marginals from $\mathbf{M}_\bullet^{curr}$ that has not been used yet. If all the present marginals chosen for the process have already been used, start a new cycle using the one used in first place.

New cycles are carried through until the following convergence criterion is attained: let $\mathbf{M}_\bullet^{At} = (M_1^{At},...,M_r^{At},...,M_R^{At})'$ be the vector of the R adjusted marginals involved in the process, after the steps a) and b) have been completed for the t-th time. Now, let $\mathbf{\Delta A}^t$ be the vector whose elements ΔA_r^t are the absolute differences between the updated marginals and the adjusted marginals after the steps a) and b) have been

completed for the t-th time i.e. $\Delta\mathbf{A}_r^t = \left| M_r^{At} - M_r^{curr} \right|$. Convergence is achieved at the t-th iteration when:

$$\Delta\mathbf{A}_r^t < \delta \quad ; \quad \forall r \, , \delta \cong 0 \tag{3.2}$$

That is, all the differences between the adjusted marginals and the present marginals chosen for the process have to be sufficiently close to zero.

We have then that the resulting adjusted counts from this process M_c^A i.e. after convergence has been reached, are the SPREE estimates or updated count estimates we were looking for. Therefore, using the notation defined in this section, we can define the SPREE estimates of a cross-tabulation as the set of counts M_r^A satisfying the set of equation:

$$\mathbf{M}_\bullet^A = \mathbf{M}_\bullet^{curr} \tag{3.3}$$

where $\mathbf{M}_\bullet^A$ is the vector of the R adjusted marginals involved in the process, after convergence has been attained.

3.2.2. SPREE and the Labour Force case

In our specific case, as it was explained in the previous section, the sub-population groups are a cross-classification of the people over 14 years old by sex, age group and state. The total counts and marginals for that cross-classification or cross-tabulation can be obtained for the Census year. In this section we will refer to Census year counts by adding apostrophe to the notation already defined i.e. M'_{ijkq}.

We do not have reliable information about the current counts M_{ijkq}. However, the census population projections provide us with information about those marginals

related to the demographics variables (sex, age, and state), that is the sub-groups $M_{ijk\bullet}$ and consequently $M_{i\bullet\bullet}$, $M_{\bullet j\bullet}$, $M_{\bullet\bullet k}$, $M_{ij\bullet}$, $M_{i\bullet k}$ and $M_{\bullet jk}$. From the LFS we can select those estimates of the present marginals that we consider "reliable" enough to be included into the SPREE process.

Example

For illustration purposes and to bring theory into LF context, we will refer throughout this chapter to the LF situation using the notation defined in section 3.1.1.

Let us suppose that we have reliable LFS estimates for the marginals $M_{\bullet\bullet kq}$, $M_{ij\bullet q}$ and accordingly $M_{i\bullet\bullet q}$ and $M_{\bullet j\bullet q}$, i.e. $\hat{M}_{ij\bullet q}$, $\hat{M}_{i\bullet\bullet q}$ and $\hat{M}_{\bullet j\bullet q}$. Let us also consider the census demography projection at the state-sex-age group level as the true marginals $M_{ijk\bullet}$ value. It is important to recall that, although the conceptual definition for the variable characterising an individual as employee, unemployed or non-active is exactly the same for both the 1990 Census and the LFS, there are practical issues suggesting that a prior adjustment of the Census data might be necessary (refer to section 2.3). It is also worth noticing that, due to the post-stratification estimation method used by the LFS, the LFS estimations counts $\hat{M}_{ijk\bullet}$ equal the census demographic projection counts $M_{ijk\bullet}$ -the LFS estimation process forces the estimates of $M_{ijk\bullet}$ to be equal to the census projections-. Therefore, we shall use $M_{ijk\bullet}$ and $\hat{M}_{ijk\bullet}$ indistinctively to denote those demography projection counts hence, although it is clear that $\hat{M}_{ijk\bullet}$ will not be object of sampling error.

Starting from the census cross-tabulation, we can use the iterative proportional fitting (IPF) algorithm to make the census marginal counts $M'_{ijk\bullet}$, $M'_{\bullet\bullet kq}$ and $M'_{ij\bullet q}$ agree with the census sub-population projection $\hat{M}_{ijk\bullet}$ and with the reliable LFS estimates of the current counts $M_{\bullet\bullet kq}$ and $M_{ij\bullet q}$, that is, $\hat{M}_{\bullet\bullet kq}$ and $\hat{M}_{ij\bullet q}$. The algorithm to obtain the SPREE estimates is as follows:

a) Adjust the initial counts $M^{'}_{ijkq}$ by the scaling factor $M_{ijk\bullet}/M^{'}_{ijk\bullet}$ to make them agree with the "present" marginal estimate $\hat{M}_{ijk\bullet}$. Let us denote the adjusted counts by $M^{''}_{ijkq}$.

b) Adjust the counts $M^{''}_{ijkq}$ by the scaling factor $M_{ij\bullet q}/M^{''}_{ij\bullet q}$ to make them agree with the "present" marginal estimate $\hat{M}_{ij\bullet q}$. Denote these adjusted counts by $M^{'''}_{ijkq}$.

c) Adjust the counts $M^{'''}_{ijkq}$ by the scaling factor $M_{\bullet\bullet kq}/M^{'''}_{\bullet\bullet kq}$ to make them agree with the "present" marginal estimate $\hat{M}_{\bullet\bullet kq}$. We denote these adjusted counts by $M^{''''}_{ijkq}$.

d) Go back to step a), using the counts $M^{''''}_{ijkq}$ as initial values.

These steps go on until convergence is achieved. The resulting cross-tabulation will thus contain the SPREE-estimated counts $\tilde{M}^{SPREE}_{ijkq}$ with the SPREE-estimated marginal $\tilde{M}_{i\bullet kq}$ and $\tilde{M}_{\bullet jkq}$ for the present time period. The marginals $\tilde{M}_{ijk\bullet}$, $\tilde{M}_{ij\bullet q}$ and $\tilde{M}_{\bullet\bullet kq}$ will all agree with the corresponding "present" marginal estimates $\hat{M}_{ijk\bullet}$, $\hat{M}_{ij\bullet q}$ and $\hat{M}_{\bullet\bullet kq}$, that is,

$$\begin{aligned} \tilde{M}_{ijk\bullet} &= \hat{M}_{ijk\bullet} \\ \tilde{M}_{ij\bullet q} &= \hat{M}_{ij\bullet q} \\ \tilde{M}_{\bullet\bullet kq} &= \hat{M}_{\bullet\bullet kq} \end{aligned} \tag{3.4}$$

Note that this procedure has left unchanged the initial structural (interaction) terms of the Census table that are not related to the marginals we have forced to agree with "present" count estimates, i.e. it has left unchanged the I-K-Q, J-K-Q and I-J-Q structural terms. For instance, when we adjust the $M^{'}_{ijkq}$ to make them agree with the "present" marginal $M_{ijk\bullet}$ - step a) in the IPF algorithm example above-:

$$M^{''}_{ijkq} = M^{'}_{ijkq}\frac{M_{ijk\bullet}}{M^{'}_{ijk\bullet}} \tag{3.5}$$

we are using the same scaling factor $M_{ijk\bullet}/M^{'}_{ijk\bullet}$ at each level of q, thus all the cross-ratios but those related to q are being changed.

To illustrate this let us write the updated cross-ratio I-J for i=1,2 and j=1,2 after the adjustment given by (3.5),

$$CR(I_{1,2}-J_{1,2})=\frac{M''_{11kq}M''_{22kq}}{M''_{21kq}M''_{12kq}}=\frac{M'_{11kq}\dfrac{M_{11k\bullet}}{M'_{11k\bullet}}M'_{22kq}\dfrac{M_{22k\bullet}}{M'_{22k\bullet}}}{M'_{21kq}\dfrac{M_{21k\bullet}}{M'_{21k\bullet}}M'_{12kq}\dfrac{M_{12k\bullet}}{M'_{12k\bullet}}}\neq\frac{M'_{11kq}M'_{22kq}}{M'_{21kq}M'_{12kq}} \tag{3.6}$$

$CR(I_{1,2}-J_{1,2})$ has obviously changed from the original one. Working the same cross-ratio for any combination of levels for I-J, I-K, J-K and I-J-K, we can easily see that those structural terms have changed. Let us now write any cross-ratio involving Q, say I-J-Q for i=1,2, j=1,2 and q=1,2:

$$\begin{aligned}CR(I_{1,2}-J_{1,2}-Q_{1,2})&=\frac{M''_{11k1}M''_{22k1}}{M''_{21k1}M''_{12k1}}\cdot\frac{M''_{21k2}M''_{12k2}}{M''_{11k2}M''_{22k2}}\\&=\frac{M'_{11k1}\dfrac{M_{11k\bullet}}{M'_{11k\bullet}}M'_{22k1}\dfrac{M_{22k\bullet}}{M'_{22k\bullet}}}{M'_{21k1}\dfrac{M_{21k\bullet}}{M'_{21k\bullet}}M'_{12k1}\dfrac{M_{12k\bullet}}{M'_{12k\bullet}}}\cdot\frac{M'_{21k2}\dfrac{M_{21k\bullet}}{M'_{21k\bullet}}M'_{12k2}\dfrac{M_{12k\bullet}}{M'_{12k\bullet}}}{M'_{11k2}\dfrac{M_{11k\bullet}}{M'_{11k\bullet}}M'_{22k2}\dfrac{M_{22k\bullet}}{M'_{22k\bullet}}}\\&\equiv\frac{M'_{11k1}M'_{22k1}}{M'_{21k1}M'_{12k1}}\cdot\frac{M'_{21k2}M'_{12k2}}{M'_{11k2}M'_{22k2}}\end{aligned} \tag{3.7}$$

This cross-ratio remains exactly the same as the initial census table. Again, working the cross-ratio for any combination of levels for I-Q, J-Q, K-Q, I-J-Q, I-K-Q, J-K-Q and I-J-K-Q, we can see that those structural terms have not changed.

When we adjust the other two marginals $\tilde{M}_{\bullet\bullet kq}=\hat{M}_{\bullet\bullet kq}$ and $\tilde{M}_{ij\bullet q}=\hat{M}_{ij\bullet q}$, we then change the structural terms in the original table associated to those marginals, i.e. K-Q, I-Q, J-Q, I-J-Q. Consequently, the only structural terms that are not changed in our specific case are I-K-Q, J-K-Q and I-J-K-Q.

On the whole, we have that when applying the SPREE method we always end up with some marginals matching "present" marginals and with some structural interaction

terms changed and the others preserved. What term changes and what remains unchanged will depend on what marginals we adjust to current reliable information.

3.3. THE SPREE METHOD AND LOG-LINEAR MODELS

3.3.1. Log-linear Representation of Cross-tabulations

We have seen in the previous section that the SPREE method involves three cross-tabulations, i.e. the target unknown cross-tabulation, the reference or "starting point" cross-tabulation and the "estimated" cross-tabulation, as well as some reliable information related to some of the marginals of the target cross-tabulation.

Nelder (1974) showed that cross-tabulations counts can be expressed or modelled as log-linear saturated models. Therefore, the three cross-tabulation involved in the SPREE method can be expressed as log-linear saturated models. In our case, the reference cross-tabulation is the Census table and its log-linear representation is,

$$Log\left(M'_{ijkq}\right) = \lambda'_0 + \lambda'_i + \lambda'_j + \lambda'_k + \lambda'_q + \lambda'_{ij} + \lambda'_{ik} + \lambda'_{iq} + \lambda'_{jk} + \lambda'_{jq} + \lambda'_{kq} + \lambda'_{ijk} + \lambda'_{ijq} + \lambda'_{ikq} + \lambda'_{jkq} + \lambda'_{ijkq} \tag{3.8}$$

This is, a log-linear model containing all interactions up to the highest interaction order (saturated). Each parameter λ' in (3.8) represents the structural terms associated with its sub-indices. For instance, the parameters λ'_{ij} correspond to the structural terms given by (3.6), related to IJ when KQ are kept constants. Likewise the parameter λ'_{ijq} correspond to the structural terms given by (3.7), related to IJQ when K is kept constant. The constant term in (3.8) is given by λ'_0.

In the same way, we can express the current but unknown cross-tabulation counts by:

$$Log\left(M_{ijkq}\right) = \lambda_0 + \lambda_i + \lambda_j + \lambda_k + \lambda_q + \lambda_{ij} + \lambda_{ik} + \lambda_{iq} + \lambda_{jk} + \lambda_{jq} + \lambda_{kq} + \lambda_{ijk} + \lambda_{ijq} + \lambda_{ikq} + \lambda_{jkq} + \lambda_{ijkq} \tag{3.9}$$

The cross-tabulation resulting from the SPREE method or "estimated" table can also be expressed as a log-linear model. We know that some of the structural terms in the SPREE-estimated table remain unaltered from the reference table. We also know that the remaining structural terms are "updated" or estimated. Therefore, we can expect that the log-linear model resembling this SPREE-estimated table will contain a mixture of original reference parameters and "updated" parameters.

As we shall prove later in this Chapter, in the LFS situation this estimated table is represented by a log-linear model with the following general structure,

$$Log\left(\tilde{M}_{ijkq}\right) = \mathbf{X}^{A}_{ijkq}\tilde{\boldsymbol{\lambda}} + \mathbf{X}^{B}_{ijkq}\boldsymbol{\lambda}' \tag{3.10}$$

where $\mathbf{X}^{A}_{ijkq}$ and $\mathbf{X}^{B}_{ijkq}$ are the zero-ones rows defining the updated and the unchanged terms, respectively, related to cell *ijkq*. Note that $\mathbf{X}^{A}_{ijkq}$ and $\mathbf{X}^{B}_{ijkq}$ are just the rows of two matrices, $\mathbf{X}^{A}$ and $\mathbf{X}^{B}$ representing a partition of the well known model matrix for log-linear models, i.e. $\mathbf{X} = [\mathbf{X}^{A} : \mathbf{X}^{B}]$. The counts $\tilde{M}_{ijkq}$ are the SPREE-estimated present counts.

In the example formulated in section 4.2.2 where the structural terms I-K-Q, J-K-Q and I-J-K-Q are preserved by the SPREE process, the explicit structure of the log-linear model (3.10) representing the SPREE-estimated table would be,

$$Log\left(\tilde{M}_{ijkq}\right) = \tilde{\lambda}_{0} + \tilde{\lambda}_{i} + \tilde{\lambda}_{j} + \tilde{\lambda}_{k} + \tilde{\lambda}_{q} + \tilde{\lambda}_{ij} + \tilde{\lambda}_{ik} + \tilde{\lambda}_{iq} + \tilde{\lambda}_{jk} + \tilde{\lambda}_{jq} + \tilde{\lambda}_{kq} + \tilde{\lambda}_{ijk} + \tilde{\lambda}_{ijq} + \lambda'_{ikq} + \lambda'_{jkq} + \lambda'_{ijkq} \tag{3.11}$$

We shall prove in the following sections that the "updated" vector of parameters $\tilde{\boldsymbol{\lambda}}$ in (3.10) are the Pseudo-Likelihood estimates for the corresponding target vector of parameter $\boldsymbol{\lambda}$ when these are constrained to equal the remaining "unchanged" parameters $\boldsymbol{\lambda}'$. Consequently, the estimates generated by the SPREE method $\tilde{M}_{ijkq}$ are in fact Pseudo-Likelihood estimates from a constrained saturated log-linear model.

This fact will enable us to develop expressions for SPREE-estimates and their respective variances based on Generalized Linear Model theory.

3.3.2. Log-linear Models: A Brief Description

Consider a set of sample counts m_c from a given cross-tabulation with $c = 1,...,C$ denoting the lexicographic order for the C cells in the table. Let us suppose that m_c follow an independent Poisson distribution,

$$f(m_c;\mu_c) = \frac{\mu_c^{m_c} \cdot e^{-\mu_c}}{m_c!}$$

with expected frequency $\mu_c = e^{\mathbf{X}_c\boldsymbol{\lambda}}$, that is,

$$Log(\mu_c) = \mathbf{X}_c\boldsymbol{\lambda} \tag{3.12}$$

where $\mathbf{X}_c$ is the c*th* row of the zero-one *CxP* model matrix $\mathbf{X}$ whose rows define the effects and/or interaction terms related to each count and $\boldsymbol{\lambda}$ is a P-vector of parameters, so that $\mu_c = \mu_c(\boldsymbol{\lambda}) = f(\boldsymbol{\lambda})$. That is, the nonlinear expression of the model for μ_c, $\mu_c = e^{\mathbf{X}_c\boldsymbol{\lambda}}$, is "linearized" and modelled as in (3.12) – a log-linear model. This is a special case of Generalized Linear Model (GLM) theory (McCullagh and Nelder 1983, Dobson 1990).

Many techniques have been developed to estimate the vector of parameters for models like (3.12); some of them are Weighted Least Squares (Grizzle et. al. 1969, Agresti 1990), Minimum Chi-Squared (Neyman 1949, Bhapkar 1966, Agresti 1990), Minimum Discrimination Information (Kullback 1959, Berkson 1972, Simon 1973, Gokhale et.al. 1978), Kernel Smoothing (Aitchison et.al 1976, Agresti 1990) and Penalized Likelihood (Good et.al 1971, Simonoff 1983, Titterington et.al. 1985).

We shall use here Maximum Likelihood Estimation (MLE) to deal with log-linear models like (3.12) (McCullagh and Nelder 1983, Agresti 1990, Dobson 1990). MLE consists in finding the vector $\tilde{\boldsymbol{\lambda}}$ that maximizes the likelihood function,

$$\begin{aligned} Log\left(L\left(\boldsymbol{\mu}(\boldsymbol{\lambda})\right)\right) &= Log\left(\prod_{c=1}^{C}\frac{\mu_c^{m_c}(\boldsymbol{\lambda})\cdot e^{-\mu_c(\boldsymbol{\lambda})}}{m_c!}\right) \\ &= \sum_{c=1}^{C} Log\left(\frac{\mu_c^{m_c}(\boldsymbol{\lambda})\cdot e^{-\mu_c(\boldsymbol{\lambda})}}{m_c!}\right) \\ &= \sum_{c=1}^{C} m_c Log(\mu_c(\boldsymbol{\lambda})) - \sum_{c=1}^{C}\mu_c(\boldsymbol{\lambda}) - \sum_{c=1}^{C} Log(m_c!) \\ &= \sum_{c=1}^{C} m_c \mathbf{X}_c\boldsymbol{\lambda} - \sum_{c=1}^{C} e^{\mathbf{X}_c\boldsymbol{\lambda}} - \text{constant} \end{aligned}$$

The likelihood equations are obtained by setting to zero the derivatives of the Log-likelihood function with respect to the unknown vector of parameters $\boldsymbol{\lambda}$:

$$\left(\frac{\partial Log(L)}{\partial\boldsymbol{\lambda}}\right) = \mathbf{0}$$

where $\left(\frac{\partial Log(L)}{\partial\boldsymbol{\lambda}}\right)$ is a P-vector whose p*th* element is given by:

$$\begin{aligned} \left(\frac{\partial Log(L)}{\partial\lambda_p}\right) &= \sum_{c=1}^{C} m_c x_{c,p} - \sum_{c=1}^{C} x_{c,p} e^{\mathbf{X}_c\boldsymbol{\lambda}} \\ &= \sum_{c=1}^{C} x_{c,p}\left(m_c - \mu_c(\boldsymbol{\lambda})\right) \end{aligned} \qquad (3.13)$$

so the p*th* likelihood equation is given by:

$$\sum_{c=1}^{C} x_{c,p} m_c = \sum_{c=1}^{C} x_{c,p}\mu_c(\boldsymbol{\lambda})$$

Let $\mathbf{m}$ be the C-vector of samples counts m_c and let $\boldsymbol{\mu}(\boldsymbol{\lambda})$ be the C-vector of expected frequencies $\mu_c(\boldsymbol{\lambda})$. The set of likelihood equations is given in matrix notation as:

$$\mathbf{X}^t\mathbf{m} = \mathbf{X}^t\boldsymbol{\mu}(\boldsymbol{\lambda}) \tag{3.14}$$

Birch (1963) showed that the likelihood equations for log-linear models are defined by equating the minimal sufficient statistics to their expected values. He also showed that for a log-linear model there exists just one set of counts m_i that both satisfies the model and makes the minimal sufficient statistics equal to their observed values. Solving (3.14) for $\boldsymbol{\lambda}$ we get the maximum likelihood estimates $\tilde{\boldsymbol{\lambda}}$ and consequently, $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}(\tilde{\boldsymbol{\lambda}})$,

$$\mathbf{X}^t\mathbf{m} = \mathbf{X}^t\boldsymbol{\mu}(\tilde{\boldsymbol{\lambda}}) \tag{3.15}$$

Note that the minimal sufficient statistics depend on the structure of the model. The zeroes and ones in each row of $\mathbf{X}^t$ define the cells for which the parameter related to that row plays any role. Therefore, $\mathbf{X}^t\mathbf{m}$ is a vector whose elements are the sums of cells defining specific marginals that correspond to parameters in the model.

3.3.3. Log-linear "Census" Models

Let us suppose now that our sample vector $\mathbf{m}$ is a vector of finite population counts $\mathbf{M} = \{M_c\}$ with $\sum_{c=1}^{C} M_c = M$. Let us suppose that the finite population, denote by Ω^M, is a sample from a super-population with M_c $c = 1,..,C$ following an independent Poisson distribution,

$$f(M_c;\mu_c) = \frac{\mu_c^{M_c}\cdot e^{-\mu_c}}{M_c!} \tag{3.16}$$

with $\mu_c = \mu_c(\boldsymbol{\lambda}) = e^{\mathbf{X}_c\boldsymbol{\lambda}}$. It is clear then that the results developed in the last section fully apply to this case. The likelihood equations given in (3.14) are now,

$$\mathbf{X}^t\mathbf{M} = \mathbf{X}^t\boldsymbol{\mu}(\boldsymbol{\lambda}) \tag{3.17}$$

Note that solving (3.17) for $\boldsymbol{\lambda}$ we actually get the vector of parameters for the log-linear model corresponding to the set of finite population counts M_c; we shall denote such a vector of parameter by $\boldsymbol{\lambda}^{\Omega^M}$; that is,

$$Log(M_c) = \mathbf{X}_c\boldsymbol{\lambda}^{\Omega^M} \tag{3.18}$$

This vector $\boldsymbol{\lambda}^{\Omega^M}$ therefore acts as the MLE for the super-population vector of parameters $\boldsymbol{\lambda}$. That is, $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}^{\Omega^M}$,

$$\mathbf{X}^t\mathbf{M} = \mathbf{X}^t\boldsymbol{\mu}(\boldsymbol{\lambda}^{\Omega^M}) = \mathbf{X}^t\boldsymbol{\mu}(\tilde{\boldsymbol{\lambda}}) \tag{3.19}$$

Example.

In our Labour Force case, assuming that the population counts M_{ijkq} follow an independent Poisson distribution[1], the counts $M_{ijk\bullet}$, $M_{\bullet\bullet kq}$ and $M_{ij\bullet q}$ would be the minimal sufficient statistics for an unsaturated log-linear model $Log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\lambda}$ containing all the two-level interaction terms and the *ijk* as well as the *ijq* three-level interaction terms (see Agresti 1990 pg.166), that is:

$$Log(\mu_{ijkq}) = \lambda_0 + \lambda_i + \lambda_j + \lambda_k + \lambda_q + \lambda_{ij} + \lambda_{ik} + \lambda_{iq} + \lambda_{jk} + \lambda_{jq} + \lambda_{kq} + \lambda_{ijk} + \lambda_{ijq} \tag{3.20}$$

Therefore, following Birch (1963) results, the equations:

$$\begin{aligned} M_{ijk\bullet} &= \mu_{ijk\bullet}\left(\tilde{\boldsymbol{\lambda}}\right) \\ M_{\bullet\bullet kq} &= \mu_{\bullet\bullet kq}\left(\tilde{\boldsymbol{\lambda}}\right) \\ M_{ij\bullet q} &= \mu_{ij\bullet q}\left(\tilde{\boldsymbol{\lambda}}\right) \end{aligned} \tag{3.21}$$

are the likelihood equations for the model (3.20) and the vector $\tilde{\boldsymbol{\lambda}}$ and the set of counts $\tilde{\mu}_{ijkq} = \tilde{\mu}_{ijkq}(\boldsymbol{\lambda}) = e^{\mathbf{X}'\tilde{\boldsymbol{\lambda}}}$ satisfying these equations are the Maximum Likelihood Estimates for $\boldsymbol{\lambda}$ and μ_{ijkq},

$$Log\left(\tilde{\mu}_{ijkq}\right) = \tilde{\lambda}_0 + \tilde{\lambda}_i + \tilde{\lambda}_j + \tilde{\lambda}_k + \tilde{\lambda}_q + \tilde{\lambda}_{ij} + \tilde{\lambda}_{ik} + \tilde{\lambda}_{iq} + \tilde{\lambda}_{jk} + \tilde{\lambda}_{jq} + \tilde{\lambda}_{kq} + \tilde{\lambda}_{ijk} + \tilde{\lambda}_{ijq} \tag{3.22}$$

3.3.4. Constrained Log-linear Estimation

We can also be interested in fitting a model where some of the higher interaction terms are believed to be known. In this case we want the fitting process to be conditioned or constrained by them.

Let $\boldsymbol{\lambda}^U$ be the P_u-vector of unknown parameters and let $\boldsymbol{\lambda}^K$ be the P_k-vector of known parameters, with $P_u + P_k = P$. Let also $\mathbf{X}^U$ and $\mathbf{X}^K$ be the zero-one model matrices whose rows define, respectively, the unknown and known effects and/or interaction terms related to each count. We have then that the log-linear model we are interested in can be expressed as:

$$Log\left(\boldsymbol{\mu}\right) = \mathbf{X}^U\boldsymbol{\lambda}^U + \mathbf{X}^K\boldsymbol{\lambda}^K \tag{3.23}$$

We can obtain the likelihood equations for a log-linear model like (3.23) in the same way as we obtained those in (3.19). Suppose again that the counts M_c follow an

[1] Some considerations about this assumption will be discussed later in Chapter 5

independent Poisson distribution as in (3.16). However, we have now that $\mu_c = e^{\mathbf{X}_c^U \boldsymbol{\lambda}^U + \mathbf{X}_c^K \boldsymbol{\lambda}^K}$, that is,

$$Log(\mu_c) = \mathbf{X}_c^U \boldsymbol{\lambda}^U + \mathbf{X}_c^K \boldsymbol{\lambda}^K \tag{3.24}$$

where $\mathbf{X}_c^U$ and $\mathbf{X}_c^K$ are the c*th* row of the zero-ones model matrix $\mathbf{X}^U$ and $\mathbf{X}^K$ respectively, so that $\mu_c = \mu_c(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K) = f(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K)$. The Log-likelihood function is:

$$\begin{aligned}
Log\left(L\left(\boldsymbol{\mu}(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K)\right)\right) &= Log\left(\prod_{c=1}^{C} \frac{\mu_c^{M_c}(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K) \cdot e^{-\mu_c(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K)}}{M_c!}\right) \\
&= \sum_{c=1}^{C} Log\left(\frac{\mu_c^{M_c}(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K) \cdot e^{-\mu_c(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K)}}{M_c!}\right) \\
&= \sum_{c=1}^{C} M_c Log(\mu_c(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K)) - \sum_{c=1}^{C} \mu_c(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K) - \sum_{c=1}^{C} Log(M_c!) \\
&= \sum_{c=1}^{C} M_c \left[\mathbf{X}_c^U \boldsymbol{\lambda}^U + \mathbf{X}_c^K \boldsymbol{\lambda}^K\right] - \sum_{c=1}^{C} e^{\mathbf{X}_c^U \boldsymbol{\lambda}^U + \mathbf{X}_c^K \boldsymbol{\lambda}^K} - \text{constant}
\end{aligned}$$

Noting that $\boldsymbol{\lambda}^K$ is a vector of known fixed constants, the likelihood equations for a log-linear model for $\boldsymbol{\mu}$ are obtained by setting to zero the derivatives of the Log-likelihood function with respect to the unknown vector of parameters $\boldsymbol{\lambda}^U$:

$$\left(\frac{\partial Log(L)}{\partial \boldsymbol{\lambda}^U}\right) = \mathbf{0}$$

where $\left(\frac{\partial Log(L)}{\partial \boldsymbol{\lambda}^U}\right)$ is a P_u-vector whose p*th* element is given by:

$$\begin{aligned}
\left(\frac{\partial Log(L)}{\partial \lambda_p^U}\right) &= \sum_{c=1}^{C} M_c x_{c,p}^U - \sum_{c=1}^{C} x_{c,p}^U e^{\mathbf{X}_c^U \boldsymbol{\lambda}^U + \mathbf{X}_c^K \boldsymbol{\lambda}^K} \\
&= \sum_{c=1}^{C} x_{c,p}^U \left(M_c - \mu_c(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K)\right)
\end{aligned} \tag{3.25}$$

so the p*th* likelihood equation is given by:

$$\sum_{c=1}^{C} x_{c,p}^{U} M_c = \sum_{c=1}^{C} x_{c,p}^{U} \mu_c(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K)$$

We have then that, in matrix notation, the set of likelihood equation is:

$$\mathbf{X}^{U'}\mathbf{M} = \mathbf{X}^{U'}\boldsymbol{\mu}(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K) \tag{3.26}$$

Solving (3.26) for $\boldsymbol{\lambda}^U$ we get the maximum likelihood estimates for $\boldsymbol{\mu}$, $\boldsymbol{\mu}(\tilde{\boldsymbol{\lambda}}^U, \boldsymbol{\lambda}^K) = \tilde{\boldsymbol{\mu}}$:

$$\mathbf{X}^{U'}\mathbf{M} = \mathbf{X}^{U'}\boldsymbol{\mu}(\tilde{\boldsymbol{\lambda}}^U, \boldsymbol{\lambda}^K) \tag{3.27}$$

As the cells taken into account for a particular equation in (3.27) depends only on $\mathbf{X}^U$, it is clear that the minimal sufficient statistics in this case are the same as those of a reduced log-linear model $Log(\boldsymbol{\mu}) = \mathbf{X}^U\boldsymbol{\lambda}^U$.

However, the ML estimation of $\boldsymbol{\lambda}^U$ and $\boldsymbol{\mu}$ in (3.27) will differ from that of the reduced model; this is due to the fact that, following Birch (1963) results expanded by Haberman (1973, 1974), the MLE of parameters and counts for a log-linear model have to satisfy not only the minimal sufficient statistics but also the structure of the model. This structure is given by the right-hand side of equation (3.24). Therefore, the MLE $\tilde{\boldsymbol{\lambda}}^U$ and consequently $\tilde{\boldsymbol{\mu}}$ have to take into account $\boldsymbol{\lambda}^K$, i.e. estimation is constrained by $\boldsymbol{\lambda}^K$. In brief, the ML estimates will consist of those values $\tilde{\boldsymbol{\lambda}}^U$ and $\tilde{\mathbf{M}}$ satisfying (3.27), that is, satisfying the minimal sufficient statistics as well as the model.

Note that working with a model like (3.12), that is, $Log(\boldsymbol{\mu}) = \mathbf{X}\,\boldsymbol{\lambda}$, can be considered as working with a particular case of the general model (3.23),

$Log(\boldsymbol{\mu}) = \mathbf{X}^U\boldsymbol{\lambda}^U + \mathbf{X}^K\boldsymbol{\lambda}^K$; in model (3.12) we do not make the assumption of knowing the value of any of the parameters in the structure of the model, i.e. all the parameters in the model belong to $\boldsymbol{\lambda}^U$, so $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\mathbf{U}}$.

Example

Let us suppose we are interested in the saturated model,

$$Log(\mu_{ijkq}) = \mu + \lambda_i + \lambda_j + \lambda_k + \lambda_q + \lambda_{ij} + \lambda_{ik} + \lambda_{iq} + \lambda_{jk} + \lambda_{jq} + \lambda_{kq} + \lambda_{ijk} + \lambda_{ijq} + \lambda_{ikq} + \lambda_{jkq} + \lambda_{ijkq}$$

and we believe we know the true values for $\lambda_{ikq}, \lambda_{jkq}$ and λ_{ijkq}. Therefore, following the notation defined above, we can write this saturated model as:

$$Log(\mu_{ijkq}) = \mathbf{X}^U_{ijkq}\boldsymbol{\lambda}^U + \mathbf{X}^K_{ijkq}\boldsymbol{\lambda}^K$$

$$Log(\mu_{ijkq}) = \lambda_0^U + \lambda_i^U + \lambda_j^U + \lambda_k^U + \lambda_q^U + \lambda_{ij}^U + \lambda_{ik}^U + \lambda_{iq}^U + \lambda_{jk}^U + \lambda_{jq}^U + \lambda_{kq}^U + \lambda_{ijk}^U + \lambda_{ijq}^U + \lambda_{ikq}^K + \lambda_{jkq}^K + \lambda_{ijkq}^K \quad (3.28)$$

In this example, the matrix $\mathbf{X}^U$ has the same structure as the matrix $\mathbf{X}$ in the model (3.20) defined for the previous example. Therefore, the general structure of the likelihood equations given by (3.27) generates, in this example, the same set of likelihood equations (3.21). However, this set of likelihood equations has now to be solved for $\boldsymbol{\lambda}^U$ conditioning on the vector of parameters $\boldsymbol{\lambda}^K$,

$$\begin{aligned} M_{ijk\cdot} &= \mu_{ijk\cdot}\left(\tilde{\boldsymbol{\lambda}}^U, \boldsymbol{\lambda}^K\right) \\ M_{\cdot\cdot kq} &= \mu_{\cdot\cdot kq}\left(\tilde{\boldsymbol{\lambda}}^U, \boldsymbol{\lambda}^K\right) \\ M_{ij\cdot q} &= \mu_{ij\cdot q}\left(\tilde{\boldsymbol{\lambda}}^U, \boldsymbol{\lambda}^K\right) \end{aligned} \quad (3.29)$$

That is, $\tilde{\boldsymbol{\lambda}}^U$ is the vector of ML estimates of $\boldsymbol{\lambda}^U$ constraint on $\boldsymbol{\lambda}^K$.

We finally have:

$$Log\left(\tilde{\mu}_{ijkq}\right) = \mathbf{X}^{U}_{ijkq}\tilde{\boldsymbol{\lambda}}^{U} + \mathbf{X}^{K}_{ijkq}\boldsymbol{\lambda}^{K}$$

$$Log\left(\tilde{\mu}_{ijkq}\right) = \tilde{\mu} + \tilde{\lambda}^{U}_{i} + \tilde{\lambda}^{U}_{j} + \tilde{\lambda}^{U}_{k} + \tilde{\lambda}^{U}_{q} + \tilde{\lambda}^{U}_{ij} + \tilde{\lambda}^{U}_{ik} + \tilde{\lambda}^{U}_{iq} + \tilde{\lambda}^{U}_{jk} + \tilde{\lambda}^{U}_{jq} + \tilde{\lambda}^{U}_{kq} + \tilde{\lambda}^{U}_{ijk} + \tilde{\lambda}^{U}_{ijq} + \\ + \lambda^{K}_{ikq} + \lambda^{K}_{jkq} + \lambda^{K}_{ijkq} \qquad (3.30)$$

with $\tilde{\mu}_{ijkq} = e^{\mathbf{X}^{U}_{ijkq}\tilde{\boldsymbol{\lambda}}^{U} + \mathbf{X}^{K}_{ijkq}\boldsymbol{\lambda}^{K}}$.

3.3.5. Pseudo-Likelihood Estimation for Log-linear Models

So far in this chapter, we have been considering the super-population parameters $\mu_c(\boldsymbol{\lambda})$ as our target for estimation. We have also used the finite population counts M_c and their model parameters $\boldsymbol{\lambda}^{\Omega^M}$ as the ML estimates for $\mu_c(\boldsymbol{\lambda})$. In practice, however, we do not know the value of the finite population counts M_c. In fact, these counts are often the real target of the analysis, as it is in our case.

Yet, it might be feasible to get reliable estimates of the marginal counts needed to solve the likelihood equations given above. Information from different sources like Administrative Registers might be available for the period of interest. If this information is available and we are willing to accept it as the "true" values, we can then apply the theory discussed in the previous sections and get the vector of parameters $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}^{\Omega^M}$ and the ML estimates $\mu(\tilde{\boldsymbol{\lambda}})$.

When no information is available from administrative registers or the same is considered of poor quality, another alternative is the use of survey estimates. Suppose we have survey estimates $\hat{\mathbf{M}}_c$ for the counts in the cross-tabulation and, although they are considered of poor precision, the aggregated marginals involved in the likelihood equations are considered reliable. We can then use the PL approach explained in

section 3.1.2 replacing finite population quantities by corresponding survey estimates in the likelihood equations (3.26) to obtain the PL estimates for the vector of parameters $\boldsymbol{\lambda}$ and the expected frequencies $\boldsymbol{\mu}(\boldsymbol{\lambda})$,

$$\mathbf{X}^t\hat{\mathbf{M}} = \mathbf{X}^t\boldsymbol{\mu}(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K) \tag{3.31}$$

Note that all we have done here is to use a vector of sampling estimates $\hat{\mathbf{M}}$ instead of $\mathbf{M}$ in the likelihood equations. In other words, we are replacing the vector of totals

$$\mathbf{X}^t\left[\mathbf{M} - \boldsymbol{\mu}(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K)\right] = \mathbf{A}\left(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K\right) = \mathbf{0} \tag{3.32}$$

by a consistent estimator,

$$\mathbf{X}^t\left[\hat{\mathbf{M}} - \boldsymbol{\mu}(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K)\right] = \hat{\mathbf{A}}\left(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K\right) = \mathbf{0} \tag{3.33}$$

Therefore, we can say that the vector of parameter estimates $\tilde{\boldsymbol{\lambda}}^U$, resulting from solving (3.31) for $\boldsymbol{\lambda}^U$, is the vector of PL estimates $\tilde{\boldsymbol{\lambda}}^U_{PL}$ generating the set of PL estimates $\tilde{\boldsymbol{\mu}}^{PL}$, with elements $\tilde{\mu}_c^{PL}$,

$$Log\left(\tilde{\mu}_c^{PL}\right) = \mathbf{X}_c^U\tilde{\boldsymbol{\lambda}}^U_{PL} + \mathbf{X}_c^K\boldsymbol{\lambda}^K \tag{3.34}$$

Note that using direct survey estimates $\hat{M}_c$ in the likelihood equations, we are in fact producing model-based estimates $\tilde{M}_c$ for the finite population counts to estimate the super-population parameters $\mu_c(\boldsymbol{\lambda})$. Provide the model holds, the model-based estimates for the finite population counts $\tilde{M}_c$ should be of a higher accuracy than the direct survey estimates; this is due to the fact that $\tilde{M}_c$ are based on the estimation of a fewer number of parameters and therefore their level of precision is higher. However, as the model departs from reality the bias component in the estimates increases. Therefore, the gains in precision when estimating the finite population

counts through a specific model like in (3.34) have to offset the increases in bias in order to justify its use instead of direct survey estimates.

Example

We can usually obtain population projections for the counts $M_{ijk\bullet}$. For counts involving Labour Force classification such as $M_{\bullet\bullet kq}$ and $M_{ij\bullet q}$, some countries produce estimates based on administrative registers such as registers of people claming benefits or registers of job seekers. If this is the case and we are willing to accept those figures as true values, we proceed as discussed in previous sections getting the vector of parameters $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}^{\Omega^M}$ and the ML estimates $\boldsymbol{\mu}(\tilde{\boldsymbol{\lambda}})$. However, we recall that this is not the case we are facing in this work. We are working under a different scenario where no extra auxiliary information other than that coming from censuses is available.

However, we do have survey information from the LFS that can be used in this process. Therefore, in our example in section 4.3.3 where we have the likelihood equations given by (3.29), we can rely on the population projections for the counts $M_{ijk\bullet}$ and on the LFS estimate $\hat{M}_{\bullet\bullet kq}$ and $\hat{M}_{ij\bullet q}$ in order to obtain the PL estimates $\tilde{M}^{PL}_{ijkq}$. In this case, the pseudo-likelihood equations would be:

$$\begin{aligned}
\hat{M}_{ijk\bullet} &= M_{ijk\bullet}(\tilde{\boldsymbol{\lambda}}^U_{PL}, \boldsymbol{\lambda}^K) \\
\hat{M}_{\bullet\bullet kq} &= M_{\bullet\bullet kq}(\tilde{\boldsymbol{\lambda}}^U_{PL}, \boldsymbol{\lambda}^K) \\
\hat{M}_{ij\bullet q} &= M_{ij\bullet q}(\tilde{\boldsymbol{\lambda}}^U_{PL}, \boldsymbol{\lambda}^K)
\end{aligned} \tag{3.35}$$

Finally, the model estimates for the finite population counts are,

$$\tilde{M}^{PL}_{ijkq} = e^{\mathbf{X}_{ijkq}\tilde{\boldsymbol{\lambda}}^U_{PL} + \mathbf{X}_{ijkq}\boldsymbol{\lambda}^K} \tag{3.36}$$

and $\tilde{M}^{PL}_{ijkq} = \tilde{\mu}^{PL}_{ijkq}$.

3.3.6. SPREE Estimates: Log-linear Model Pseudo-Likelihood Estimates

We have seen that the minimal sufficient statistics implied by (3.31) will depend on the structure of the model and:

a) Those minimal sufficient statistics will always consist of those marginals for which we assume to have reliable current estimates. They are also related to the highest interaction terms in $\boldsymbol{\lambda}^U$ for each factor in the model.
b) The likelihood equations given by (3.31) force the marginals of the to-be-estimated table of counts to match the minimal sufficient statistics.
c) Solving (3.31) for $\boldsymbol{\lambda}^U$ given $\boldsymbol{\lambda}^K$, gives us PL estimates $\tilde{\boldsymbol{\lambda}}^U_{PL}$ and thereby PL count estimates $\tilde{\mathbf{M}}_{PL}$.

We now focus our attention on the particular case of a *saturated* model of the form of (3.24). Let us consider the following two notes.

In point a), note that the marginals involved in the minimal sufficient statistics are the same marginals belonging to the vector $\mathbf{M}^{curr}_{\bullet}$ in the SPREE method procedure explained in section 3.2.1.

In point b), note that the likelihood equations given by (3.31) will always have the same structure as those equations implied by (3.3) i.e. the set of equations the adjusted counts $M^A_{ab,\ldots,n}$ have to satisfy for those counts to be the SPREE estimates (see section 3.2.1).

Therefore, taking into account these comments, point c) also implies the adjusted counts $M^A_{ab,\ldots,n}$ resulting from the SPREE process have to be the PL count estimates $\tilde{\mathbf{M}}_{PL}$ obtained using the estimated Log-linear model,

$$Log\left(\tilde{\mathbf{M}}_{PL}\right) = \mathbf{X}^U \tilde{\boldsymbol{\lambda}}^U_{PL} + \mathbf{X}^K \boldsymbol{\lambda}^K \tag{3.37}$$

where $\tilde{\lambda}^{U}_{PL}$ represent the "changed" or "updated" terms $\tilde{\lambda}$ and λ^{K} the "unchanged" terms $\tilde{\lambda}^{'}$ in the resulting SPREE model as denoted in (3.10).

Example

Let us return to the SPREE algorithm discussion. In the example in section 3.2.2, we discussed the case where a table from a previous census was to be updated using the LFS estimates $\hat{M}_{ijk\bullet}$, $\hat{M}_{\bullet\bullet kq}$ and $\hat{M}_{ij\bullet q}$. We saw in that example that the IPF algorithm forces the following equalities to be true:

$$\begin{aligned}\tilde{M}_{ijk\bullet} &= \hat{M}_{ijk\bullet}\\ \tilde{M}_{ij\bullet q} &= \hat{M}_{ij\bullet q}\\ \tilde{M}_{\bullet\bullet kq} &= \hat{M}_{\bullet\bullet kq}\end{aligned} \tag{3.38}$$

which has exactly the same structure as (3.35) for the PL-estimates under the model,

$$\begin{aligned}Log\left(\mu_{ijkq}\right) = \lambda_0^U + \lambda_i^U + \lambda_j^U + \lambda_k^U + \lambda_q^U + \lambda_{ij}^U + \lambda_{ik}^U + \lambda_{iq}^U + \lambda_{jk}^U + \lambda_{jq}^U + \lambda_{kq}^U + \lambda_{ijk}^U + \lambda_{ijq}^U + \\ + \lambda_{ikq}^K + \lambda_{jkq}^K + \lambda_{ijkq}^K\end{aligned} \tag{3.39}$$

We know from section 3.3.1 that the set of counts produced by the SPREE method in that example have the structural terms I-K-Q, J-K-Q and I-J-K-Q preserved from the census table. We also know from section 3.3.1 that these structural terms are related to the interaction terms in the log-linear representation of the census cross-tabulation that we denoted as $\lambda_{ikq}^{'}$, $\lambda_{jkq}^{'}$ and $\lambda_{ijkq}^{'}$. Therefore, if we set

$$\begin{aligned}\lambda_{ikq}^{K} &= \lambda_{ikq}^{'}\\ \lambda_{jkq}^{K} &= \lambda_{jkq}^{'}\\ \lambda_{ijkq}^{K} &= \lambda_{ijkq}^{'}\end{aligned}$$

and solve the likelihood equations (3.35) for λ^{U}, we just have to invoke the result of Birch (1963) where he showed that for a log-linear model there exists just one set of

counts that both satisfies the model and ensures the minimal sufficient statistics are equal to the observed values, to prove that we will get from the constrained log-likelihood PL-estimation the same set of estimated counts as those we get from the SPREE process,

$$\tilde{M}_{ijkq}^{SPREE} = M_{ijkq}(\tilde{\boldsymbol{\lambda}}_{PL}^{U}, \boldsymbol{\lambda}^{K}) = \mu_{ijkq}(\tilde{\boldsymbol{\lambda}}_{PL}^{U}, \boldsymbol{\lambda}^{K}) = e^{\mathbf{X}_{ijkq}^{U}\tilde{\boldsymbol{\lambda}}_{PL}^{U} + \mathbf{X}_{ijkq}^{K}\boldsymbol{\lambda}^{K}}$$

3.4. UNSATURATED SPREE

Suppose we conclude that some of the higher order interaction structures do not play any role in the overall structure of our cross-classification; yet, we feel that the higher interaction structures among the remaining terms are worth preserving. We can still use the SPREE algorithm in this case to get the target count estimates; however, this requires us to first make a suitable modification to the "reference" or "starting" table.

This modified table has to be one having: a) the same structural terms as those to be either updated or preserved from the original table and b) the remaining structural terms not to be either updated or preserved set to one. We get to this modified table by building first a table from the marginals related to those effects to be either updated or preserved, and then expanding that table to the original dimension of the starting cross-tabulation ensuring the remaining structural effects are kept to one.

This modified table can be represented as a non-saturated log-linear model where those parameters equivalent to those not to be either preserved or updated are not present, i.e. they equal zero. Here, note that structural terms (odd ratios) equalizing one in the cross-tabulation translate into parameters equalizing zero in the log-linear representation of such cross-tabulation. As this modified table is the one used as the starting table, this approach is called "unsaturated SPREE".

The rest of the process would be exactly as it has been explained in this chapter. We apply the IPF algorithm to the modified starting table and end up with a table of count

estimates with the preserved structural terms from the original table and the updated structural terms as planned and the remaining structural terms equal to one.

However, building that modified table might be cumbersome. Instead, we can use Log-linear models to do this. Let $\boldsymbol{\lambda}^U$ again be the set of parameters considered as "unknown" and $\boldsymbol{\lambda}^K$ the set of parameters considered as "known" in a log-linear model for the expected frequencies matrix of dimension *C*, that is,

$$Log(\boldsymbol{\mu}) = \mathbf{X}^U\boldsymbol{\lambda}^U + \mathbf{X}^K\boldsymbol{\lambda}^K \tag{3.40}$$

We recall that $\boldsymbol{\lambda}^U$ and $\boldsymbol{\lambda}^K$ represent the interaction structures to be updated and to be changed respectively in the SPREE procedure. In section 3.3.4 we showed that the general set of likelihood equations for this kind of models is given by:

$$\mathbf{X}^t\mathbf{M} = \mathbf{X}^t\boldsymbol{\mu}(\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^K) \tag{3.41}$$

However, note that we are now working under the assumption that some of the highest order interaction terms do not play any role in the structure of the cross-tabulation. Therefore, our model (3.40) is now a non-saturated log-linear model with $\boldsymbol{\lambda}^U$ and $\boldsymbol{\lambda}^K$ being of dimension P_u and P_k respectively with $(P_u + P_k) < P$.

We obtain the PL estimates for the target counts as it was explained in section 3.3.5, that is, using the set of direct estimates $\hat{\mathbf{M}}$ instead of $\mathbf{M}$ in (3.41), solving for $\boldsymbol{\lambda}^U$, so that we get the vector $\tilde{\boldsymbol{\lambda}}^U_{PL}$,

$$\mathbf{X}^{U^t}\hat{\mathbf{M}} = \mathbf{X}^{U^t}\boldsymbol{\mu}(\tilde{\boldsymbol{\lambda}}^U_{PL}, \boldsymbol{\lambda}^K) \tag{3.42}$$

and finally using the equation:

$$\tilde{\mathbf{M}}_{PL} = e^{\mathbf{X}^U\tilde{\boldsymbol{\lambda}}^U_{PL} + \mathbf{X}^K\boldsymbol{\lambda}^K} \tag{3.43}$$

Again, the cells taken into account for a particular equation in (3.42) depend only on $\mathbf{X}^U$ so the minimal sufficient statistics are the same as those of a reduced log-linear

model consisting only of $\boldsymbol{\lambda}^{U}$. Note in (3.42) that ML estimation of $\boldsymbol{\lambda}^{U}$ is also dependent on the value of $\boldsymbol{\lambda}^{K}$.

Example

In our example, suppose we assume that the highest order structural terms *ijkq* in our cross-tabulation can be considered as not playing any role in the overall structure of the table; i.e. they are equal to one. Let us also suppose that we want to preserve the structural terms *ijq* and *ikq* from the census cross-tabulation updating the remaining terms making use of the corresponding LFS direct estimate marginals.

We will use the log-linear approach to get the current count estimators. The modified census table can be obtained by fitting the unsaturated log-linear model,

$$Log\left(M^{"}_{ijkq}\right) = \mathbf{X}_{ijkq}\boldsymbol{\lambda}^{\Omega"} = \lambda_0^{\Omega"} + \lambda_i^{\Omega"} + \lambda_j^{\Omega"} + \lambda_k^{\Omega"} + \lambda_q^{\Omega"} + \lambda_{ij}^{\Omega"} + \lambda_{ik}^{\Omega"} + \lambda_{iq}^{\Omega"} + \lambda_{jk}^{\Omega"} + \lambda_{jq}^{\Omega"} + \lambda_{kq}^{\Omega"} + \lambda_{ijk}^{\Omega"} + \lambda_{ijq}^{\Omega"} + \lambda_{ikq}^{\Omega"} + \lambda_{jkq}^{\Omega"} \tag{3.44}$$

Therefore, $M^{"PL}_{ijkq} = e^{\mathbf{X}_{ijkq}\tilde{\boldsymbol{\lambda}}^{\cdot}_{PL}}$ are the counts in the modified "starting" table. Note that (3.44) does not contain the parameter λ_{ijkq}.

The current cross-tabulation counts are suppose to be well represented by the following unsaturated log-linear model,

$$Log\left(M_{ijkq}\right) = \mathbf{X}_{ijkq}\boldsymbol{\lambda}^{\Omega} = \lambda_0^{\Omega} + \lambda_i^{\Omega} + \lambda_j^{\Omega} + \lambda_k^{\Omega} + \lambda_q^{\Omega} + \lambda_{ij}^{\Omega} + \lambda_{ik}^{\Omega} + \lambda_{iq}^{\Omega} + \lambda_{jk}^{\Omega} + \lambda_{jq}^{\Omega} + \lambda_{kq}^{\Omega} + \lambda_{ijk}^{\Omega} + \lambda_{ijq}^{\Omega} + \lambda_{ikq}^{\Omega} + \lambda_{jkq}^{\Omega} \tag{3.45}$$

Note that this model, as model (3.44), does not contain the highest interaction term λ_{ijkq}^{Ω}. We now treat the terms related to the interaction structures to be preserved as "known", so that (3.45) can be written as,

$$Log\left(M_{ijkq}\right)=\mathbf{X}_{ijkq}^{U}\boldsymbol{\lambda}^{\Omega^U}+\mathbf{X}_{ijkq}^{K}\boldsymbol{\lambda}^{\Omega^K}=\lambda_0^{\Omega^U}+\lambda_i^{\Omega^U}+\lambda_j^{\Omega^U}+\lambda_k^{\Omega^U}+\lambda_q^{\Omega^U}+\lambda_{ij}^{\Omega^U}+\lambda_{ik}^{\Omega^U}+\lambda_{iq}^{\Omega^U}+$$
$$+\lambda_{jk}^{\Omega^U}+\lambda_{jq}^{\Omega^U}+\lambda_{kq}^{\Omega^U}+\lambda_{ijk}^{\Omega^U}+\lambda_{ijq}^{\Omega^U}+\lambda_{ikq}^{\Omega^K}+\lambda_{jkq}^{\Omega^K} \tag{3.46}$$

We now assume that the current counts M_{ijkq} follow a Poisson distribution with expected frequencies μ_{ijkq} so we can use log-linear PL estimation to estimate the model,

$$Log\left(\mu_{ijkq}\right)=\mathbf{X}_{ijkq}^{U}\boldsymbol{\lambda}^{U}+\mathbf{X}_{ijkq}^{K}\boldsymbol{\lambda}^{K}=\lambda_0^{U}+\lambda_i^{U}+\lambda_j^{U}+\lambda_k^{U}+\lambda_q^{U}+\lambda_{ij}^{U}+\lambda_{ik}^{U}+\lambda_{iq}^{U}+\lambda_{jk}^{U}+\lambda_{jq}^{U}+$$
$$+\lambda_{kq}^{U}+\lambda_{ijk}^{U}+\lambda_{ijq}^{U}+\lambda_{ikq}^{K}+\lambda_{jkq}^{K} \tag{3.47}$$

As we have already seen, this procedure will give us PL estimates for the target model (3.46). Therefore, considering $\lambda_{ikq}^{K}=\lambda_{ikq}^{\Omega^*}$ and $\lambda_{jkq}^{K}=\lambda_{jkq}^{\Omega^*}$ and using the LFS direct counts estimates $\hat{M}_{ijkq}$ to solve (3.42) for $\boldsymbol{\lambda}^{U}$, we get the PL parameter estimates $\tilde{\boldsymbol{\lambda}}_{PL}^{U}$. The likelihood equations are the same as in (3.35); however, the resulting $\tilde{\boldsymbol{\lambda}}_{PL}^{U}$ and consequently the resulting PL current count estimates $\tilde{M}_{ijkq}^{PL}$ are different in this case from those we would have obtained in (3.35). This is due to the fact that the likelihood equations are solved here constraining on $\boldsymbol{\lambda}^{K}=(\lambda_{ikq}^{K},\lambda_{jkq}^{K})^t$ whilst in the saturated example they are solved constraining on $\boldsymbol{\lambda}^{K}=(\lambda_{ikq}^{K},\lambda_{jkq}^{K},\lambda_{ijkq}^{K})^t$.

Finally, the set of PL current count estimates are given by the equation:

$$\tilde{\mathbf{M}}_{PL}=e^{\mathbf{X}^{U}\tilde{\boldsymbol{\lambda}}_{PL}^{U}+\mathbf{X}^{K}\boldsymbol{\lambda}^{K}}=e^{\mathbf{X}^{U}\tilde{\boldsymbol{\lambda}}_{PL}^{\Omega^U}+\mathbf{X}^{K}\boldsymbol{\lambda}^{\Omega^*}} \tag{3.48}$$

3.5. LOG-LINEAR MODELLING WITHOUT STRUCTURAL INFORMATION

We could face the situation where no information from reference tables is available. That is, there are not interactions from any reference table that we want to preserve.

In this case we can still make use of log-linear models to get current count estimates although we now have to rely only on direct estimates of current marginals or any other source of information regarding current marginals we assume as reliable. That is, we have to rely on a traditional unsaturated log-linear model,

$$Log\left(\mu_c\right) = \mathbf{X}_c\boldsymbol{\lambda} \tag{3.49}$$

where $\boldsymbol{\lambda} = \boldsymbol{\lambda}^U$ is a P_u-vector with $P_u < P$. The likelihood equations are again given by (3.42) but without conditioning on any interaction term $\boldsymbol{\lambda}^K$,

$$\mathbf{X}^{U'}\hat{\mathbf{M}} = \mathbf{X}^{U'}\boldsymbol{\mu}(\tilde{\boldsymbol{\lambda}}^U_{PL}) \tag{3.50}$$

The set of PL estimates are given by the equation:

$$\tilde{\mathbf{M}}_{PL} = e^{\mathbf{X}^U\tilde{\boldsymbol{\lambda}}^U_{PL}} \tag{3.51}$$

Example

Providing there is a close relationship between the preserved census year structures and the respective current structures, the procedure explained in previous sections should yield significantly better estimates of the count M_{ijkq} than those offered by the direct design estimators.

The problem is that Latin American countries usually have unstable economies that do not guarantee the preservation over time of structures such as the labour force,

especially at small area levels. In the last fifteen years, for instance, Venezuela has experienced important economic changes that have led to dramatic transformations in the dynamics of its labour force. As a simple illustration of this fact, we can mention that the Venezuelan Unemployment Rate was 10.4% in 1990, dropping to 6.4% in 1993 and registering 14.5% in 1999. The Activity Rate dramatically rose from 59.4% in 1990 to 65.7% in 1999. People occupied in either small businesses or as self-employed increased 10.3% (from 42.1% to 52.4%) in just nine years. These indicators suggest that the dynamics governing the behaviour of the labour market have changed considerably since the last census year. This phenomenon can be even more evident at local levels. For instance, the discovery of new oil layers in the south east of the country at the beginning of the nineties motivated important internal changes in some states that started registering significant growth in population. These arguments suggest that preserving structures from the census year as those implied by $\lambda'_{ikq}, \lambda'_{jkq}$ and λ'_{ijkq} in (3.8) may not be a sensible strategy to follow. However, this has to be verified before drawing any definitive conclusion.

Suppose now that there is no Census table that can be used as a reference for the SPREE process. Suppose also we feel that the expected frequencies of the Poisson model generating the current set of counts is sufficiently explained by the following unsaturated log-linear model:

$$Log\left(\mu_{ijkq}\right) = \mathbf{X}\boldsymbol{\lambda} = \mu + \lambda_i + \lambda_j + \lambda_k + \lambda_q + \lambda_{ij} + \lambda_{ik} + \lambda_{iq} + \lambda_{jk} + \lambda_{jq} + \lambda_{kq} + \lambda_{ijk} + \lambda_{ijq} \tag{3.52}$$

That is, the highest order interaction structures *jkq ikq* and *ijkq* in our cross-tabulation can be considered as not playing any role in the overall structure of the table, so we set them to one and consequently to zero in the log-linear model.

The likelihood equations for this model are again the same as in (3.35). The resulting $\tilde{\boldsymbol{\lambda}}_{PL}$ and PL current count estimates $\tilde{M}^{PL}_{ijkq}$ will be different from those shown in the previous examples as the likelihood equations in those cases are solved constraining on "known" parameters $\boldsymbol{\lambda}^K$.

The PL current count estimates are now given by the expression:

$$\tilde{M}_{ijkq}^{PL} = e^{\mathbf{x}_{ijkq}\tilde{\lambda}_{PL}} \tag{3.53}$$

3.6. THE SPREE METHOD AND MULTINOMIAL LOGISTIC MODELS

Let again the set of population counts M_c follow a Poisson distribution $P(\mu_c)$. It is well know that the distribution of M_c given $M = \sum_{c=1}^{C} M_c$ is multinomial (M,π_c), with $\pi_c = \mu_c/\mu$; $\mu = \sum_{c=1}^{C} \mu_c$. If we assume one of the variables defining the cross-tabulation a response variable and we split the set of population counts into the G groups defined by the remaining variables in the table, so that we rename the C counts as $M_{gc'}$; $g = 1,..,G$; $c' = 1,..,C'$, the conditional distribution of $M_{gc'}$ given $M_{g\bullet} = \sum_{c' \in g} M_{c'}$ is a product multinomial $(M_{g\bullet},\pi_{c'/g})$, with $\pi_{c'/g} = \mu_{gc'}/\mu^{g}$; $\mu^{g} = \sum_{c' \in g} \mu_{gc'}$. Note that $C = GxC'$.

Suppose $\pi_{c'/g} = \pi_{c'/g}(\boldsymbol{\beta}) = h(\mathbf{Z}_c, \boldsymbol{\beta})$ is a general response model for $\pi_{c'/g}$, where $\boldsymbol{\beta}$ is a T-vector of "effect" parameters and $\mathbf{Z}_c$ is the cth row of the zero-one CxT model matrix $\mathbf{Z}$ defining the effects and/or interaction terms related to each cell. A product multinomial logistic model for $\pi_{c'/g}$ is given by,

$$Log\left(\frac{\pi_{c'/g}}{\pi_{base/g}}\right) = \mathbf{Z}_{gc'}\boldsymbol{\beta} = \mathbf{Z}_c\boldsymbol{\beta} \tag{3.54}$$

where $\pi_{base/g}$ is any of the C' expected response proportion of group g selected to be used as the base category. We can also express (3.54) for the case in which we assume we know some of the effects and/or interaction terms in $\boldsymbol{\beta}$,

$$Log\left(\frac{\pi_{c'/g}}{\pi_{base/g}}\right) = \mathbf{Z}_{gc'}^{U}\boldsymbol{\beta}^{U} + \mathbf{Z}_{gc'}^{K}\boldsymbol{\beta}^{K} \tag{3.55}$$

The proportion $\pi_{c'/g}$ is given by the expression,

$$\pi_{c'/g} = \frac{e^{\mathbf{Z}^U_{gc'}\boldsymbol{\beta}^U + \mathbf{Z}^K_{gc'}\boldsymbol{\beta}^K}}{\sum_{c'=1}^{C'} e^{\mathbf{Z}^U_{gc'}\boldsymbol{\beta}^U + \mathbf{Z}^K_{gc'}\boldsymbol{\beta}^K}} \tag{3.56}$$

We can easily extend the SPREE method to the case of product multinomial logistic regression noting that,

$$Log\left(\frac{\pi_{c'/g}}{\pi_{base/g}}\right) = Log\left(\frac{\mu_{gc'}/\mu^g}{\mu_{gbase}/\mu^g}\right) = Log\left(\frac{\mu_{gc'}}{\mu_{gbase}}\right)$$

and by properties of logarithms,

$$Log\left(\frac{\mu_{gc'}}{\mu_{gbase}}\right) = Log\left(\mu_{gc'}\right) - Log\left(\mu_{gbase}\right)$$

that is, a log-linear model for $\mu_{gc'} = \mu_c$ minus the same model for another count $\mu_{baseg} = \mu_b$; thus, assuming a general log linear model for μ_c like the one in (3.24), we have that,

$$\begin{aligned} Log\left(\frac{\pi_{c'/g}}{\pi_{base/g}}\right) &= \left(\mathbf{X}^U_{gc'}\boldsymbol{\lambda}^U + \mathbf{X}^K_{gc'}\boldsymbol{\lambda}^K\right) - \left(\mathbf{X}^U_{gbase}\boldsymbol{\lambda}^U + \mathbf{X}^K_{gbase}\boldsymbol{\lambda}^K\right) \\ &= \left(\mathbf{X}^U_{gc'}\boldsymbol{\lambda}^U - \mathbf{X}^U_{gbase}\boldsymbol{\lambda}^U\right) + \left(\mathbf{X}^K_{gc'}\boldsymbol{\lambda}^K - \mathbf{X}^K_{gbase}\boldsymbol{\lambda}^K\right) \\ &= \left(\mathbf{X}^U_{gc'} - \mathbf{X}^U_{gbase}\right)\boldsymbol{\lambda}^U + \left(\mathbf{X}^K_{gc'} - \mathbf{X}^K_{gbase}\right)\boldsymbol{\lambda}^K \end{aligned} \tag{3.57}$$

Therefore, from (3.55) and (3.57) we have that,

$$\left(\mathbf{X}^U_{gc'} - \mathbf{X}^U_{gbase}\right)\boldsymbol{\lambda}^U + \left(\mathbf{X}^K_{gc'} - \mathbf{X}^K_{gbase}\right)\boldsymbol{\lambda}^K = \mathbf{Z}^U_{gc'}\boldsymbol{\beta}^U + \mathbf{Z}^K_{gc'}\boldsymbol{\beta}^K \tag{3.58}$$

As we have seen in this chapter, finding the PL estimates for the log-linear model $Log\left(\mu_{gc'}\right)=\mathbf{X}^{U}_{gc'}\boldsymbol{\lambda}^{U}+\mathbf{X}^{K}_{gc'}\boldsymbol{\lambda}^{K}$ is equivalent to carrying out a SPREE process preserving the structural terms related to $\boldsymbol{\lambda}^{K}$ from the reference table. Therefore, if we substitute the PL estimate $\tilde{\boldsymbol{\lambda}}^{U^{PL}}$ into (3.58) we get the structure of the logit model equivalent to the SPREE process,

$$Log\left(\frac{\tilde{\pi}^{PL}_{c'/g}}{\tilde{\pi}^{PL}_{base/g}}\right)=\left(\mathbf{X}^{U}_{gc'}-\mathbf{X}^{U}_{gbase}\right)\tilde{\boldsymbol{\lambda}}^{U^{PL}}+\left(\mathbf{X}^{K}_{gc'}-\mathbf{X}^{K}_{gbase}\right)\boldsymbol{\lambda}^{K}=\mathbf{Z}^{U}_{gc'}\tilde{\boldsymbol{\beta}}^{U^{PL}}+\mathbf{Z}^{K}_{gc'}\boldsymbol{\beta}^{K} \quad (3.59)$$

Recall that when we fit (3.59) we are working with counts $M_{gc'}$ following a product multinomial distribution $(M^{g},\pi_{c'/g})$ so that we assume $M_{g.}$ fixed. Therefore, the log-linear models $Log\left(\mu_{gc'}\right)$ we are interested in are those containing the interaction terms λ^{U}_{g} as they force, through the likelihood equations, the marginals $\tilde{M}_{g.}$ to agree with the population marginals $M_{g.}$.

The PL estimates for the proportions $\tilde{\pi}^{PL}_{c'/g}$, which are also the PL estimates for the finite population proportions $P_{c'/g}=M_{gc'}/M_{g.}$, are given by,

$$\tilde{\pi}^{PL}_{c'/g}=\frac{e^{\mathbf{Z}^{U}_{gc'}\tilde{\boldsymbol{\beta}}^{U^{PL}}+\mathbf{Z}^{K}_{gc'}\boldsymbol{\beta}^{K}}}{\sum_{c'=1}^{C'}e^{\mathbf{Z}^{U}_{gc'}\tilde{\boldsymbol{\beta}}^{U^{PL}}+\mathbf{Z}^{K}_{gc'}\boldsymbol{\beta}^{K}}} \quad (3.60)$$

We shall illustrate this using our LFS case.

Example

Consider the saturated log-linear model for μ_{ijkq} in (3.39). We have seen that fitting such a model conditioning on $\boldsymbol{\lambda}^{K}$ we get PL estimates $\tilde{\mu}^{PL}_{ijkq}$ equivalent to those

obtained by carrying out the SPREE method using the LFS estimates $\hat{M}_{ijk\cdot}$, $\hat{M}_{\cdot\cdot kq}$ and $\hat{M}_{ij\cdot q}$. Therefore, we can obtain PL estimate for $\pi_{q/ijk}$ and therefore for $P_{q/ijk}$ as,

$$\tilde{\pi}_{q/ijk}^{PL} = \frac{\tilde{\mu}_{ijkq}^{PL}}{\sum_{q=1}^{3} \tilde{\mu}_{ijkq}^{PL}}$$

Another way to get the PL estimate for $\pi_{q/ijk}$ is by fitting the equivalent product multinomial logistic model. The response variable is the category in the labour force which we denoted by *q=1,2,3*. We know that $\pi_{1/ijk} + \pi_{2/ijk} + \pi_{3/ijk} = 1$. We can then write the following product multinomial logistic model for $\pi_{q/ijk}$ using *q=3* as the base category,

$$Log\left(\frac{\pi_{q/ijk}}{\pi_{3/ijk}}\right) = \mathbf{X}_{ijkq}\boldsymbol{\beta} \tag{3.61}$$

where $\boldsymbol{\beta}$ is the *T=(IJK)(Q-1)*-vector of parameters. The proportion $\pi_{q/ijk}$ is given by,

$$\pi_{q/ijk} = \frac{e^{\mathbf{X}_{ijkq}\boldsymbol{\beta}}}{\sum_{q=1}^{3} \mathbf{X}_{ijkq}\boldsymbol{\beta}} \tag{3.62}$$

Noting that $\pi_{q/ijk} = \mu_{ijkq} / \mu_{ijk\cdot}$, we can write (3.61) as follows:

$$Log\left(\frac{\pi_{q/ijk}}{\pi_{3/ijk}}\right) = Log\left(\frac{\mu_{ijkq}}{\mu_{ijk3}}\right) = Log\left(\mu_{ijkq}\right) - Log\left(\mu_{ijk3}\right) =$$

$$= \left(\lambda_0^U + \lambda_i^U + \lambda_j^U + \lambda_k^U + \lambda_q^U + \lambda_{ij}^U + \lambda_{ik}^U + \lambda_{iq}^U + \lambda_{jk}^U + \lambda_{jq}^U + \lambda_{kq}^U + \lambda_{ijk}^U + \lambda_{ijq}^U + \lambda_{ikq}^K + \lambda_{jkq}^K + \lambda_{ijkq}^K\right)$$

$$-\left(\lambda_0^U + \lambda_i^U + \lambda_j^U + \lambda_k^U + \lambda_3^U + \lambda_{ij}^U + \lambda_{ik}^U + \lambda_{i3}^U + \lambda_{jk}^U + \lambda_{j3}^U + \lambda_{k3}^U + \lambda_{ijk}^U + \lambda_{ij3}^U + \lambda_{ik3}^K + \lambda_{jk3}^K + \lambda_{ijk3}^K\right)$$

Rearranging,

$$Log\left(\frac{\pi_{q/ijk}}{\pi_{3/ijk}}\right)=(\lambda_q^U-\lambda_3^U)+(\lambda_{iq}^U-\lambda_{i3}^U)+(\lambda_{jq}^U-\lambda_{j3}^U)+(\lambda_{kq}^U-\lambda_{k3}^U)$$
$$+(\lambda_{ijq}^U-\lambda_{ij3}^U)+(\lambda_{ikq}^K-\lambda_{ik3}^K)+(\lambda_{jkq}^K-\lambda_{jk3}^K)+(\lambda_{ijkq}^K-\lambda_{ijk3}^K) \quad (3.63)$$

so that,

$$Log\left(\frac{\pi_{q/ijk}}{\pi_{3/ijk}}\right)=\alpha_q^U+\beta_{iq}^U+\beta_{jq}^U+\beta_{kq}^U+\beta_{ijq}^U+\beta_{ikq}^K+\beta_{jkq}^K+\beta_{ijkq}^K \quad (3.64)$$

Fitting (3.64) conditioning on $\boldsymbol{\beta}^K$ -this will be explained in the next chapter-, we get the PL estimates,

$$Log\left(\frac{\tilde{\pi}_{q/ijk}^{PL}}{\tilde{\pi}_{3/ijk}^{PL}}\right)=\tilde{\alpha}_q^{U^{PL}}+\tilde{\beta}_{iq}^{U^{PL}}+\tilde{\beta}_{jq}^{U^{PL}}+\tilde{\beta}_{kq}^{U^{PL}}+\tilde{\beta}_{ijq}^{U^{PL}}+\beta_{ikq}^K+\beta_{jkq}^K+\beta_{ijkq}^K \quad (3.65)$$

where $\tilde{\alpha}_q^{U^{PL}}$ is the PL estimate for the constant term for the qth labour force group and $\tilde{\beta}_{iq}^{U^{PL}}$, $\tilde{\beta}_{jq}^{U^{PL}}$, $\tilde{\beta}_{kq}^{U^{PL}}$ and $\tilde{\beta}_{ijq}^{U^{PL}}$ are the corresponding sex, age group and state PL estimates effects and the sex-age PL estimates interactions effects at the different levels of q. The parameters β_{ikq}^K, β_{jkq}^K and β_{ijkq}^K correspond to the sex-state, age-state and sex-age-state interactions effects at the different levels of q for the Census year, which we planned to preserve.

Note that the log-linear model $Log\left(\mu_{ijkq}\right)$ we use here contains the interaction terms λ_{ijk} as necessary terms given the assumption that the $M_{ijk.}$ are fixed.

The PL estimates $\tilde{\pi}^{PL}_{q/ijk}$, and consequently the PL estimates for $P_{q/ijk} = M_{ijkq} / M_{ijk\bullet}$, are finally given by,

$$\tilde{\pi}^{PL}_{q/ijk} = \frac{e^{\tilde{\alpha}^{U^{PL}}_{q} + \tilde{\beta}^{U^{PL}}_{iq} + \tilde{\beta}^{U^{PL}}_{jq} + \tilde{\beta}^{U^{PL}}_{kq} + \tilde{\beta}^{U^{PL}}_{ijq} + \beta^{K}_{ikq} + \beta^{K}_{jkq} + \beta^{K}_{ijkq}}}{\sum_{q=1}^{3} e^{\tilde{\alpha}^{U^{PL}}_{q} + \tilde{\beta}^{U^{PL}}_{iq} + \tilde{\beta}^{U^{PL}}_{jq} + \tilde{\beta}^{U^{PL}}_{kq} + \tilde{\beta}^{U^{PL}}_{ijq} + \beta^{K}_{ikq} + \beta^{K}_{jkq} + \beta^{K}_{ijkq}}} \tag{3.66}$$

CHAPTER 4

PARAMETER AND VARIANCE ESTIMATION FOR THE SPREE METHOD

In the previous chapter, we discussed the link between the SPREE method and both log-linear models and logit models. These methods allow us to produce alternative count estimates based on models when traditional direct estimates are considered unreliable and no auxiliary information other than from past censuses is available.

The application of these methods results in smoothed estimates of the total counts M_{ijkq} by the shrinking the direct estimates toward the average values defined by the direct estimates of marginal counts specified by the terms in the model. These model estimates should be better than the direct design estimates, providing the terms in the model explain well enough the structure of the current cross-tabulation. That is, the variance of these estimators should be lower than the variance of the direct estimators, but that difference must be sufficiently larger than the magnitude of the bias arising from the misspecification of the model for the model estimates to be preferred.

The main advantage of having established the link between the SPREE method and log-linear and logit models is the possibility of using the Generalized Linear Model (GLM) theory to estimate parameters and calculate variances estimators for the SPREE process.

In this chapter we first discuss a new idea for the estimation of SPREE parameters and variances making use of the well known experimental design concept of "exposures". The appeal of this idea lies in its practical convenience as it can be carried out using standard statistical software.

We then complement the theory developed in the previous chapter by describing the theory behind the parameter and variance estimation process. A Rao-Scott (1981) chi-squared approximation is defined in order to assess the goodness of fit of the models proposed as representation of the cross-tabulation. Finally, alternatives for a diagnostics process are discussed in order to obtaining measures that give us some insight into the presence of outlying cells and influential points when these models are to be implemented in practical work.

4.1. SPREE ESTIMATION AND VARIANCE ESTIMATES: PRACTICAL COMPUTATION

The IPF algorithm described in the previous Chapter is the procedure traditionally used to fit SPREE models. However, that algorithm does not allow us to get parameter estimates and variances. We now describe a simpler method that allows us to obtain the target count estimates and their variances without prior knowledge of the model parameters from the census data. This approach can be carried out using standard statistical software.

4.1.1. A new practical approach to SPREE computation: The Exposure-based Method

We know from the previous Chapter that both the census counts M'_c (or any table counts being used as a reference table) and the unknown current year counts M_c can be expressed as saturated log-linear models:

$$Log\left(M_c\right) = \mathbf{X}_c \boldsymbol{\lambda}^{\Omega} \tag{4.1}$$

$$Log\left(M_c^{'}\right) = \mathbf{X}_c \boldsymbol{\lambda}^{\Omega'} \tag{4.2}$$

where $\boldsymbol{\lambda}^{\Omega}$ and $\boldsymbol{\lambda}^{\Omega'}$ are the C-vectors of parameters for each saturated model. We also know that (4.1) and (4.2) are respectively ML estimates of the super-population model,

$$Log\left(\mu_c\right) = \mathbf{X}_c \boldsymbol{\lambda} \tag{4.3}$$

$$Log\left(\mu_c^{'}\right) = \mathbf{X}_c \boldsymbol{\lambda}^{'} \tag{4.4}$$

Let us consider a table whose C cells values are the ratios $\mu_c / M_c^{'}$, that is, the relative change between the expected value of the current count and the census year count for the cth-cell. Consider now the logarithm of the ratios $\mu_c / M_c^{'}$. Treating these ratios as counts, we can also model them using a saturated log-linear model,

$$Log\left(\frac{\mu_c}{M_c^{'}}\right) = \mathbf{X}_c \boldsymbol{\lambda}^{r} \tag{4.5}$$

where $\boldsymbol{\lambda}^{r}$ is the C-vector of parameters. We can get PL estimates for (4.5) using the same procedure explained in the previous Chapter (see Section 3.3.3). Therefore, using the direct estimates $\hat{M}_c$ we can obtain PL estimates for M_c,

$$Log\left(\frac{\tilde{\mu}_c^{PL}}{M_c^{'}}\right) = \mathbf{X}_c \tilde{\boldsymbol{\lambda}}^{r^{PL}} \tag{4.6}$$

We shall highlight here the similarity of (4.6) to a technique widely used in experimental design when the outcome of a variable is known to be correlated to an "exposure" value (see e.g. Agresti 1990). Let y_i, $i=1,2,...,n$, be a lexicographic-ordered set of variables or counts following a Poisson distribution with mean Y_i, from

a given cross-tabulation; that is, $y_i \sim P(Y_i)$. Let the mean Y_i be proportional to a quantity or "exposure" value E_i. Modelling the ratios Y_i / E_i allows us to compare the means Y_i allowing for the distortions produced by the E_i. Therefore, Y_i / E_i can be modelled as:

$$Log\left(\frac{Y_i}{E_i}\right) = \mathbf{X}_i \boldsymbol{\lambda}$$
$$Log\left(Y_i\right) - Log\left(E_i\right) = \mathbf{X}_i \boldsymbol{\lambda} \tag{4.7}$$

and the means Y_i can be obtained as:

$$Y_i = E_i \cdot e^{\mathbf{X}_i \boldsymbol{\lambda}} \tag{4.8}$$

Using the sample counts y_i we can obtain ML estimate of $\boldsymbol{\lambda}$ and thus of Y_i:

$$\tilde{Y}_i = E_i \cdot e^{\mathbf{X}_i \tilde{\boldsymbol{\lambda}}} \tag{4.9}$$

The term $Log\left(E_i\right)$ in (4.7) is usually called an 'offset'. Most currently available statistical software allow fitting models using an 'offset', and so obtaining count estimates like the set $\tilde{Y}_i$ above is a fairly simple task to carry out.

In (4.5) $M_c^{'}$ acts as the 'exposure' value we assume to be correlated to the current expected frequencies μ_c and $Log\left(M_c^{'}\right)$ in then equivalent to an offset. Therefore, the PL estimates for the current expected frequencies are given by,

$$\tilde{\mu}_c^{PL} = M_c^{'} \cdot e^{\mathbf{X}_c \tilde{\boldsymbol{\lambda}}^{PL}}$$

Let us now look into the structure of (4.5). By properties of logarithms, equation (4.5) is the logarithm of the current expected frequency μ_c minus the logarithm of the census count $M_c^{'}$,

$$Log\left(\frac{\mu_c}{M_c^{'}}\right) = Log\left(\mu_c\right) - Log\left(M_c^{'}\right) \tag{4.10}$$

It follows that, replacing (4.1) and (4.2) into (4.10),

$$\begin{aligned} Log\left(\frac{\mu_c}{M_c^{'}}\right) &= \mathbf{X}_c\boldsymbol{\lambda} - \mathbf{X}_c\boldsymbol{\lambda}^{\Omega'} \\ &= \mathbf{X}_c\left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{\Omega'}\right) \end{aligned} \tag{4.11}$$

Let us write the vectors of parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^{\Omega'}$ as a vectors consisting of two sub-vectors $\boldsymbol{\lambda}^{A}$, $\boldsymbol{\lambda}^{B}$ and $\boldsymbol{\lambda}^{'A}$, $\boldsymbol{\lambda}^{'B}$ respectively. That is:

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}^{A} \\ \boldsymbol{\lambda}^{B} \end{bmatrix}^{t} \quad ; \quad \boldsymbol{\lambda}^{\Omega'} = \begin{bmatrix} \boldsymbol{\lambda}^{\Omega'A} \\ \boldsymbol{\lambda}^{\Omega'B} \end{bmatrix}$$

so that (4.3) and (4.2) are now,

$$Log\left(\mu_c\right) = \mathbf{X}_c^{A}\boldsymbol{\lambda}^{A} + \mathbf{X}_c^{B}\boldsymbol{\lambda}^{B} \tag{4.12}$$

$$Log\left(M_c^{'}\right) = \mathbf{X}_c^{A}\boldsymbol{\lambda}^{\Omega'A} + \mathbf{X}_c^{B}\boldsymbol{\lambda}^{\Omega'B} \tag{4.13}$$

Here, $\boldsymbol{\lambda}^{A}$ and $\boldsymbol{\lambda}^{\Omega'A}$ are P_A-vectors containing the first P_A parameters of "lower dimension" in $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^{\Omega'}$ respectively. On the other hand, $\boldsymbol{\lambda}^{B}$ and $\boldsymbol{\lambda}^{\Omega'B}$ are the P_B-vectors containing the remaining P_B parameters in $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^{\Omega'}$; $P_A + P_B = C$. Thus, we can re-write (4.11) as,

$$Log\left(\frac{\mu_c}{M_c^{'}}\right) = \mathbf{X}_c\begin{pmatrix} \boldsymbol{\lambda}^{A} - \boldsymbol{\lambda}^{\Omega'A} \\ \boldsymbol{\lambda}^{B} - \boldsymbol{\lambda}^{\Omega'B} \end{pmatrix} \tag{4.14}$$

and given $\mathbf{X}_c = \left[\mathbf{X}_c^A, \mathbf{X}_c^B\right]$,

$$Log\left(\frac{\mu_c}{M_c^{'}}\right) = \mathbf{X}_c^A\left(\boldsymbol{\lambda}^A - \boldsymbol{\lambda}^{\Omega'A}\right) + \mathbf{X}_c^B\left(\boldsymbol{\lambda}^B - \boldsymbol{\lambda}^{\Omega'B}\right) \tag{4.15}$$

We have then that $\mathbf{X}_c\boldsymbol{\lambda}^r$ in (4.5) can be written as $\mathbf{X}_c^A\left(\boldsymbol{\lambda}^A - \boldsymbol{\lambda}^{\Omega'A}\right) + \mathbf{X}_c^B\left(\boldsymbol{\lambda}^B - \boldsymbol{\lambda}^{\Omega'B}\right)$. Therefore, the PL estimates in (4.6) are in fact obtained by solving the corresponding likelihood equations for vectors $\boldsymbol{\lambda}^A$ and $\boldsymbol{\lambda}^B$ conditioning on the vectors $\boldsymbol{\lambda}^{\Omega'A}$ and $\boldsymbol{\lambda}^{\Omega'B}$. In this case we can rewrite (4.6) as,

$$Log\left(\frac{\tilde{\mu}_c^{PL}}{M_c^{'}}\right) = \mathbf{X}_c^A\left(\tilde{\boldsymbol{\lambda}}^{A^{PL}} - \boldsymbol{\lambda}^{\Omega'A}\right) + \mathbf{X}_c^B\left(\tilde{\boldsymbol{\lambda}}^{B^{PL}} - \boldsymbol{\lambda}^{\Omega'B}\right) \tag{4.16}$$

Let us assume now that $\boldsymbol{\lambda}^B = \boldsymbol{\lambda}^{\Omega'B}$ i.e. the P_B parameters of the highest interaction order defining the structure of both the "actual" and the "reference" cross-tabulation can be considered as having the same value. In this case, the relative change $\mu_c / M_c^{'}$ can be modelled as the unsaturated log-linear model:

$$Log\left(\frac{\mu_c}{M_c^{'}}\right) = \mathbf{X}_c^A\left(\boldsymbol{\lambda}^A - \boldsymbol{\lambda}^{\Omega'A}\right) \tag{4.17}$$

and (4.16) would be,

$$Log\left(\frac{\tilde{\mu}_c^{PL}}{M_c^{'}}\right) = \mathbf{X}_c^A\left(\tilde{\boldsymbol{\lambda}}^{A^{PL}} - \boldsymbol{\lambda}^{\Omega'A}\right) \tag{4.18}$$

Now, noting that by logarithm properties (4.18) can be re-written as:

$$Log\left(\tilde{\mu}_c^{PL}\right) = \mathbf{X}_c^A\left(\tilde{\boldsymbol{\lambda}}^{A^{PL}} - \boldsymbol{\lambda}^{\Omega'A}\right) + Log\left(M_c^{'}\right) \tag{4.19}$$

by (4.13) we have that:

$$Log\left(\tilde{\mu}_c^{PL}\right) = \mathbf{X}_c^A\left(\tilde{\boldsymbol{\lambda}}^{A^{PL}} - \boldsymbol{\lambda}^{\Omega' A}\right) + \mathbf{X}_c^A\boldsymbol{\lambda}^{\Omega' A} + \mathbf{X}_c^B\boldsymbol{\lambda}^{\Omega' B}$$

$$Log\left(\tilde{\mu}_c^{PL}\right) = \mathbf{X}_c^A\tilde{\boldsymbol{\lambda}}^{A^{PL}} + \mathbf{X}_c^B\boldsymbol{\lambda}^{\Omega' B} \qquad (4.20)$$

It is important to bear in mind that $\mathbf{X}_c^A\tilde{\boldsymbol{\lambda}}^{A^{PL}}$ in (4.20) comes from (4.18), that is, $\tilde{\boldsymbol{\lambda}}^{A^{PL}}$ is the PL estimated vector of parameter for the underline log-linear model (4.17). These estimates are obtained by solving the corresponding likelihood equation for $\boldsymbol{\lambda}^A$ conditioning on the value of $\boldsymbol{\lambda}^{\Omega' A}$.

Following (4.19) and (4.20) we can write the following expression for the PL estimates of the expected frequencies,

$$\tilde{M}_c^{PL} = \tilde{\mu}_c^{PL} = M_c^{'} \cdot e^{\mathbf{X}_c^A\left(\tilde{\boldsymbol{\lambda}}^{A^{PL}} - \boldsymbol{\lambda}^{\Omega' A}\right)} = e^{\mathbf{X}_c^A\tilde{\boldsymbol{\lambda}}^{A^{PL}} + \mathbf{X}_c^B\boldsymbol{\lambda}^{\Omega' B}} \qquad (4.21)$$

We now recall the results of Birch (1963) where he shows that for a log-linear model there exists just one set of counts that both satisfies the model and ensures that the expected values of the minimal sufficient statistics are equal to their observed values. Therefore, if the set of counts given by (4.21) are the PL estimates for the current expected frequencies then they have to satisfy the likelihood equation for the model,

$$Log\left(\mu_c\right) = \mathbf{X}_c^A\boldsymbol{\lambda}^A + \mathbf{X}_c^B\boldsymbol{\lambda}^{\Omega' B}$$

where $\boldsymbol{\lambda}^{\Omega' B}$ is considered as a constant. Note that this is the same set of PL estimated counts as those obtained in section 3.3.5 in the previous chapter, where $\boldsymbol{\lambda}^A$ is considered as the "unknown" parameters and $\boldsymbol{\lambda}^{\Omega' B}$ is considered as the "known" parameters. Consequently, fitting a log-linear model using as an "exposure" variable the set of census counts is equivalent to fitting the same log-linear models as those described in chapter 3 and is therefore related to application of the SPREE process.

The advantage of this approach is that we do not need to know the assumed "known" parameters to fit the log-linear model with exposure values but only the counts from the reference table; this makes the estimation process simpler than the one implied by the original approach in section 3.3.5 where we have to find the census log-linear structure first to get the "known" parameters values.

Example

Let us consider again our example where the current set of counts M_{ijkq} and the set of counts from a previous census M'_{ijkq} are represented as:

$$Log\left(M_{ijkq}\right) = \mathbf{X}\boldsymbol{\lambda} = \mu + \lambda_i + \lambda_j + \lambda_k + \lambda_q + \lambda_{ij} + \lambda_{ik} + \lambda_{iq} + \lambda_{jk} + \lambda_{jq} + \lambda_{kq} + \lambda_{ijk} + \lambda_{ijq} + \lambda_{ikq} + \lambda_{jkq} + \lambda_{ijkq} \quad (4.22)$$

$$Log\left(M'_{ijkq}\right) = \mathbf{X}\boldsymbol{\lambda}' = \mu' + \lambda'_i + \lambda'_j + \lambda'_k + \lambda'_q + \lambda'_{ij} + \lambda'_{ik} + \lambda'_{iq} + \lambda'_{jk} + \lambda'_{jq} + \lambda'_{kq} + \lambda'_{ijk} + \lambda'_{ijq} + \lambda'_{ikq} + \lambda'_{jkq} + \lambda'_{ijkq} \quad (4.23)$$

Let us consider the ratio M_{ijkq} / M'_{ijkq}, that is, the relative change between the current count and the census year count for the *ijkq*-cell and let us suppose that ratio can be modelled as:

$$Log\left(\frac{M_{ijkq}}{M'_{ijkq}}\right) = \mathbf{X}\boldsymbol{\lambda}^r = \mu^r + \lambda_i^r + \lambda_j^r + \lambda_k^r + \lambda_q^r + \lambda_{ij}^r + \lambda_{ik}^r + \lambda_{iq}^r + \lambda_{jk}^r + \lambda_{jq}^r + \lambda_{kq}^r + \lambda_{ijk}^r + \lambda_{ijq}^r \quad (4.24)$$

That is, we assume that the relative change M_{ijkq} / M'_{ijkq} is sufficiently explained by the first order factors, the second order interactions and the *ijk* and the *ijq* third order interactions in the cross-tabulation without any other interaction term playing an important role. This means that the remaining third and fourth order terms in (4.22) and (4.23) are the same.

We get estimates $\tilde{M}_{ijkq}$ of the present counts M_{ijkq} by estimating $\boldsymbol{\lambda}^r$ in (4.24), so we have:

$$Log\left(\tilde{M}_{ijkq}\right) = Log\left(M^{'}_{ijkq}\right) + \mathbf{X}_{ijkq}\tilde{\boldsymbol{\lambda}}^r \tag{4.25}$$

and thus,

$$\tilde{M}_{ijkq} = M^{'}_{ijkq} \cdot e^{\mathbf{X}_{ijkq}\tilde{\lambda}^r} \tag{4.26}$$

Note that (4.25) is in fact:

$$\begin{aligned} Log\left(\tilde{M}_{ijkq}\right) &= \mu^{'} + \lambda^{'}_i + \lambda^{'}_j + \lambda^{'}_k + \lambda^{'}_q + \lambda^{'}_{ij} + \lambda^{'}_{ik} + \lambda^{'}_{iq} + \lambda^{'}_{jk} + \lambda^{'}_{jq} + \lambda^{'}_{kq} + \\ &+ \lambda^{'}_{ijk} + \lambda^{'}_{ijq} + \lambda^{'}_{ikq} + \lambda^{'}_{jkq} + \lambda^{'}_{ijkq} + \mu^r + \tilde{\lambda}^r_i + \tilde{\lambda}^r_j + \tilde{\lambda}^r_k + \tilde{\lambda}^r_q + \\ &+ \tilde{\lambda}^r_{ij} + \lambda^r_{ik} + \lambda^r_{iq} + \lambda^r_{jk} + \lambda^r_{jq} + \lambda^r_{kq} + \lambda^r_{ijk} + \lambda^r_{ijq} \\ &= (\mu^{'} + \mu^r) + (\lambda^{'}_i + \tilde{\lambda}^r_i) + (\lambda^{'}_j + \tilde{\lambda}^r_j) + (\lambda^{'}_k + \tilde{\lambda}^r_k) + (\lambda^{'}_q + \tilde{\lambda}^r_q) + \\ &+ \left(\lambda^{'}_{ij} + \tilde{\lambda}^r_{ij}\right) + \left(\lambda^{'}_{ik} + \tilde{\lambda}^r_{ik}\right) + \left(\lambda^{'}_{iq} + \tilde{\lambda}^r_{iq}\right) + \left(\lambda^{'}_{jk} + \tilde{\lambda}^r_{jk}\right) + \left(\lambda^{'}_{jq} + \tilde{\lambda}^r_{jq}\right) + \\ &+ \left(\lambda^{'}_{kq} + \tilde{\lambda}^r_{kq}\right) + \left(\lambda^{'}_{ijk} + \tilde{\lambda}^r_{ijk}\right) + \left(\lambda^{'}_{ijq} + \tilde{\lambda}^r_{ijq}\right) + \lambda^{'}_{ikq} + \lambda^{'}_{jkq} + \lambda^{'}_{ijkq} \end{aligned} \tag{4.27}$$

If we use the direct estimates $\hat{M}_{ijkq}$ to obtain $\tilde{\boldsymbol{\lambda}}^r$, we will have that $\tilde{\boldsymbol{\lambda}}^r = \tilde{\boldsymbol{\lambda}}^r_{PL}$ has to satisfy the same conditions given in (3.35) as well as the model. Consequently, $\tilde{M}_{ijkq} = \tilde{M}^{PL}_{ijkq}$ is the same set of counts obtained either using the estimated Log-linear model (3.37) with $\boldsymbol{\lambda}^K = \left(\lambda^{'}_{ikq}, \lambda^{'}_{jkq}, \lambda^{'}_{ijkq}\right)^t$ or using the relevant SPREE process as in Section 3.3.

4.1.2. Unsaturated SPREE

If as exposure variable we use counts from a "modified" reference table in the same fashion as those tables discussed in section 3.4, i.e. census tables with some of the highest interaction structures deleted, we get current count estimates equivalent to the "unsaturated SPREE" estimates in section 3.4.

Let us suppose that some of the parameters in vector $\lambda^{\Omega' B}$ in the log linear representation of the census counts (4.13) are considered to be zero, so that $P_A + P_B < C$ and let again λ^B be equal to $\lambda^{\Omega' B}$, i.e. $\lambda^B = \lambda^{\Omega' B}$. In this case the structure of equation (4.17) and the estimates in (4.18) will remain the same. However, note here that the set of census counts $M_c^{'}$ in the denominator are now those from the modified table, which should not differ that much from the original ones.

The structure of the PL expected frequencies (4.21) remains also the same but now the vector $\lambda^{\Omega' B}$ has some terms missing or equal to zero. Therefore, application of Birch's results means that the set of counts given by (4.21) has to satisfy the likelihood equation for the log-linear model equivalent to the unsaturated SPREE as in section 3.4. Note that if the assumption that the missing terms do not play a significant role in the model is sensible, then the resulting PL estimates from the unsaturated case will be close to those from the saturated one.

4.1.3. Logit Models

In section 3.6 we extended the SPREE method to product multinomial logistic regression. In doing so, we noted that the logarithm of the proportion ratios $\pi_{c'/g}/\pi_{base/g}$ -we follow now the notation defined in section 3.6 for groups g and categories c'-, is equal to the ratio of the related cell counts $\mu_{gc'}/\mu_{gbase}$, that is,

$$Log\left(\frac{\pi_{c'/g}}{\pi_{base/g}}\right) = Log\left(\frac{\mu_{gc'}}{\mu_{gbase}}\right) = Log\left(\mu_{gc'}\right) - Log\left(\mu_{gbase}\right) \qquad (4.28)$$

with corresponding PL estimates given by,

$$Log\left(\frac{\tilde{\pi}^{PL}_{c'/g}}{\tilde{\pi}^{PL}_{base/g}}\right) = Log\left(\tilde{\mu}^{PL}_{gc'}\right) - Log\left(\tilde{\mu}^{PL}_{gbase}\right) \qquad (4.29)$$

Suppose we now use the exposure variable approach to model the expected frequencies. Using the same correspondence in notation between the log-linear and the logit cases as in section 3.6, we have that from (4.19),

$$Log\left(\tilde{\mu}_{gc'}^{PL}\right)=\mathbf{X}_{gc'}^{A}\left(\tilde{\boldsymbol{\lambda}}^{A^{PL}}-\boldsymbol{\lambda}^{\Omega' A}\right)+Log\left(M_{gc'}^{'}\right) \tag{4.30}$$

Therefore we can write the right-hand side in (4.29) –recall that "base" is one of the c' response categories- as,

$$\left[\mathbf{X}_{gc'}^{A}\left(\tilde{\boldsymbol{\lambda}}^{A^{PL}}-\boldsymbol{\lambda}^{\Omega' A}\right)+Log\left(M_{gc'}^{'}\right)\right]-\left[\mathbf{X}_{gbase}^{A}\left(\tilde{\boldsymbol{\lambda}}^{A^{PL}}-\boldsymbol{\lambda}^{\Omega' A}\right)+Log\left(M_{gbase}^{'}\right)\right]$$

so that rearranging terms,

$$Log\left(\frac{\tilde{\pi}_{c'/g}^{PL}}{\tilde{\pi}_{base/g}^{PL}}\right)=\left(\mathbf{X}_{gc'}^{A}-\mathbf{X}_{gbase}^{A}\right)\left(\tilde{\boldsymbol{\lambda}}^{A^{PL}}-\boldsymbol{\lambda}^{\Omega' A}\right)+Log\left(\frac{M_{gc'}^{'}}{M_{gbase}^{'}}\right) \tag{4.31}$$

From section 4.6 we know that $\left(\mathbf{X}_{gc'}^{A}-\mathbf{X}_{gbase}^{A}\right)\tilde{\boldsymbol{\lambda}}^{A^{PL}}$ can be expressed in logistic notation as $\mathbf{Z}_{gc'}^{A}\tilde{\boldsymbol{\beta}}^{A^{PL}}$. In the same way, $\left(\mathbf{X}_{gc'}^{A}-\mathbf{X}_{gbase}^{A}\right)\boldsymbol{\lambda}^{\Omega' A}$ can be denoted in logistic notation as $\mathbf{Z}_{gc'}^{A}\tilde{\boldsymbol{\beta}}^{\Omega' A}$. We have thus that (4.31) can be written as,

$$\begin{aligned} Log\left(\frac{\tilde{\pi}_{c'/g}^{PL}}{\tilde{\pi}_{base/g}^{PL}}\right)&=\mathbf{Z}_{gc'}^{A}\left(\tilde{\boldsymbol{\beta}}^{A^{PL}}-\boldsymbol{\beta}^{\Omega' A}\right)+Log\left(\frac{M_{gc'}^{'}}{M_{gbase}^{'}}\right)\\ &=\mathbf{Z}_{gc'}^{A}\tilde{\boldsymbol{\beta}}^{E^{PL}}+Log\left(\frac{M_{gc'}^{'}}{M_{gbase}^{'}}\right)\end{aligned} \tag{4.32}$$

where $\tilde{\boldsymbol{\beta}}^{E^{PL}}=\left(\tilde{\boldsymbol{\beta}}^{A^{PL}}-\boldsymbol{\beta}^{\Omega' A}\right)$ is the vector of PL estimates of parameters of the logit model constrained on the set of constants $Log\left(M_{gc'}^{'}/M_{gbase}^{'}\right)$.

The logarithm on the right hand side of (4.32) is the logarithm of the numerator minus the logarithm of the denominator, that is, the difference of logarithms of census counts. We know that those logarithms can be modelled as,

$$Log\left(M^{'}_{gc'}\right)=\mathbf{X}^{A}_{gc'}\boldsymbol{\lambda}^{\Omega' A}+\mathbf{X}^{B}_{gc'}\boldsymbol{\lambda}^{\Omega' B} \tag{4.33}$$

Therefore we have that,

$$\begin{aligned} Log\left(\frac{M^{'}_{gc'}}{M^{'}_{gbase}}\right) &= \left(\mathbf{X}^{A}_{gc'}\boldsymbol{\lambda}^{\Omega' A}+\mathbf{X}^{B}_{gc'}\boldsymbol{\lambda}^{\Omega' B}\right)-\left(\mathbf{X}^{A}_{gbase}\boldsymbol{\lambda}^{\Omega' A}+\mathbf{X}^{B}_{gbase}\boldsymbol{\lambda}^{\Omega' B}\right) \\ &= \left(\mathbf{X}^{A}_{gc'}-\mathbf{X}^{A}_{gbase}\right)\boldsymbol{\lambda}^{\Omega' A}+\left(\mathbf{X}^{B}_{gc'}-\mathbf{X}^{B}_{gbase}\right)\boldsymbol{\lambda}^{\Omega' B} \end{aligned} \tag{4.34}$$

Substituting (4.34) into (4.31) we therefore obtain,

$$\begin{aligned} Log\left(\frac{\tilde{\pi}^{PL}_{c'/g}}{\tilde{\pi}^{PL}_{base/g}}\right) &= \left(\mathbf{X}^{A}_{gc'}-\mathbf{X}^{A}_{gbase}\right)\tilde{\boldsymbol{\lambda}}^{A^{PL}}+\left(\mathbf{X}^{B}_{gc'}-\mathbf{X}^{B}_{gbase}\right)\boldsymbol{\lambda}^{\Omega' B} \\ &= \mathbf{Z}^{A}_{gc'}\tilde{\boldsymbol{\beta}}^{A}+\mathbf{Z}^{B}_{gc'}\tilde{\boldsymbol{\beta}}^{B} \end{aligned} \tag{4.35}$$

which is the PL estimator of (3.57) already proved as the equivalent to carry out a SPREE process updating the structural terms related to $\boldsymbol{\lambda}^{A}=\boldsymbol{\lambda}^{U}$ whilst preserving the structural terms related to $\boldsymbol{\lambda}^{B}=\boldsymbol{\lambda}^{k}$ from a reference table.

It follows that the PL estimator of the logit model given in (4.32) is the equivalent to such a SPREE estimator. Note that here again we do not need the assumed "known" parameters to fit (4.32) but only the counts from the reference table; as in the log-linear situation described above, this makes the estimation process simple since there is not need to find the census log-linear structure first to get the "known" parameters values.

As Logit models deal directly with proportions, the idea of an "exposure" variable is not as natural as it is for log-linear models, thus we do not find this option for logit

models in statistical software. However, the ratios $Log\left(M'_{gc'}/M'_{gbase}\right)$ can be introduced in the fitting process as one of the independent variables with coefficient equal to one. This is easily done when a "parameter constraint" option is available in the statistical software.

Finally, from (4.31) and (4.35) we have that the set of PL estimates for the proportion $\tilde{\pi}^{PL}_{c'/g}$ are given by,

$$\tilde{\pi}^{PL}_{c'/g} = \frac{e^{\mathbf{Z}^A_{gc'}\tilde{\boldsymbol{\beta}}^{A^{PL}} + \mathbf{Z}^B_{gc'}\boldsymbol{\beta}^B}}{\sum_{c'=1}^{C'} e^{\mathbf{Z}^A_{gc'}\tilde{\boldsymbol{\beta}}^{A^{PL}} + \mathbf{Z}^B_{gc'}\boldsymbol{\beta}^B}} = \frac{e^{\left(\mathbf{X}^A_{gc'} - \mathbf{X}^A_{gbase}\right)\left(\tilde{\boldsymbol{\lambda}}^{A^{PL}} - \boldsymbol{\lambda}^{\Omega' A}\right) + Log\left(\frac{M'_{gc'}}{M'_{gbase}}\right)}}{\sum_{c'=1}^{C'} e^{\left(\mathbf{X}^A_{gc'} - \mathbf{X}^A_{gbase}\right)\left(\tilde{\boldsymbol{\lambda}}^{A^{PL}} - \boldsymbol{\lambda}^{\Omega' A}\right) + Log\left(\frac{M'_{gc'}}{M'_{gbase}}\right)}} \tag{4.36}$$

As we have already mentioned, there are many statistical packages that can be used to fit the models described in this chapter. These packages offer options to take into account the complexity of the sample design in order to obtain the correct direct estimators and parameter estimator quantities. The variances of the model-based estimate counts $\hat{M}_c$ can be obtained by writing programs not necessarily too complex. The structure of these variances is formally described below for the case of multinomial logistic models, which is the one we shall use for the simulation study carried out in this document.

4.2. PRODUCT MULTINOMIAL LOGISTIC MODELS

We now formally define the parameter estimation, variance estimation and model assessment processes for the general class of Logistic models we have proposed for sub-groups estimation. We do this using the notation specified for the particular situation of concern in this study, that is, the estimation of the proportion of people in each of the three categories (Employed, Unemployed and Non-active) needed for the calculation of the rates within each sub-population.

4.2.1. General Structure

We shall start by defining a general model structure that covers all the logistic models discussed in this document in the context of our LFS situation.

Let $U=\{u_1,u_1,...,u_v,...,u_M\}$ denote the Venezuelan population over 14 years old of size M for a specified reference time.

Let U be partitioned into $ijk=C$ groups (sub-populations) of sizes $M_{ijk\cdot}$ as specified in section 3.1, that is, $U=\{U_{1.1.1},U_{1.1.2},...,U_{ijk},...,U_{2.4.23}\}$ where $U_{ijk}=\{u_{ijk,1},u_{ijk,2},...,u_{ijk,v},...,u_{ijk,M_{ijk\cdot}}\}$.

Let y denote the variable "labour force status" of three mutually exclusive and exhaustive possible outcomes (1=Employee, 2=Unemployed or 3=Non-Active). Let $y_{ijk,1},y_{ijk2},...,y_{ijk,v},...,y_{ijk,M_{ijk}}$ be the values of y for the $M_{ijk\cdot}$ elements in group ijk and let us express M_{ijkq} (already defined in section 3.1) as:

$$M_{ijkq}=\sum_{v=1}^{M_{ijk\cdot}} I\left(y_{ijk,v}\right) \quad ; \quad I\left(y_{ijk,v}\right)=\begin{cases}1 & \text{if } y_{ijk,v}=q\\ 0 & \text{otherwise}\end{cases}$$

We will consider the set of counts M_{ijkq} as an independent sample from a super-population that follows a multinomial distribution with $P(y_{ijk,v}=q)=\pi_{q/ijk}$ and $\sum_{q=1}^{3}\pi_{q/ijk}=1$.

Suppose $\pi_{q/ijk}$ is related to a P-vector of dummy variables $\mathbf{X}_{ijkq}=(x_{ijkq,1},...,x_{ijkq,p},...,x_{ijkq,P})$ defining the effects and/or interaction terms related to count M_{ijkq} with $P=(I\cdot J\cdot K)(Q-1)$, so that $\mathbf{X}=\{\mathbf{X}_{ijkq}\}$ is the zero-one $(I\cdot J\cdot K\cdot Q)\times P$ saturated model matrix. Let also $\boldsymbol{\beta}=(\beta_1,\beta_2,...,\beta_p,...,\beta_P)'$ be the P-

vector of super-population parameters for this saturated model, such that, $\pi_{q/ijk} = \pi_{q/ijk}(\boldsymbol{\beta}) = h(\mathbf{X}_{ijkq}, \boldsymbol{\beta})$ is a general response model for $\pi_{q/ijk}$.

We now define $\mathbf{X}_{ijkq} = (\mathbf{X}^U_{ijkq}, \mathbf{X}^K_{ijkq}, \mathbf{X}^O_{ijkq})$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}^{U'}, \boldsymbol{\beta}^{K'}, \boldsymbol{\beta}^{O'})^t$ so that $\mathbf{X} = (\mathbf{X}^U, \mathbf{X}^K, \mathbf{X}^O)$. Here, $\boldsymbol{\beta}^U$ is a P_u-sub-vector containing the parameters in $\boldsymbol{\beta}$ that are unknown, $\boldsymbol{\beta}^K$ is a P_k-sub-vector containing the non-zero parameters from $\boldsymbol{\beta}$ that are known -so there is no need for them to be estimated-, and $\boldsymbol{\beta}^O$ is a P_o-sub-vector containing the parameters from $\boldsymbol{\beta}$ that are zero; $P = (P_u + P_k + P_o)$. Accordingly, $\mathbf{X}^U$, $\mathbf{X}^K$ and $\mathbf{X}^O$ are the $(I \cdot J \cdot K \cdot Q) \times P_u$, $(I \cdot J \cdot K \cdot Q) \times P_k$ and $(I \cdot J \cdot K \cdot Q) \times P_o$ model sub-matrices related to $\boldsymbol{\beta}^U$, $\boldsymbol{\beta}^K$ and $\boldsymbol{\beta}^O$ respectively. We note that either P_k or P_o (or both) can be equal to zero.

We use Generalised Linear Model (GLM) theory to specify the structure of $\pi_{q/ijk}(\boldsymbol{\beta})$, $g(\pi_{q/ijk}) = \sum_{p=1}^{P} x_{ijkq,p}\, \beta_p$ with logistic link function $g(\pi_{q/ijk}(\boldsymbol{\beta})) = \ln\left(\pi_{q/ijk}(\boldsymbol{\beta}) / \pi_{3/ijk}(\boldsymbol{\beta})\right)$. We have then that the general structure of the model is:

$$\ln\left(\frac{\pi_{q/ijk}(\boldsymbol{\beta})}{\pi_{3/ijk}(\boldsymbol{\beta})}\right) = \mathbf{X}_{ijkq}\boldsymbol{\beta} \tag{4.37}$$

with $\pi_{q/ijk}(\boldsymbol{\beta}^U, \boldsymbol{\beta}^K, \boldsymbol{\beta}^O)$ given by,

$$\pi_{q/ijk}(\boldsymbol{\beta}) = \frac{e^{\mathbf{X}_{ijkq}\boldsymbol{\beta}}}{\sum_{q=1}^{3} e^{\mathbf{X}_{ijkq}\boldsymbol{\beta}}} = \frac{e^{\mathbf{X}_{ijkq}\boldsymbol{\beta}}}{1 + \sum_{q=1}^{2} e^{\mathbf{X}_{ijkq}\boldsymbol{\beta}}} \tag{4.38}$$

4.2.2. Parameters and Variance Estimators: general form

4.2.2.a. Standard ML estimation

The multinomial distribution of the M_{ijkq} can be expressed as follows:

$$p\left(M_{ijk1}, M_{ijk2}, M_{ijk3} / \left(\sum_{q=1}^{3} M_{ijkq} = M_{ijk\cdot}\right)\right) = \left(\frac{M_{ijk\cdot}!}{\prod_{q=1}^{3} M_{ijkq}!}\right) \cdot \prod_{q=1}^{3} \pi_{q/ijk}(\boldsymbol{\beta})^{M_{ijkq}} \quad (4.39)$$

We want to estimate the unknown sub-vector $\boldsymbol{\beta}^U$ within vector $\boldsymbol{\beta}$. The ML estimate for $\boldsymbol{\beta}^U$ is the vector $\tilde{\boldsymbol{\beta}}^U$ that maximizes the likelihood function,

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \left[\left(\frac{M_{ijk\cdot}!}{\prod_{q=1}^{3} M_{ijkq}!}\right) \cdot \prod_{q=1}^{3} \pi_{q/ijk}(\boldsymbol{\beta})^{M_{ijkq}} \right] \quad (4.40)$$

Taking the logarithm of (4.40) and differentiating with respect to β_p^U, we get:

$$\begin{aligned} \left(\frac{\partial L}{\partial \beta_p^U}\right) &= \sum_{i=1}^{2} \sum_{j=1}^{4} \sum_{k=1}^{23} \left[M_{ijkq} x_{ijkq,p}^U - M_{ijk\cdot} x_{ijkq,p}^U \frac{e^{\mathbf{X}_{ijkq}^t \boldsymbol{\beta}}}{\sum_{q=1}^{3} e^{\mathbf{X}_{ijkq}^t \boldsymbol{\beta}}} \right] \\ &= \sum_{i=1}^{2} \sum_{j=1}^{4} \sum_{k=1}^{23} \left[M_{ijkq} x_{ijkq,p}^U - M_{ijk\cdot} x_{ijkq,p}^U \pi_{q/ijk}(\boldsymbol{\beta}) \right] \quad (4.41) \\ &= \sum_{i=1}^{2} \sum_{j=1}^{4} \sum_{k=1}^{23} \left[x_{ijkq,p}^U \left(M_{ijkq} - M_{ijk\cdot} \pi_{q/ijk}(\boldsymbol{\beta}) \right) \right] \end{aligned}$$

In matrix notation, let us consider again the cells ijk ordered lexicographically as $c = 1, \ldots, C$ where *C=184*. Let $\mathbf{X}_c^U = \mathbf{X}_{ijk}^U = \begin{pmatrix} \mathbf{X}_{ijk1}^U \\ \mathbf{X}_{ijk2}^U \end{pmatrix}$, $\boldsymbol{\pi}_c(\boldsymbol{\beta}) = \begin{pmatrix} \pi_{1/ijk}(\boldsymbol{\beta}) \\ \pi_{2/ijk}(\boldsymbol{\beta}) \end{pmatrix}$ and

$$\mathbf{p}_c = \begin{pmatrix} p_{1/ijk} \\ p_{2/ijk} \end{pmatrix}, \text{ such that: } \mathbf{X}^U = \begin{pmatrix} \mathbf{X}_1^U \\ \vdots \\ \mathbf{X}_c^U \\ \vdots \\ \mathbf{X}_C^U \end{pmatrix}, \ \boldsymbol{\pi}(\boldsymbol{\beta}) = \begin{pmatrix} \boldsymbol{\pi}_1(\boldsymbol{\beta}) \\ \vdots \\ \boldsymbol{\pi}_c(\boldsymbol{\beta}) \\ \vdots \\ \boldsymbol{\pi}_C(\boldsymbol{\beta}) \end{pmatrix}, \ \mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_c \\ \vdots \\ \mathbf{p}_C \end{pmatrix}$$

that is, $\mathbf{X}^U$ is defined as before, $\boldsymbol{\pi}(\boldsymbol{\beta})$ is the IJK -vector of probabilities $\pi_{q/ijk}(\boldsymbol{\beta})$ and $\mathbf{p}$ is the IJK -vector of sample proportions $p_{q/ijk}$ defined in section 3.1. Finally, let $\mathbf{M} = (M_1, ..., M_c, ..., M_C)^t$ be the IJK-vector of cell counts $M_c = M_{ijk\circ}$. Thus, we can write (4.41) as follows:

$$\left(\frac{\partial \mathbf{L}}{\partial \boldsymbol{\beta}^U}\right) = \mathbf{X}^{U'}\mathbf{D}(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\beta}^U, \boldsymbol{\beta}^K, \boldsymbol{\beta}^O)) \tag{4.42}$$

where $\mathbf{D} = diag(\mathbf{M}) \otimes \mathbf{I}_2$ with $\otimes$ denoting Kronecker product. The ML estimates for $\boldsymbol{\beta}$ are given by $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}^{U'}, \boldsymbol{\beta}^{K'}, \boldsymbol{\beta}^{O'})^t$ and consequently $\boldsymbol{\pi}(\tilde{\boldsymbol{\beta}}) = \tilde{\boldsymbol{\pi}}$ where $\tilde{\boldsymbol{\beta}}^U$ is obtained by setting the P_u-vector (4.42) to the P_u-vector of zeros $\mathbf{0}$ and solving for $\boldsymbol{\beta}^U$. This leads to the P_u likelihood equations,

$$\mathbf{X}^{U'}\mathbf{D}\mathbf{p} = \mathbf{X}^{U'}\mathbf{D}\boldsymbol{\pi}(\boldsymbol{\beta}^U, \boldsymbol{\beta}^K, \boldsymbol{\beta}^O) \tag{4.43}$$

We can get estimators of the covariance matrices for $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\pi}}$ by using Taylor approximation or by combining standard GLM results with those of Royall (1986). Noting that,

$$\mathbf{X}^t\mathbf{D}(\mathbf{p} - \boldsymbol{\pi}(\tilde{\boldsymbol{\beta}})) \cong \left(\frac{\partial \mathbf{L}}{\partial \boldsymbol{\beta}^U}\right) + \left(\frac{\partial^2 \mathbf{L}}{\partial \boldsymbol{\beta}^{U2}}\right)(\tilde{\boldsymbol{\beta}}^U - \boldsymbol{\beta}^U)$$

and

$$\boldsymbol{\pi}(\tilde{\boldsymbol{\beta}}) \cong \boldsymbol{\pi}(\boldsymbol{\beta}) + \left(\frac{\partial \boldsymbol{\pi}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^U}\right)(\tilde{\boldsymbol{\beta}}^U - \boldsymbol{\beta}^U)$$

By definition, $\mathbf{X}^t\mathbf{D}(\mathbf{p}-\boldsymbol{\pi}(\tilde{\boldsymbol{\beta}}))=\mathbf{0}$, so assuming that the inverse of $\frac{\partial^2\mathbf{L}}{\partial\boldsymbol{\beta}^{U2}}$ exists, we have that,

$$(\tilde{\boldsymbol{\beta}}^U-\boldsymbol{\beta}^U)\sim\left(-\frac{\partial^2\mathbf{L}}{\partial\boldsymbol{\beta}^{U2}}\right)^{-1}\left(\frac{\partial\mathbf{L}}{\partial\boldsymbol{\beta}^U}\right) \tag{4.44}$$

and

$$\left(\boldsymbol{\pi}(\tilde{\boldsymbol{\beta}})-\boldsymbol{\pi}(\boldsymbol{\beta})\right)\sim\left(\frac{\partial\boldsymbol{\pi}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^U}\right)(\tilde{\boldsymbol{\beta}}^U-\boldsymbol{\beta}^U) \tag{4.45}$$

Thus regarding $\left(-\frac{\partial^2\mathbf{L}}{\partial\boldsymbol{\beta}^{U2}}\right)^{-1}$ as a constant, we finally obtain,

$$\tilde{\mathbf{Cov}}(\tilde{\boldsymbol{\beta}}^U)\cong\left(-\frac{\partial^2\mathbf{L}}{\partial\boldsymbol{\beta}^{U2}}\right)^{-1}\hat{\mathbf{Cov}}\left(\frac{\partial\mathbf{L}}{\partial\boldsymbol{\beta}^U}\right)\left(-\frac{\partial^2\mathbf{L}}{\partial\boldsymbol{\beta}^{U2}}\right)^{-1} \tag{4.46}$$

and

$$\tilde{\mathbf{Cov}}(\boldsymbol{\pi}(\tilde{\boldsymbol{\beta}}))=\left(\frac{\partial\boldsymbol{\pi}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^U}\right)\tilde{\mathbf{Cov}}(\tilde{\boldsymbol{\beta}}^U)\left(\frac{\partial\boldsymbol{\pi}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^U}\right) \tag{4.47}$$

where $\hat{\mathbf{Cov}}$ in (4.46) denotes design estimated variances and covariances, that is, not depending on the model and $\left(-\frac{\partial^2\mathbf{L}}{\partial\boldsymbol{\beta}^{U2}}\right)=\left(\mathbf{X}^t\mathbf{D}\boldsymbol{\Delta}\mathbf{X}\right)$, $\left(\frac{\partial\boldsymbol{\pi}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^U}\right)=\boldsymbol{\Delta}\mathbf{X}$, with $\boldsymbol{\Delta}=Blockdiag\left(Diag(\mathbf{p}_c)-\mathbf{p}_c\mathbf{p}_c^t\right)$.

4.2.2.b. Allowing for complex sampling design: PL estimation

In reality, we do not have information about the whole population U but only information from the LFS sample. Let $S=\{s_1,s_1,...,s_v,...,s_m\}$ denote a sample of U obtained using a specific sampling design. We have then that S is also partitioned into $ijk=C$ groups of sizes $m_{ijk.}$ as specified in section 3.1, that is, $S=\{S_{1.1.1},S_{1.1.2},...,S_{ijk},...,S_{2.4.23}\}$ where $S_{ijk}=\{s_{ijk,1},s_{ijk,2},...,s_{ijk,v},...,s_{ijk,M_{ijk.}}\}$. We shall denote the values of y for the $m_{ijk.}$ elements in group ijk as $y_{ijk,1},y_{ijk2},...,y_{ijk,v},...,y_{ijk,m_{ijk}}$ and will also denote m_{ijkq} (already defined in section 3.1) as:

$$m_{ijkq}=\sum_{v=1}^{m_{ijk.}} I\left(y_{ijk,v}\right) \quad ; \quad I\left(y_{ijk,v}\right)=\begin{cases}1 \text{ if } y_{ijk,v}=q\\ 0 \text{ otherwise}\end{cases}$$

If the sample counts m_{ijkq} follow a multinomial distribution with $P(y_{ijk,v}=q)=\pi_{q/ijk}$ and $\sum_{q=1}^{3}\pi_{q/ijk}=1$, we could apply the theory described earlier in this section to obtain parameter and variance estimates. However, the LFS sampling design is a complex one (see section 2.2.2) involving clustering of elements as well as stratification. The assumption that $y_{ijk,1},...,y_{ijk,m_{ijk}}$ are independent may not necessarily be true. We also have that $E(y_{ijk,v})$ may not equal $m_{ijk}\boldsymbol{\pi}_{ijk}$. This affects the distribution function $P(m_{ijk1},m_{ijk2},m_{ijk3})$, which is the conditional distribution of the population given the sampling design and not necessarily the same as (4.39). As the strata do not cross the cells, these do not affect the distribution function although we should bear in mind the assumption behind the post-stratification process used at the LFS estimation stage (see section 2.2.4); these post-strata do cross the cells. Moreover, the clustering present in the sampling design also affects the distribution function $P(m_{ijk1},m_{ijk2},m_{ijk3})$.

To avoid the complex task of specifying a model for the sample data -which might also bring some misspecification problems to the inference (Skinner et. al. 1989)- it is

customary to use the Pseudo-Maximum Likelihood approach (PL) explained in section 3.1.2. Suppose we have observed the whole population so that the sample vector $\mathbf{p}$ is in fact the finite population vector of proportion $\mathbf{P}$. In that case we can simply apply all the theory described in this section to calculate the "census" or population vector $\boldsymbol{\beta}^{\Omega}$ for the target population (see Binder 1983), which itself is an estimate of the super-population parameter $\boldsymbol{\beta}$ defined above. However, since we have not observed the whole population but a sample from it, we must first calculate direct estimates $\hat{\mathbf{P}}$ of relevant population quantities and use them in (4.42) to obtain 'pseudo' maximum likelihood estimations $\tilde{\boldsymbol{\beta}}_{PL} = (\tilde{\boldsymbol{\beta}}^{U^t}, \boldsymbol{\beta}^{K^t}, \boldsymbol{\beta}^{O^t})^t$ of the vector parameter. Formally, (4.42) can be written:

$$\left(\frac{\partial \mathbf{L}}{\partial \boldsymbol{\beta}^{U}}\right) = \mathbf{X}^t \mathbf{D}(\hat{\mathbf{P}} - \boldsymbol{\pi}(\boldsymbol{\beta}^{U}, \boldsymbol{\beta}^{K}, \boldsymbol{\beta}^{O})) \tag{4.48}$$

where $\hat{\boldsymbol{\pi}} = (\hat{\boldsymbol{\pi}}_1, ..., \hat{\boldsymbol{\pi}}_c, ..., \hat{\boldsymbol{\pi}}_C)^t$ with $\hat{\boldsymbol{\pi}}_c = (\hat{\pi}_{1/ijk}, \hat{\pi}_{2/ijk})$, $\hat{\pi}_{q/ijk}$ given by (2.8) but calculated at sub-group levels and $\mathbf{D} = diag(\mathbf{M}) \otimes \mathbf{I}_2$ with $\mathbf{M} = (M_1, ..., M_c, ..., M_C)^t$ (the census population projections for sub-groups). Setting (4.48) equal to the vector of zeroes $\mathbf{0}$, we get the likelihood equations,

$$\mathbf{X}^{U^t} \mathbf{D} \hat{\mathbf{P}} = \mathbf{X}^{U^t} \mathbf{D} \boldsymbol{\pi}(\boldsymbol{\beta}^{U}, \boldsymbol{\beta}^{K}, \boldsymbol{\beta}^{O}) \tag{4.49}$$

Note that (4.49) depends on the individual estimated counts $(M_{ijk.} \cdot \hat{\pi}_{q/ijk})$ only via appropriate aggregates. Solving (4.49) for $\boldsymbol{\beta}^{U}$ gives the PL-estimates $\tilde{\boldsymbol{\beta}}_{PL} = (\tilde{\boldsymbol{\beta}}^{U^t}_{PL}, \boldsymbol{\beta}^{K^t}, \boldsymbol{\beta}^{0^t})^t$ of the model parameters, which can then be substituted in (4.38) to obtain the model (PL) estimates $\tilde{\pi}^{PL}_{q/ijk} = \pi_{q/ijk}(\tilde{\boldsymbol{\beta}}_{PL})$. Note that $\tilde{\boldsymbol{\beta}}_{PL}$ is also a model-based estimate for the "census" vector $\boldsymbol{\beta}^{\Omega}$, i.e. $\tilde{\boldsymbol{\beta}}_{PL} = \tilde{\boldsymbol{\beta}}^{\Omega}_{PL}$. Accordingly, $\tilde{\pi}^{PL}_{q/ijk}$ a the model-based estimate of the finite population proportion $P_{q/ijk}$, $\tilde{P}^{PL}_{q/ijk} = P_{q/ijk}(\tilde{\boldsymbol{\beta}}^{\Omega}_{PL})$. The model-based (PL) estimates for the sub-group counts M_{ijkq} are therefore $\tilde{M}^{PL}_{ijkq} = \tilde{\pi}^{PL}_{q/ijk} M_{ijk\cdot}$.

We now write $\mathbf{Cov}(\tilde{\boldsymbol{\beta}}_{PL})$ as,

$$\mathbf{Cov}(\tilde{\boldsymbol{\beta}}_{PL}) \cong \left(-\frac{\partial^2 \mathbf{L}}{\partial \boldsymbol{\beta}^{U2}}\right)^{-1} Cov\left(\frac{\partial \mathbf{L}}{\partial \boldsymbol{\beta}^{U}}\right)\left(-\frac{\partial^2 \mathbf{L}}{\partial \boldsymbol{\beta}^{U2}}\right)^{-1} \tag{4.50}$$

with

$$\hat{\mathbf{Cov}}\left(\frac{\partial \mathbf{L}}{\partial \boldsymbol{\beta}^{U}}\right) = \hat{\mathbf{Cov}}\left(\mathbf{X}^{U'}\,\mathbf{D}\,(\hat{\mathbf{P}} - \boldsymbol{\pi}(\boldsymbol{\beta}\,))\right)$$
$$= \left(\mathbf{X}^{U'}\,\mathbf{D}\,\mathbf{Cov}(\hat{\mathbf{P}})\,\mathbf{D}\,\mathbf{X}^{U}\right)$$

where,

$$\left(-\frac{\partial^2 \mathbf{L}}{\partial \boldsymbol{\beta}^{U2}}\right) = \left(\mathbf{X}^{t}\mathbf{D}\Delta\mathbf{X}\right); \qquad \text{with } \Delta = Blockdiag\left(Diag(\boldsymbol{\pi}_c) - \boldsymbol{\pi}_c\boldsymbol{\pi}_c^{t}\right).$$

We estimate (4.50) by using $\tilde{\boldsymbol{\pi}}_c^{PL} = \boldsymbol{\pi}_c(\tilde{\boldsymbol{\beta}}_{PL})$ for $\boldsymbol{\pi}_c = \boldsymbol{\pi}_c(\boldsymbol{\beta}_{PL})$ and $\hat{\mathbf{Cov}}(\hat{\mathbf{P}})$ for $\mathbf{Cov}(\hat{\mathbf{P}})$. $\hat{\mathbf{Cov}}(\hat{\mathbf{P}})$ is given by (2.17) and (2.18) for the proportion case at sub-group level. Therefore, $\tilde{\mathbf{Cov}}(\tilde{\boldsymbol{\beta}}_{PL})$ is,

$$\tilde{\mathbf{Cov}}(\tilde{\boldsymbol{\beta}}_{PL}) \cong \left(\mathbf{X}^{t}\mathbf{D}\tilde{\Delta}\mathbf{X}\right)^{-1}\left(\mathbf{X}^{t}\,\mathbf{D}\,\hat{\mathbf{Cov}}(\hat{\mathbf{P}})\,\mathbf{D}\,\mathbf{X}\right)\left(\mathbf{X}^{t}\mathbf{D}\tilde{\Delta}\mathbf{X}\right)^{-1} \tag{4.51}$$

with $\tilde{\Delta} = Blockdiag\left(Diag(\tilde{\boldsymbol{\pi}}_c) - \tilde{\boldsymbol{\pi}}_c\tilde{\boldsymbol{\pi}}_c^{t}\right)$. Similarly, the covariance matrix of $\boldsymbol{\pi}(\tilde{\boldsymbol{\beta}}_{PL})$ is given by,

$$\mathbf{Cov}\left(\boldsymbol{\pi}(\tilde{\boldsymbol{\beta}}_{PL})\right) = \left(\frac{\partial \boldsymbol{\pi}(\boldsymbol{\beta}\,)}{\partial \boldsymbol{\beta}^{U}}\right)\mathbf{Cov}(\tilde{\boldsymbol{\beta}}_{PL}^{U})\left(\frac{\partial \boldsymbol{\pi}(\boldsymbol{\beta}\,)}{\partial \boldsymbol{\beta}^{U}}\right); \qquad \text{with } \left(\frac{\partial \boldsymbol{\pi}(\boldsymbol{\beta}\,)}{\partial \boldsymbol{\beta}^{U}}\right) = \Delta\mathbf{X}$$

which is estimated by,

$$\tilde{\mathbf{Cov}}\left(\boldsymbol{\pi}(\tilde{\boldsymbol{\beta}}_{PL})\right) = \tilde{\Delta}\mathbf{X}\left(\mathbf{X}^{t}\mathbf{D}\tilde{\Delta}\mathbf{X}\right)^{-1}\left(\mathbf{X}^{t}\,\mathbf{D}\,\hat{\mathbf{Cov}}(\hat{\mathbf{P}})\,\mathbf{D}\,\mathbf{X}\right)\left(\mathbf{X}^{t}\mathbf{D}\tilde{\Delta}\mathbf{X}\right)^{-1}\mathbf{X}^{t}\tilde{\Delta} \tag{4.52}$$

4.2.3. Parameters and Variance Estimators: specific situations

The general model structure (4.37) covers all the logistic models described in this document. The systematic part of the model, i.e. $\mathbf{X\beta} = \mathbf{X}(\boldsymbol{\beta}^{U'}, \boldsymbol{\beta}^{K'}, \boldsymbol{\beta}^{O'})^t$, will have different shapes depending on the specific situation we want to address.

For logistic models equivalent to "saturated" or traditional SPREE estimation, $\mathbf{X\beta}$ will contain the full range of parameters for the saturated model, i.e. $P_o = 0$ and $P_u + P_k = P$, so that $\mathbf{X\beta} = \mathbf{X}(\boldsymbol{\beta}^{U'}, \boldsymbol{\beta}^{K'})^t$. The estimation process follows as it has been described above but without $\boldsymbol{\beta}^O$. As we have seen, the number of equations involved in (4.49) is equal to the number of parameter in $\boldsymbol{\beta}^U$; however, the process of solving for $\boldsymbol{\beta}^U$ is conditioned on $\boldsymbol{\beta}^K$. Note that if we use the "offset" approach described at the beginning of this chapter, $\boldsymbol{\beta}^K$ then consists of the set of census log-ratios $Log\left(M'_{gc'}/M'_{gbase}\right)$.

For logistic models equivalent to what we have called "Unsaturated SPREE" (section 3.4), some of the higher order interaction effects are considered to be zero. In this case we have $P_o > 0$ and $P_k > 0$ with $P_u + P_k + P_o = P$, so that $\mathbf{X\beta} = \mathbf{X}(\boldsymbol{\beta}^{U'}, \boldsymbol{\beta}^{K'}, \boldsymbol{\beta}^{O'})^t$ as described above.

If we decide not to preserve structural terms from the reference cross-tabulation but still set some of the higher order interaction effects to zero, the equivalent logistic model has $P_o > 0$ and $P_k = 0$ with $P_u + P_o = P$. Here, $\mathbf{X\beta} = \mathbf{X}(\boldsymbol{\beta}^{U'}, \boldsymbol{\beta}^{O'})^t$ and the estimation process follows as it has been described above but without $\boldsymbol{\beta}^K$. Again, the number of equations involved in (4.49) will be equal to the number of parameter in $\boldsymbol{\beta}^U$ and the process of solving for $\boldsymbol{\beta}^U$ is conditioned on $\boldsymbol{\beta}^O$.

4.2.4. Goodness of Fit

The statistics most widely used to test goodness of fit of model based estimates and certain hypothesized values are the Pearson chi-squared test statistics X_P^2 and the Likelihood ratio statistics X_{LR}^2. In our case these are defined as follows:

$$X_P^2 = m\sum_{c=1}^{C} W_c \sum_{q=1}^{3} \frac{\left(\hat{\pi}_{q/c} - \tilde{\pi}_{q/c}\right)^2}{\tilde{\pi}_{q/c}} \qquad \text{and} \qquad X_{LR}^2 = 2m\sum_{c=1}^{C} W_c \sum_{q=1}^{3} Log\left[\frac{\hat{\pi}_{q/c}}{\tilde{\pi}_{q/c}}\right]$$

where $W_c = M_c/M$. We can also test nested hypotheses where a non-saturated model *G1* defined in terms of a *R-vector* of parameters (*R<P)* is assumed to hold by setting $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, where $\boldsymbol{\beta}_1$ is a R_1-vector and $\boldsymbol{\beta}_2$ is a R_2-vector, $R = R_1 + R_2$. In this case we test the hypothesis $\boldsymbol{\beta}_2 = 0$, by testing the goodness of fit of a model *G2* consisting only of the *R1-vector* of parameters $\boldsymbol{\beta}_1$, given that *G1* holds. Let $\tilde{\pi}_{q/ijk}$ denote the *PL*-estimates for *G1* and also let $\tilde{\tilde{\pi}}_{q/ijk}$ be the *PL*-estimates for *G2*. The Pearson chi-squared statistic and the Likelihood ratio statistics for the nested hypothesis are then as follows:

$$X_P^2(G2/G1) = m\sum_{c=1}^{C} W_c \sum_{q=1}^{3} \frac{\left(\tilde{\pi}_{q/c} - \tilde{\tilde{\pi}}_{q/c}\right)^2}{\tilde{\tilde{\pi}}_{q/c}} \quad \text{and} \quad X_{LR}^2(G2/G1) = 2m\sum_{c=1}^{C} W_c \sum_{q=1}^{3} Log\left[\frac{\tilde{\pi}_{q/c}}{\tilde{\tilde{\pi}}_{q/c}}\right]$$

Under multinomial sampling the distribution of both X_P^2 and X_{LR}^2 are asymptotically chi-squared with degrees of freedom $[IJK(Q-1)-P] = (368-P)$. Likewise, for the nested case, under multinomial sampling the distribution of both X_P^2 and X_{LR}^2 are asymptotically chi-squared with degrees of freedom $(R - R_1) = R_2$. However, since we have a complex sample design, the distribution of both X_P^2 and X_{LR}^2 is shown by Rao and Scott (1981) to be asymptotically equivalent to a weighted sum of 368-P (R_2 in

the nested case) independent chi-squared random variables each with one degree of freedom. That is, $X_P^2 \approx \delta_1 Z_1 + \cdots + \delta_{368-P} Z_{368-P}$, $X_{LR}^2 \approx \delta_1 Z_1 + \cdots + \delta_{368-P} Z_{368-P}$, $X_P^2(G2/G1) \approx \delta_1 Z_1 + \cdots + \delta_{R_2} Z_{R_2}$ and $X_{LR}^2(G2/G1) \approx \delta_1 Z_1 + \cdots + \delta_{R_2} Z_{R_2}$, with $Z_i \approx \chi_1^2$ where the weights $(\delta_1 \geq \cdots \geq \delta_{368-P} > 0)$ are the eigenvalues of the design effects matrix,

$$\nabla = \left(\mathbf{B}'\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{B}\right)^{-1}\left(\mathbf{B}'\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{Cov}(\hat{\boldsymbol{\pi}})\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{B}\right)$$

where $\mathbf{B}$ is the matrix which complements $\mathbf{X}$ to form the zero-one model matrix for the saturated model. In general, $\mathbf{B}$ can be any (IJKQ x *df*) full rank matrix satisfying $\mathbf{B}'\mathbf{X} = \mathbf{0}$, where *df*=degree of freedom equal to *368-P* or R_2 in the nested case. Note that the weights δ_i are equal to one in the case of product multinomial sampling, so they can be interpreted as `generalized design effects'.

We obtain estimators of δ_i, $\hat{\delta}_i$, through the matrix $\hat{\nabla}$, that is,

$$\hat{\nabla} = \left(\mathbf{B}'\mathbf{D}\tilde{\boldsymbol{\Delta}}^{-1}\mathbf{B}\right)^{-1}\left(\mathbf{B}'\mathbf{D}\tilde{\boldsymbol{\Delta}}^{-1}\hat{\mathbf{Cov}}(\hat{\boldsymbol{\pi}})\mathbf{D}\tilde{\boldsymbol{\Delta}}^{-1}\mathbf{B}\right)$$

The sum of the *df* terms $\hat{\delta}_i$ is given by $tr(\hat{\nabla})$ (e.g. Harville 1997, p.539). Therefore, a Rao-Scott first-order correction for the statistics X_P^2 and X_{LR}^2, which is asymptotically distributed as chi-squared with *df* degree of freedom is given by:

$$X_P^2\left(\hat{\delta}.\right) = \frac{X_P^2}{\hat{\delta}.} \qquad \text{and} \qquad X_{LR}^2\left(\hat{\delta}.\right) = \frac{X_{LR}^2}{\hat{\delta}.} \tag{4.53}$$

with

$$\hat{\delta}. = \sum_{i=1}^{df} \frac{\hat{\delta}_i}{df} = \frac{tr(\hat{\nabla})}{df}$$

This correction works well if the variability of the terms $\hat{\delta}_i$ is not large. A Rao-Scott second-order correction that takes into account this variability is given by:

$$X_P^2\left(\hat{\delta}.,\hat{a}\right)=\frac{X_p^2(\hat{\delta}.)}{1+\hat{a}^2} \qquad \text{and} \qquad X_{LR}^2\left(\hat{\delta}.,\hat{a}\right)=\frac{X_{LR}^2(\hat{\delta}.)}{1+\hat{a}^2} \tag{4.54}$$

where $\hat{a}^2$ is the coefficient of variation of the $\hat{\delta}_i$, that is,

$$\hat{a}^2=\frac{\sum_{i=1}^{df}\hat{\delta}_i^2}{(df)\hat{\delta}.^2}-1=\frac{(df)tr(\hat{\nabla}^2)}{tr(\hat{\nabla})^2}-1$$

Both $X_P^2\left(\hat{\delta}.\right)$ and $X_{LR}^2\left(\hat{\delta}.\right)$ are asymptotically distributed as chi-squared random variables with $(df)/(1+\hat{a}^2)$ degrees o freedom.

The second-order correction has been shown to perform well in different situations in several empirical studies (e.g. Rao and Scott 1981, Roberts et.al. 1987, Thomas and Rao 1987, Rao and Thomas 1999). An alternative test that takes into account the complex design used in selecting the sample is based on the Wald Statistic:

$$X_W^2=(\tilde{\gamma}_{PL})^t\,\tilde{\mathbf{Cov}}(\tilde{\gamma}_{PL})^{-1}(\tilde{\gamma}_{PL})$$

Here $\tilde{\gamma}_{PL}$ is the pseudo-likelihood estimate of the $[IJK(Q-1)-P]$-vector of parameters that we want to test being equal to zero and $\tilde{\mathbf{Cov}}(\tilde{\gamma}_{PL})$ is the estimator of the variance-covariance matrix of $\tilde{\gamma}_{PL}$ defined as in (4.51).

However, X_W^2 has been found to perform poorly in several empirical studies, especially when the degree of freedom for the estimated covariance matrix is not large compared with the number of cells in the table (e.g. Rao and Scott 1981, Fay 1985, Thomas and Rao 1987, Rao and Thomas 1999, Molina and Skinner 1992); furthermore, it is not defined if any of the $\hat{\pi}_{q/c}$ are equal to zero as occurs in the LFS data.

4.2.5. Diagnostics

A critical assessment of the models to be considered in this study can be carried out by obtaining measures that give us some insight into the presence of outlying cells and influential points.

By identifying outlying cells we can get a rough idea of the cells that are poorly explained by the model. We do not expect any model to accurately explain the behaviour of every cell but we do expect a good model to account for most of them. The relative importance between states and demographic groups tend to vary depending on the analyst and the kind of analysis he/she is going to undertake. For instance, having reliable estimates about the Unemployment rates relative to the population under 45 years old will be more important than those relative to the over 45 years old groups. This is especially true in countries with "pyramid-like" age population structures like Venezuela. Thus, the final judgement on the usefulness of a model has to take into account this relative importance which depends on the realities of each country. That is why we are especially concerned here with the detection of patterns like, for instance, specific states or sex-age groups that the model seems unable to explain.

Extreme points in the design space can affect the usefulness of a model. This can have an important influence in the structure of the model and consequently on the accuracy of estimation. This is the reason for our interest in the existence of influential points in our data. Should these points exist, some assessment regarding the influence they have on the estimation process needs to be undertaken, with the aim of establishing if the predicting power of the model can be improved by removing such cells from the estimation process.

We shall follow Pregibon (1981) in define some useful diagnostic measures in this regard. He bases his suggestions mainly on measures that are easily calculated by using data naturally obtained during the fitting process. This is highly convenient in most of the practical situations when human and economic resources are limited, as it is in many national statistics offices. Although his work is based on a maximum

likelihood fit of a logistic regression model, it is also valid for any within the exponential family, even if pseudo-maximum likelihood has been used, as is the case in this thesis. For the remainder of this section, the sub-script *c=1,...,C(Q-1)* will denote the lexicography order of the product multinomial table omitting the cell related to the *Qth* labour force category.

Standardised residuals are commonly used as a first attempt to identify outliers. Following the notation used so far in this thesis, let $\mathbf{r} = \hat{\mathbf{P}} - \tilde{\boldsymbol{\pi}}_{PL}$ be the $IJK(Q-1)$-vector of residuals. We define the standardised residuals $e_{cq} = r_{cq} / \hat{V}^{1/2}_{cq,cq}(r_{cq})$ where $\hat{V}^{1/2}_{cq,cq}(r_{cq})$, the estimated standard error for the residual r_{cq}, is given by the squared root of the *c*th component of the diagonal of the matrix:

$$\hat{\mathbf{Cov}}(\mathbf{r}) = \left[\mathbf{I} - \hat{\Delta}\mathbf{X}\left(\mathbf{X}^t\mathbf{D}\hat{\Delta}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{D}\right]\hat{\mathbf{Cov}}(\hat{\pi})\left[\mathbf{I} - \hat{\Delta}\mathbf{X}\left(\mathbf{X}^t\mathbf{D}\hat{\Delta}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{D}\right]^t$$

We obtain this matrix by noting that from (4.45) and (4.44) we have

$$\left(\pi(\tilde{\boldsymbol{\beta}}_{PL}) - \pi(\boldsymbol{\beta})\right) \sim \Delta\mathbf{X}(\tilde{\boldsymbol{\beta}}_{PL} - \boldsymbol{\beta}) \quad \text{and} \quad \left(\tilde{\boldsymbol{\beta}}_{PL} - \boldsymbol{\beta}\right) = \left(\mathbf{X}^t\mathbf{D}\Delta\mathbf{X}\right)\mathbf{X}^t\mathbf{D},$$

so that,

$$\begin{aligned}\hat{\pi} - \tilde{\pi}_{PL} &\sim (\hat{\pi} - \pi) - \hat{\Delta}\mathbf{X}\left(\mathbf{X}^t\mathbf{D}\hat{\Delta}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{D}(\hat{\pi} - \pi) \\ &\sim \left[\mathbf{I} - \hat{\Delta}\mathbf{X}\left(\mathbf{X}^t\mathbf{D}\hat{\Delta}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{D}\right](\hat{\pi} - \pi)\end{aligned}$$

Assuming normality of the standardised residual distribution, we can then compare these standardised residuals with the values of the standard normal distribution, looking for potential outliers and extreme points. There are different ways of carrying out this comparison suggested in the specialised literature (e.g. Pregibon 1981, McCullagh and Nedler 1983, Dobson 1990, Agresti 1990, Draper and Smith 1998). We can for instance plot the ordered standardised residuals against the expected

normal order statistics. However, it is worth mentioning that cells with $\hat{\pi}_{cq} = 0$ and $\hat{\pi}_{cq} = 1$ makes the distribution of **r** skew and so makes the assumption of normality unreliable. A plot of the standardised residuals against the fitted values, excluding the points close to either zero or one, can give us an idea of the validity of the normality assumption related to the distribution of **r**. This plot should show an evenly spread pattern.

Pregibon (1981) suggests an alternative method of detecting outliers that avoids problems with estimates that are close to zero and one. It consists in the use of components of the chi-squared statistic $X_{P,c}$ or $X_{LR,c}$ noting the fact that large values of these components suggest potential outliers. As we are concern with the complex sampling design case, it seems appropriate to use components of the Rao-Scott second order correction (or first order correction in cases where the second order correction can not be calculated) $X_{P,c}\left(\hat{\delta}_{\cdot}, \hat{a}\right)$ and $X_{LR,c}\left(\hat{\delta}_{\cdot}, \hat{a}\right)$ for this purpose as Roberts et. al. (1987) did for complex data in a binomial context. Again, different plots can be used as a visual check for outlying points.

To detect influential points, Pregibon (1981) suggests the use of the projection matrix **M** that in our complex design case is given by the expression:

$$\begin{aligned}\mathbf{M} &= \left[\mathbf{I} - \mathbf{D}^{1/2}\hat{\boldsymbol{\Delta}}^{1/2}\mathbf{X}\left(\mathbf{X}^t\mathbf{D}\hat{\boldsymbol{\Delta}}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{D}^{1/2}\hat{\boldsymbol{\Delta}}^{1/2}\right] \\ &= \left[\mathbf{I} - \mathbf{H}\right]\end{aligned} \qquad (4.55)$$

The interpretation of this matrix is similar to that of $\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ in traditional regression analysis, in the sense that the diagonal of the second term gives an indicator of the influence of each point in the design space. This can easily be seen if we consider the fitting process as carried out by iterative re-weighted least-squares or by Newton-Raphson Methods (see Pregibon 1981, p.712). Therefore, small values m_{cc} in the diagonal of **M** will identify potential influential cells. A visual examination of a plot of m_{cc} against c can give us an idea of the existence of such

cells. Hoaglin & Welch (1978) suggested looking at cells with $m_{cc} < 1-[(2-P)/C]$ as potentially influential for the linear case. Pregibon (1981) used the same criteria as a rough cut-off.

An alternative plot that can summarise information about outlying cells and extreme points in the design space is a plot of the values in the diagonal of the **H** matrix, h_{cc}, against $X^2_{P,c}\left(\hat{\delta}.,\hat{a}\right)/X^2_P\left(\hat{\delta}.,\hat{a}\right)$ or $X^2_{LR,c}\left(\hat{\delta}.,\hat{a}\right)/X^2_{LR}\left(\hat{\delta}.,\hat{a}\right)$. Note also that:

$$\frac{X^2_{P,c}\left(\hat{\delta}.,\hat{a}\right)}{X^2_P\left(\hat{\delta}.,\hat{a}\right)} = \frac{X^2_{P,c}/\hat{\delta}.(1+\hat{a}^2)}{X^2_P/\hat{\delta}.(1+\hat{a}^2)} = \frac{X^2_{P,c}}{X^2_P}$$

and

$$\frac{X^2_{LR,c}\left(\hat{\delta}.,\hat{a}\right)}{X^2_{LR}\left(\hat{\delta}.,\hat{a}\right)} = \frac{X^2_{LR,c}/\hat{\delta}.(1+\hat{a}^2)}{X^2_{LR}/\hat{\delta}.(1+\hat{a}^2)} = \frac{X^2_{LR,c}}{X^2_{LR}}$$

If influential points are detected, further investigation about how heavily these points influence the estimation process should be carried out. The impact can be measured with respect to parameter estimates, fitted values or chi-squared statistics. In the study reported in this thesis we aim to obtain good estimates of rates and are not interested in the specific structure and dynamics of the underlying factors determining these rates. Therefore, we will not be concerned with the impact of influential points on parameter estimates, unless it involves significant impact on fitted values.

To measure the effect of extreme points on the fitted values as well as on the goodness-of-fit statistics, Pregibon (1981) again used a generalisation of the traditional regression diagnostic. Similar results were also implemented by Roberts et. al. (1987) for complex data in the binomial context.

Let $\tilde{\boldsymbol{\beta}}_{PL}(c)$ be the pseudo-likelihood estimate of the parameter vector $\boldsymbol{\beta}$ calculated without taking into account the c-th cell. Likewise, let $\tilde{\boldsymbol{\pi}}_{PL}(c) = \boldsymbol{\pi}\left(\tilde{\boldsymbol{\beta}}_{PL}(c)\right)$ be the

vector of fitted values using $\tilde{\boldsymbol{\beta}}_{PL}(c)$. A measure of the effect of cell c-th on the fitted value l-th can be obtained as follows,

$$\Delta_l X^2_{P,c}\left(\hat{\delta}.,\hat{a}\right)=X^2_{P,l-c}\left(\hat{\delta}.,\hat{a}\right)-X^2_{P,l}\left(\hat{\delta}.,\hat{a}\right)$$

or

$$\Delta_l X^2_{LR,c}\left(\hat{\delta}.,\hat{a}\right)=X^2_{LR,l-c}\left(\hat{\delta}.,\hat{a}\right)-X^2_{LR,l}\left(\hat{\delta}.,\hat{a}\right)$$

where $X^2_{P,l-c}\left(\hat{\delta}.,\hat{a}\right)$ and $X^2_{LR,l-c}\left(\hat{\delta}.,\hat{a}\right)$ are the contributions of cell l-th to the Pearson and Likelihood Ratio goodness-of-fit statistics respectively when cell c-th has been removed from the fitting process. Pregibon (1981) uses the following approximation to $\Delta X^2_{LR,l}=X^2_{LR,l-c}-X^2_{LR,l}$ for non-complex binomial data:

$$\Delta_l X^2_{LR,c}\cong\frac{2X_{P,c}h_{lc}X_{P,l}}{1-h_{ll}}+\frac{X^2_{P,l}h^2_{lc}}{\left(1-h_{ll}\right)^2} \tag{4.56}$$

where h_{lc} denote the element l,c of the matrix for non-complex binomial data that is equivalent to the **H** matrix in (4.55). For our complex independent multinomial data we use the h_{lc} elements from the **H** matrix in (4.55) and $X^2_{P,l}\left(\hat{\delta}.,\hat{a}\right)$ and $X^2_{P,c}\left(\hat{\delta}.,\hat{a}\right)$ instead of $X^2_{P,l}$ and $X^2_{P,c}$ in (4.56). This leads to the following expression:

$$\Delta_l X^2_{LR,c}\left(\hat{\delta}.,\hat{a}\right)\cong\frac{2X_{P,c}\left(\hat{\delta}.,\hat{a}\right)h_{lc}X_{P,l}\left(\hat{\delta}.,\hat{a}\right)}{1-h_{ll}}+\frac{X^2_{P,l}\left(\hat{\delta}.,\hat{a}\right)h^2_{lc}}{\left(1-h_{ll}\right)^2} \tag{4.57}$$

For each potential influential point, we can obtain a set of C measures $\Delta_l X^2_{LR,c}\left(\hat{\delta}.,\hat{a}\right)$. Then we can plot each of these sets against c for a visual inspection of the impact of each influential point on every other cell. Note that negatives $\Delta_l X^2_{LR,c}\left(\hat{\delta}.,\hat{a}\right)$ values indicate an improvement in the c-th cell fit due to removing cell l from the fitting

process. The opposite is also true, positive $\Delta_l X^2_{LR,c}\left(\hat{\delta}.,\hat{a}\right)$ values indicate a worsening in the c-th cell fit due to removing cell l from the fitting process.

Regarding the effect of cell c-th on the goodness-of-fit value, we can obtain useful measures from the following expressions:

$$\Delta_l X^2_P\left(\hat{\delta}.,\hat{a}\right)=X^2_P\left(\hat{\delta}.,\hat{a}\right)-X^2_{P,-l}\left(\hat{\delta}.,\hat{a}\right)$$

or

$$\Delta_l X^2_{LR}\left(\hat{\delta}.,\hat{a}\right)=X^2_{LR}\left(\hat{\delta}.,\hat{a}\right)-X^2_{LR,-l}\left(\hat{\delta}.,\hat{a}\right)$$

They indicate changes in the value of the goodness-of-fit statistics (Pearson Chi-squared and Likelihood Ratio) due to deleting the l-th cell from the fitting process. Again, Pregibon (1981) uses an approximation to $\Delta X^2_{LR}=X^2_{LR}-X^2_{LR,-l}$ for non-complex binomial data:

$$\Delta_l X^2_{LR}\cong X^2_{LR,l}+\frac{X^2_{P,l}h_{ll}}{1-h_{ll}} \tag{4.58}$$

He also gives an approximation for changes in the Pearson Chi-squared statistics $\Delta X^2_P=X^2_P-X^2_{P,-l}$, that is:

$$\Delta_l X^2_P\cong\frac{X^2_{P,l}}{1-h_{ll}} \tag{4.59}$$

though he warns about its inferiority compared with $\Delta_l X^2_{LR}$ due to the fact that X^2_P does not necessarily decrease as data is removed from the fitting process.

As we did in (4.57), we write an approximation to $\Delta_l X^2_{LR}\left(\hat{\delta}.,\hat{a}\right)$ as follows,

$$\Delta_l X^2_P\left(\hat{\delta}.,\hat{a}\right)\cong\frac{X^2_{P,l}\left(\hat{\delta}.,\hat{a}\right)}{1-h_{ll}} \tag{4.60}$$

A plot of $\Delta_l X^2_{LR}\left(\hat{\delta}_{\cdot}, \hat{a}\right)$ against l then provides us with a visual tool to examine the magnitude of the changes in the Likelihood Ratio statistics when different cells are removed from the fitting process.

CHAPTER 5

EMPIRICAL RESULTS FOR VENEZUELAN LFS

In this Chapter we carry out an empirical analysis of the different models that can be used to obtain SPREE estimates, including unsaturated SPREE and conventional Logistic models.

We first discuss some important issues about the structure of the population and its implication in the estimation process. Next, we carry out an empirical analysis and provide diagnostics for different potential models using data from the Venezuelan 1981 and 1990 Population Census. We then describe and discuss the results from a simulation study based on the Venezuelan 1990 Census; this simulation study is designed to explore the design-based properties of a number of competing estimators based on the theory developed in previous sections. Finally, we briefly discuss some issues related to using "time" as an extra dimension in the SPREE process.

5.1. CONSIDERATIONS ABOUT THE STRUCTURE OF THE POPULATION

Throughout this document, we have taken the count M_{ijkq} to be a realisation from a super-population following an independent Poisson distribution with expected value and variance equal to μ_{ijkq}. Likewise, when we condition on the marginals $M_{ijk.}$, we have assumed these counts are distributed as product multinomial with

$E(M_{ijkq}) = M_{ijk.}\pi_{q/ijk}$ and $V(M_{ijkq}) = M_{ijk.}\pi_{q/ijk}(1-\pi_{q/ijk})$; in this case we also assume that each individual in the finite population is independent and identical distributed as a product multinomial with $P(y_{ijk,v} = q) = \pi_{q/ijk}$ and $\sum_{q=1}^{3}\pi_{q/ijk} = 1$. Furthermore, we also assume that $\pi_{q/ijk}$ are related to a set of variables $\mathbf{X}_{ijkq} = (x_{ijkq,1},..,x_{ijkq,p},..,x_{ijkq,P})$ allowing us to model $\pi_{q/ijk}$ as $\pi_{q/ijk} = \pi_{q/ijk}(\boldsymbol{\beta}) = h(\mathbf{X}_{ijkq},\boldsymbol{\beta})$ with $\boldsymbol{\beta} = (\beta_1,..,\beta_p,..,\beta_P)^t$ a P-vector of unknown super-population parameters.

Since in real situations we work with survey samples, we described a Pseudo-Maximum Likelihood approach to deal with the distortion caused on the survey data distribution by sampling designs like the LFS, which involves strata, clusters and unequal selection probabilities. This would "reassemble" the finite population distribution, producing estimates $\tilde{\pi}_{q/ijk}^{PL}$ and their variance estimates allowing for the adjustment needed for the sampling design. This theoretical formulation implies that if we knew the finite population counts M_{ijkq} we would be able to estimate the model parameters $\pi_{q/ijk}(\boldsymbol{\beta})$ and their variance applying standard estimation procedures like ML estimation.

However, the theoretical process generating the finite or "census" population counts can be more complex than the one described here. This seems plausible taking into account the clustering pattern human beings follow regarding allocation, the internal regionalization of a country regarding economical activities as well as others cultural patterns found in different societies. For instance, we have that the "geographical" or "spatial" component is likely to be more complex than the physical division implied by "states". It is noticeable the socio-economic differences found in Latin-American countries between "main" cities and the rest of the country. Moreover, within those "main" cities, it is likely to find the usual clustering structure regarding socio-economic aspects with a between cluster heterogeneity certainly higher than what can be found in developed countries. Others factors related to individual characteristics or internal households composition might also play a role in the super-population structure. Educational level and relationship with other household members are examples of potentially important variables in explaining that structure.

In those situations, units from the same sub-population *ijk* might in fact have different expected values π given by the "true" more complex super-population model. In other words, the super-population distribution generating the finite population might have different parameters π_q depending not only on the variables *ijk* but also on other variables. This obviously violates the distributional assumptions about M_{ijkq} and $y_{ijk,v}$ specified above. Elements in the finite population belonging to a specific sub-group *ijk* are not longer independent and the set of counts M_{ijkq} are not distributed as product multinomial $(M_{ijk\cdot}, \pi_{q/ijk})$ any more.

Given these circumstances, our first reaction would be to change the working model in favour of a model that better represents the "real" dynamic governing labour force variables. The original target parameters $\pi_{q/ijk}$ would be disaggregated so new parameters $\pi_{q/ijkd}$, with d representing one or more than one extra dimension, would become the target of analysis. This can be accomplished by adding an extra dimension to the working cross-tabulation, corresponding to adding fixed parameters to the model. Alternatively, we can add random parameters to the model at the extra-dimension level or treat geographical information as "extra-levels" in the analysis (multilevel modelling). The approach we should follow will depend on the available information regarding the sampled element as well as the "geographic" strata and/or cluster relevant units. This process would lead to a better understanding of the geographical and socio-economic interaction of the process in the population. For a comprehensive discussion on this subject see Chapter 10 in Skinner et. al. (1989).

At this point it is necessary to go back to the initial formulation and targets of this study. Its main goal is to produce estimates for the main indicators required by the national statistics office in Venezuela (INE) as well as in most Latin American countries. For different reasons, these indicators are needed at state level and disaggregated by gender and age-groups; we therefore require estimates of the counts M_{ijkq} or the proportions $P_{q/ijk} = M_{ijkq}/M_{ijk\cdot}$ needed to construct the labour force rates that are the ultimate target of this work. This fact does not mean that a deeper

analysis and understanding of the dynamics of the labour force is not necessary. On the contrary, this would obviously be of great benefit to analysts and to national statistics systems. However, due to restrictions in the data available to us for this study –INE policy regarding data disclosure is far from an open one- as well as restriction in time we will not attempt this sort of analysis, even though we acknowledge its importance.

We still have to address the impact on variance estimation and hypothesis testing from misspecification of the model. It is pertinent to stress that our target parameters are the finite population counts M_{ijkq} and proportions $P_{q/ijk}$ and not the model proportions $\pi_{q/ijk}$. We use a model for $\pi_{q/ijk}$ as a theoretical tool to get estimates for $P_{q/ijk}$ and regard the sampling design as the only impediment to carrying out appropriate analysis of the finite population parameters using standard statistical procedures. Therefore, the PL approach and the adjustment made to the Pearson and LR chi-squared statistics (refer to Chapter 4) should be all we need to compute parameter and variance estimates and to carry out tests of hypotheses for the finite population parameters. Those procedures take care of the effect that the sampling design has on the data so we can carry out an appropriate analysis with focus on our finite population aims.

This point is particularly important when working with SPREE estimation. The assumed "known" parameters in the model have a real interest when the focus is on the analysis of the finite population. They are used to improve the quality of the finite population estimates even if the model is not appropriate for describing the super-population process. Let us consider the logistic representation of the finite population set of probabilities $P_{q/ijk}$ given by,

$$Log\left(\frac{P_{q/ijk}}{P_{3/ijk}}\right)=\alpha_q^{\Omega}+\beta_{iq}^{\Omega}+\beta_{jq}^{\Omega}+\beta_{kq}^{\Omega}+\beta_{ijq}^{\Omega}+\beta_{ikq}^{\Omega}+\beta_{jkq}^{\Omega}+\beta_{ijkq}^{\Omega} \qquad (5.1)$$

Suppose we know the interaction terms β_{ijq}^{K}, β_{ikq}^{K}, β_{jkq}^{K} and β_{ijkq}^{K} from a set of finite population proportions $P_{q/ijk}^{'}$ from a previous period of time, which are roughly the same as their equivalent current interaction terms, that is,

$$\begin{aligned}\beta_{ijq}^{\Omega} &= \beta_{ijq}^{K} \\ \beta_{ikq}^{\Omega} &= \beta_{ikq}^{K} \\ \beta_{jkq}^{\Omega} &= \beta_{jkq}^{K} \\ \beta_{ijkq}^{\Omega} &= \beta_{ijkq}^{K}\end{aligned}$$

We formulate this model for the super-population generating the finite population and apply the methods described in previous chapters, so that using the LFS sample we get the PL estimates for the super-population logits, that is,

$$Log\left(\frac{\tilde{\pi}_{q/ijk}^{PL}}{\tilde{\pi}_{3/ijk}^{PL}}\right) = \tilde{\alpha}_{q}^{PL} + \tilde{\beta}_{iq}^{PL} + \tilde{\beta}_{jq}^{PL} + \tilde{\beta}_{kq}^{PL} + \beta_{ijq}^{K} + \beta_{ikq}^{K} + \beta_{jkq}^{K} + \beta_{ijkq}^{K} \qquad (5.2)$$

which are also the estimates for the finite population logits in (5.1),

$$Log\left(\frac{\tilde{P}_{q/ijk}^{PL}}{\tilde{P}_{3/ijk}^{PL}}\right) = \tilde{\alpha}_{q}^{\Omega^{PL}} + \tilde{\beta}_{iq}^{\Omega^{PL}} + \tilde{\beta}_{jq}^{\Omega^{PL}} + \tilde{\beta}_{kq}^{\Omega^{PL}} + \beta_{ijq}^{K} + \beta_{ikq}^{K} + \beta_{jkq}^{K} + \beta_{ijkq}^{K} \qquad (5.3)$$

The set of PL estimates $\tilde{\pi}_{q/ijk}^{PL} = \tilde{P}_{q/ijk}^{PL}$ should therefore be of high precision for the finite population proportions $P_{q/ijk}$ no matter whether the explanatory variables in the model are appropriate or not. Similarly, the variance estimators and adjusted tests described in the previous chapter should be the appropriate when estimating finite population characteristics, even though this might not be the case when inference about super-population quantities is of primary interest.

5.2. MODELS STRUCTURES FOR SPREE AND RELATED MODELS

We have seen in previous chapters that the log-linear and logistic models equivalent to the SPREE method consist of two set of parameters, the unknown parameters and the assumed known parameters. The latter are those related to the structural terms we want to preserve from the reference cross-tabulation. We have also seen that the estimation process is conditioned on those "known" parameters; however, the structure of the likelihood equations will depend only on those margins related to the "unknown" parameters.

The decision here is what structural terms we want to preserve and what structural term we will update in the estimation process. This decision is related to the information available to us regarding the margins of the cross-tabulation as well as the confidence we have about the stability over time of the structure of the reference table. The decision is likely to be a compromise between the following two statements.

Firstly, to update specific structural terms we must have reliable information regarding the current marginal counts related to those terms. Available information will basically depend on population projections based on the latest census, vital statistics and the LFS and its sampling design.

Secondly, to preserve specific structural terms we should have the confidence those terms have remained sufficiently stable between the two points in time involved in the analysis. In general, this confidence will weaken as the gap between those two points in time gets bigger. How fast it weakens will depend on the dynamic of the variables governing the socio-economics process in the country.

Another factor to take into account is the fact that the LFS traditionally produces direct estimates for labour force indicators for sex-age groups at the national level. These indicators are widely published and the LFS design ensures that these estimates have good reliability. It is thus necessary that the model estimated counts agree with the sex-age group direct estimates at the national level. This suggests that the set of

pseudo-likelihood equations related to the fitting process for the log-linear working model should satisfy the equalities $\hat{M}_{ij\bullet q} = M_{ij\bullet q}(\tilde{\boldsymbol{\lambda}}^{U}_{PL}, \boldsymbol{\lambda}^{K})$. For this to be possible, the interaction terms I-J-Q have to be regarded as part of the "unknown" terms in the model, that is, the term λ^{U}_{ijq} has to be present in our log-linear model, which is equivalent to saying that the interaction term β^{U}_{ijq} has to be in our logistic model.

It follows that the choice should be made among the four following logistic models,

$$Log\left(\frac{\pi_{q/ijk}}{\pi_{3/ijk}}\right) = \alpha^{U}_{q} + \beta^{U}_{iq} + \beta^{U}_{jq} + \beta^{U}_{kq} + \beta^{U}_{ijq} + \beta^{K}_{ikq} + \beta^{K}_{jkq} + \beta^{K}_{ijkq} \tag{5.4}$$

$$Log\left(\frac{\pi_{q/ijk}}{\pi_{3/ijk}}\right) = \alpha^{U}_{q} + \beta^{U}_{iq} + \beta^{U}_{jq} + \beta^{U}_{kq} + \beta^{U}_{ijq} + \beta^{U}_{ikq} + \beta^{K}_{jkq} + \beta^{K}_{ijkq} \tag{5.5}$$

$$Log\left(\frac{\pi_{q/ijk}}{\pi_{3/ijk}}\right) = \alpha^{U}_{q} + \beta^{U}_{iq} + \beta^{U}_{jq} + \beta^{U}_{kq} + \beta^{U}_{ijq} + \beta^{K}_{ikq} + \beta^{U}_{jkq} + \beta^{K}_{ijkq} \tag{5.6}$$

$$Log\left(\frac{\pi_{q/ijk}}{\pi_{3/ijk}}\right) = \alpha^{U}_{q} + \beta^{U}_{iq} + \beta^{U}_{jq} + \beta^{U}_{kq} + \beta^{U}_{ijq} + \beta^{U}_{ikq} + \beta^{U}_{jkq} + \beta^{K}_{ijkq} \tag{5.7}$$

Model (5.4) is the simplest model of this kind assuring an agreement between the model estimates and the LFS sex-age group direct estimates at the national level whilst model (5.7) is the most complex.

Other structures like the independence model (5.8) below can also be considered, however, their use does not guarantee the agreement explained above. Should we decide to use them, we would have to carry out a calibration process after the estimates have been obtained in order to obtain the desired agreement. For this reason, model (5.8) would have to prove significantly better than (5.4) to justify a more complex estimation process.

$$Log\left(\frac{\pi_{q/ijk}}{\pi_{3/ijk}}\right) = \alpha^{U}_{q} + \beta^{U}_{iq} + \beta^{U}_{jq} + \beta^{U}_{kq} + \beta^{K}_{ijq} + \beta^{K}_{ikq} + \beta^{K}_{jkq} + \beta^{K}_{ijkq} \tag{5.8}$$

Unsaturated SPREE models will result from "deleting" k-terms from the models above. Different combinations can be obtained depending on what and how many k-terms are deleted.

5.3. CENSUS-BASED COMPARATIVE ANALYSIS

We have seen in Chapter 2 that the most recent census available in Venezuela at the moment is the one carried out in 1990[1]. That represents a gap of over ten years with respect to the present year, and does not sound too promising when thinking of structural terms as "preserved" over time.

We will use the 1990 and 1981 censuses in a first attempt to explore the behaviour of the SPREE models in different situations.

In this section we explore empirically the SPREE estimation process when the gap in time is nine years. We shall assume we do not know the 1990 counts but only some of its aggregated marginals. We will use these marginals and the 1981 census table to produce SPREE estimates and Unsaturated SPREE estimates for 1990. These estimations will be calculated using the "exposure" approach proposed in Chapter 4 using the software STATA 7.0[2].

Let M^{81}_{ijkq} and M^{90}_{ijkq} be the census'81 and the "unknown" census'90 counts for sex category i, age group j, state k and labour force classification q. We want to get SPREE estimates for M^{90}_{ijkq} assuming we know a set of 1990 marginal aggregates.

[1] A Census was carried out in Venezuela in the year 2001; however, no database and only results at the national level were available at the time this document was produced.
[2] Stata Corporation, Texas, Release 2001

We know that using the following saturated logistic models,

$$Log\left(\frac{P_{q/ijk}^{81}}{P_{3/ijk}^{81}}\right)=\mathbf{X}_{ijkq}\boldsymbol{\beta}^{81^{\Omega}}=\alpha_{q}^{81^{\Omega}}+\beta_{iq}^{81^{\Omega}}+\beta_{jq}^{81^{\Omega}}+\beta_{kq}^{81^{\Omega}}+\beta_{ijq}^{81^{\Omega}}+\beta_{ikq}^{81^{\Omega}}+\beta_{jkq}^{81^{\Omega}}+\beta_{ijkq}^{81^{\Omega}} \qquad (5.9)$$

and

$$Log\left(\frac{P_{q/ijk}^{90}}{P_{3/ijk}^{90}}\right)=\mathbf{X}_{ijkq}\boldsymbol{\beta}^{90^{\Omega}}=\alpha_{q}^{90^{\Omega}}+\beta_{iq}^{90^{\Omega}}+\beta_{jq}^{90^{\Omega}}+\beta_{kq}^{90^{\Omega}}+\beta_{ijq}^{90^{\Omega}}+\beta_{ikq}^{90^{\Omega}}+\beta_{jkq}^{90^{\Omega}}+\beta_{ijkq}^{90^{\Omega}} \qquad (5.10)$$

we can express the finite population proportions $P_{q/ijk}^{81}=M_{ijkq}^{81}/M_{ijk\cdot}^{81}$ and $P_{q/ijk}^{90}=M_{ijkq}^{90}/M_{ijk\cdot}^{90}$ as follows,

$$P_{q/ijk}^{81}=\frac{e^{\mathbf{X}_{ijkq}\boldsymbol{\beta}^{81^{\Omega}}}}{\sum_{q=1}^{Q}e^{\mathbf{X}_{ijkq}\boldsymbol{\beta}^{81^{\Omega}}}} \qquad (5.11)$$

and

$$P_{q/ijk}^{90}=\frac{e^{\mathbf{X}_{ijkq}\boldsymbol{\beta}^{90^{\Omega}}}}{\sum_{q=1}^{Q}e^{\mathbf{X}_{ijkq}\boldsymbol{\beta}^{90^{\Omega}}}} \qquad (5.12)$$

We recall that (5.9) and (5.10) are also estimators of the super-population model parameters,

$$Log\left(\frac{\pi_{q/ijk}^{81}}{\pi_{3/ijk}^{81}}\right)=\mathbf{X}_{ijkq}\boldsymbol{\beta}^{81} \quad \text{and} \quad Log\left(\frac{\pi_{q/ijk}^{90}}{\pi_{3/ijk}^{90}}\right)=\mathbf{X}_{ijkq}\boldsymbol{\beta}^{90}$$

In order to explore the characteristics of different model structures, we shall fit the following SPREE models estimating their parameters for 1990, using the Census data as described above:

SPREE (a) $$Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right)=\alpha^{90}_{q}+\beta^{90}_{iq}+\beta^{90}_{jq}+\beta^{90}_{kq}+\beta^{81^{\Omega}}_{ijq}+\beta^{81^{\Omega}}_{ikq}+\beta^{81^{\Omega}}_{jkq}+\beta^{81^{\Omega}}_{ijkq}$$

SPREE (b) $$Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right)=\alpha^{90}_{q}+\beta^{90}_{iq}+\beta^{90}_{jq}+\beta^{90}_{kq}+\beta^{90}_{ijq}+\beta^{81^{\Omega}}_{ikq}+\beta^{81^{\Omega}}_{jkq}+\beta^{81^{\Omega}}_{ijkq}$$

SPREE (c) $$Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right)=\alpha^{90}_{q}+\beta^{90}_{iq}+\beta^{90}_{jq}+\beta^{90}_{kq}+\beta^{90}_{ijq}+\beta^{90}_{ikq}+\beta^{81^{\Omega}}_{jkq}+\beta^{81^{\Omega}}_{ijkq}$$

SPREE (d) $$Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right)=\alpha^{90}_{q}+\beta^{90}_{iq}+\beta^{90}_{jq}+\beta^{90}_{kq}+\beta^{90}_{ijq}+\beta^{81^{\Omega}}_{ikq}+\beta^{90}_{jkq}+\beta^{81^{\Omega}}_{ijkq}$$

We will also fit the following Unsaturated SPREE models:

(a)-(b) $$Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right)=\alpha^{90}_{q}+\beta^{90}_{iq}+\beta^{90}_{jq}+\beta^{90}_{kq}+\beta^{81^{\Omega}}_{ijq}$$

(a)-(c) $$Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right)=\alpha^{90}_{q}+\beta^{90}_{iq}+\beta^{90}_{jq}+\beta^{90}_{kq}+\beta^{81^{\Omega}}_{ijq}+\beta^{81^{\Omega}}_{ikq}$$

(a)-(d) $$Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right)=\alpha^{90}_{q}+\beta^{90}_{iq}+\beta^{90}_{jq}+\beta^{90}_{kq}+\beta^{81^{\Omega}}_{ijq}+\beta^{81^{\Omega}}_{jkq}$$

(b)-(c) $$Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right)=\alpha^{90}_{q}+\beta^{90}_{iq}+\beta^{90}_{jq}+\beta^{90}_{kq}+\beta^{90}_{ijq}+\beta^{81^{\Omega}}_{ikq}$$

(b)-(d) $$Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right)=\alpha^{90}_{q}+\beta^{90}_{iq}+\beta^{90}_{jq}+\beta^{90}_{kq}+\beta^{90}_{ijq}+\beta^{81^{\Omega}}_{jkq}$$

(c)-(d) $$Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right)=\alpha^{90}_{q}+\beta^{90}_{iq}+\beta^{90}_{jq}+\beta^{90}_{kq}+\beta^{90}_{ijq}+\beta^{90}_{ikq}+\beta^{81^{\Omega}}_{jkq}$$

We compare these estimates with the target finite population 1990 counts, and with the following conventional Logistic models (no preservation of structure):

$$\text{(a)} \qquad Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right) = \alpha^{90}_{q} + \beta^{90}_{iq} + \beta^{90}_{jq} + \beta^{90}_{kq}$$

$$\text{(b)} \qquad Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right) = \alpha^{90}_{q} + \beta^{90}_{iq} + \beta^{90}_{jq} + \beta^{90}_{kq} + \beta^{90}_{ijq}$$

$$\text{(c)} \qquad Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right) = \alpha^{90}_{q} + \beta^{90}_{iq} + \beta^{90}_{jq} + \beta^{90}_{kq} + \beta^{90}_{ijq} + \beta^{90}_{ikq}$$

$$\text{(d)} \qquad Log\left(\frac{\pi^{90}_{q/ijk}}{\pi^{90}_{3/ijk}}\right) = \alpha^{90}_{q} + \beta^{90}_{iq} + \beta^{90}_{jq} + \beta^{90}_{kq} + \beta^{90}_{ijq} + \beta^{90}_{jkq}$$

In what follows we refer to the conventional Logistic model (a), the SPREE model (a) and the Unsaturated SPREE model (a)-(b), (a)-(c), (a)-(d), as "a-based models". Likewise, we refer to the conventional Logistic model (b), the SPREE model (b) and the Unsaturated SPREE model (b)-(c), (b)-(d), as "b-based models". Finally, the conventional Logistic model (c) and the SPREE model (c) are called "c-based models" whilst the conventional Logistic model (d) and the SPREE model (d) are called "d-based models".

Table 7.1 in the Appendix shows the Absolute Relative Bias of the proportion estimates $\tilde{P}^{90(m)}_{q/ijk} = \tilde{\pi}^{90(m)}_{q/ijk}$ for all the models *(m)* specified above. The Absolute Relative Bias (ARB) for the estimates generated by a model is given by,

$$ARB(\tilde{P}^{90(m)}_{q/ijk}) = \frac{\left|\tilde{P}^{90(m)}_{q/ijk} - P^{90}_{q/ijk}\right|}{P^{90}_{q/ijk}} \; x\,100 \qquad (5.13)$$

The ARB value represents the absolute relative bias that an estimator based on a particular model will have when applied to LFS sample in 1990. As we are interested

in the outcome of a multinomial variable, we also consider the subgroups averages of the ARBs of the three LFS proportion (Table 7.2), that is,

$$\overline{ARB_{ijk}}(\tilde{P}_{q/ijk}^{90^{(m)}}) = \frac{1}{3}\sum_{q=1}^{3} ARB(\tilde{P}_{q/ijk}^{90^{(m)}}) \tag{5.14}$$

Table 5.1 shows the aggregated differences $D_q^{(ms)-(ml)}$ in ARB between the SPREE estimates *(ms)* and the conventional Logistic estimates *(ml)* for each LF category *q=1,2,3* i.e. *q=Employee, Unemployed and Non-active.* The values in that table are given by,

$$D_q^{(ms)-(ml)} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left[ARB(\tilde{P}_{q/ijk}^{90^{(ms)}}) - ARB(\tilde{P}_{q/ijk}^{90^{(ml)}}) \right] \tag{5.15}$$

where the *(ms)-(ml)* combinations are:

(ms)	*(ml)*
SPREE (a)	*(a)*
SPREE (b)	*(b)*
SPREE (c)	*(c)*
SPREE (d)	*(d)*
USPREE (a)-(b)	*(a)*
USPREE (a)-(c)	*(a)*
USPREE (a)-(d)	*(a)*
USPREE (b)-(c)	*(b)*
USPREE (b)-(d)	*(b)*

Table 5.1
Venezuela - Aggregated Differences $D_q^{(ms)-(ml)}$
For 1990-1981 SPREE Models and Unsaturated SPREE

q (LF)	1990 Census Prop.	MODELS *(ms)*								
		SPREE				Unsaturated SPREE				
		(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
1	47.5	13.3	0.0	-1.3	2.3	14.4	14.4	14.3	1.0	-1.0
2	7.5	-6.7	4.8	1.7	6.2	-10.5	-8.5	-10.3	4.1	1.2
3	45.0	6.0	0.9	-0.9	1.7	6.1	6.6	5.6	2.1	-0.6

In general, we can see from table 5.1 that the conventional Logistic models recorded lower ARBs than the SPREE models (1990-1981).

We can also look at the proportion of subgroups (table 5.2) where a gain in ARB with respect to the conventional Logistic models was observed, that is, the proportion $PG_q^{(ms)-(ml)}$ of subgroups with $D_{ijkq}^{(ms)-(ml)} < 0$,

$$PG_q^{(ms)-(ml)} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} I\left(D_{ijkq}^{(ms)-(ml)}\right)}{184} \tag{5.16}$$

where,

$$I\left(D_{ijkq}^{(ms)-(ml)}\right) = \begin{cases} 1 & \text{if } D_{ijkq}^{(ms)-(ml)} < 0 \\ 0 & \text{otherwise} \end{cases}$$

We recall that $(I \cdot J \cdot K) = C = 184$.

Table 5.2 shows that despite the overall superiority of a-based models evident in Table 5.1, this improved performance comes from less than the 50% of subgroups. Furthermore, Table 7.1 in the appendix shows the irregularity of the behaviour of estimates from one subgroup to another when using different a-based models. For the rest of the models, i.e. b-based, c-based and d-based models it seems clear that not only does the use of SPREE and Unsaturated SPREE models not lead to an overall improvement in ARB but on the contrary, it actually worsen the overall outcome.

Table 5.2

Venezuela - Proportions $PG_q^{(ms)-(ml)}$ of subgroups with $D_{ijkq}^{(ms)-(ml)} < 0$
For 1990-1981 SPREE Models and Unsaturated SPREE Models.

q	Prop	MODELS								
		SPREE				Unsaturated SPREE				
		(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
1	47.5	16.8	52.7	57.6	37.0	9.2	17.9	12.0	42.4	61.4
2	7.5	48.4	31.5	41.3	29.3	45.7	48.4	46.2	33.7	37.0
3	45.0	39.1	50.0	60.9	38.6	40.2	35.3	40.2	35.9	61.4

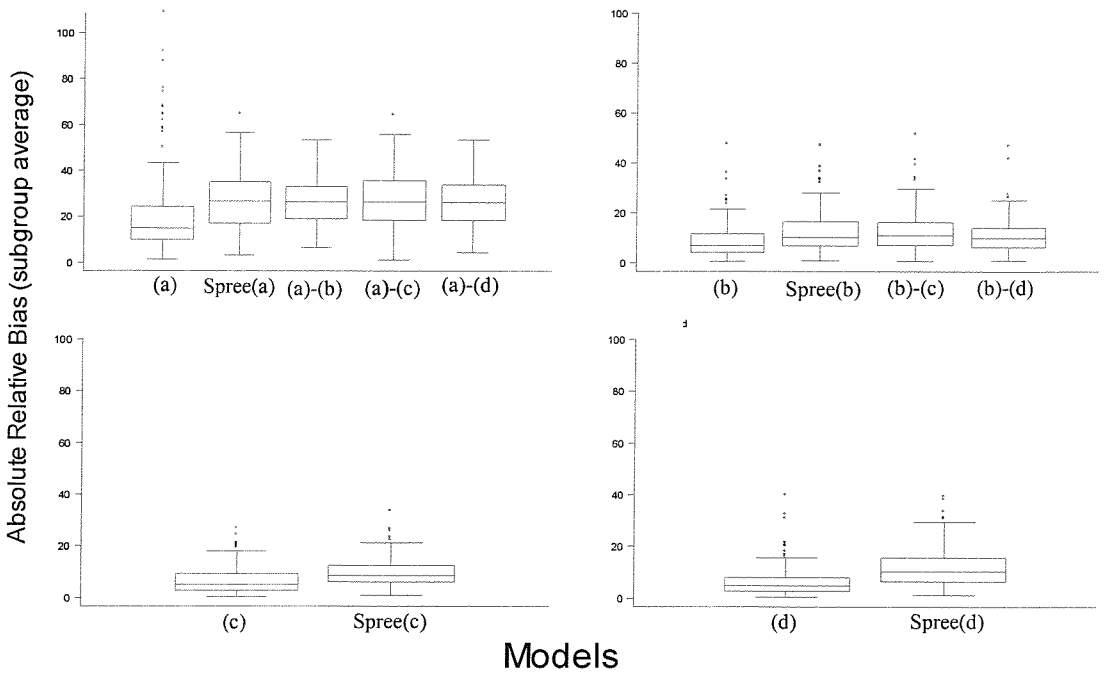

Figure 5.1. *Box plots of the subgroups* $\overline{ARB_{ijk}}$ *distribution, for conventional Logistic models, 1990-1981 SPREE Models and Unsaturated SPREE Models.*

A better picture of the differences between models is obtained from Figure 5.1 which shows box plots of the distribution of the subgroups $\overline{ARB_{ijk}}$ for each model. These are separated into four charts for a-based, b-based, c-based and d-based models. The graph clearly shows that using reference information with a 9 years gap is not only unnecessary but also inadequate. Unfortunately, we do not have the required data to carry out the same exercise with a smaller gap. Instead, later in this chapter we will carry out a simulation study with samples from the Census'90 in order to assess how these models perform in a census year.

If a good reference table for a particular model is available for the period of interest, i.e. the preservation assumptions are sufficiently acceptable, we can use the appropriate SPREE model so that the bias shown in Figure 5.1 is significantly reduced as we shall later show in the simulation study. In that case we are almost guaranteed superior quality estimates in terms of Mean Squared Error ($MSE = Variance + Bias^2$)

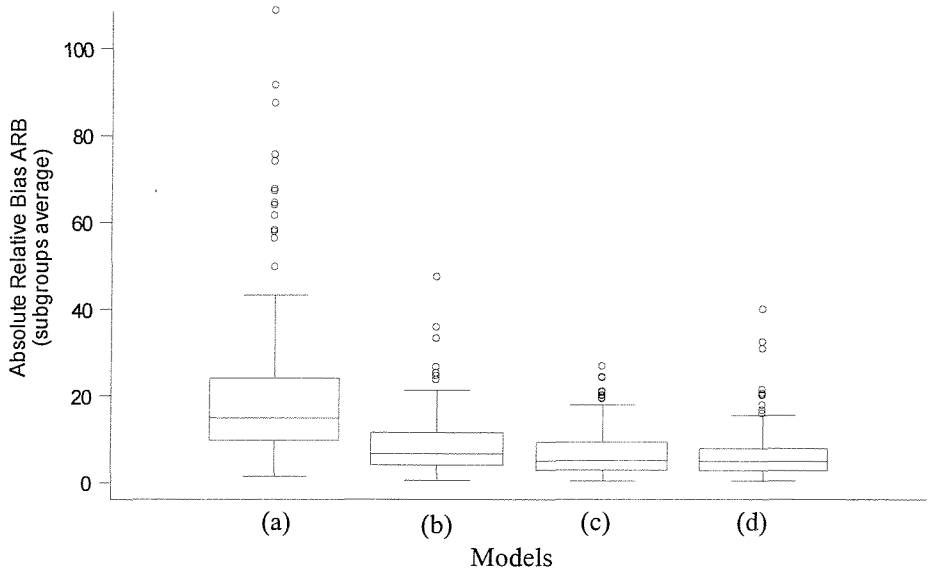

Figure 5.2 *Box Plots of the subgroups ARB average distribution, for conventional Logistic models (a), (b), (c) and (d).*

when using the sample from the LFS compared with the direct estimates obtained from this survey.

As an alternative procedure, should such a table not be available, we are interested in exploring possible gains from using conventional Logistic models like models (a), (b), (c) and (d). The idea is to assess these four models to understand the differences between them in terms of the bias of the corresponding model estimates This will also allow us to determine potential variants of these models that might provide a better fit.

Figure 5.2 shows the box plots of the distribution of the subgroups ARB average ($\overline{ARB_{ijk}}$) for the conventional Logistic models (a), (b), (c) and (d). As we expect, the more complex the model the lower its overall bias. However, the most important characteristic we can observe in Figure 5.2 is the appreciable impact that inclusion of the sex-age interaction has on the fitting process –model (b)-. That impact is not so dramatic when we add either the age-state interaction –model (c)- or the sex-state interaction –model (d)- to the sex-age interaction model (b).

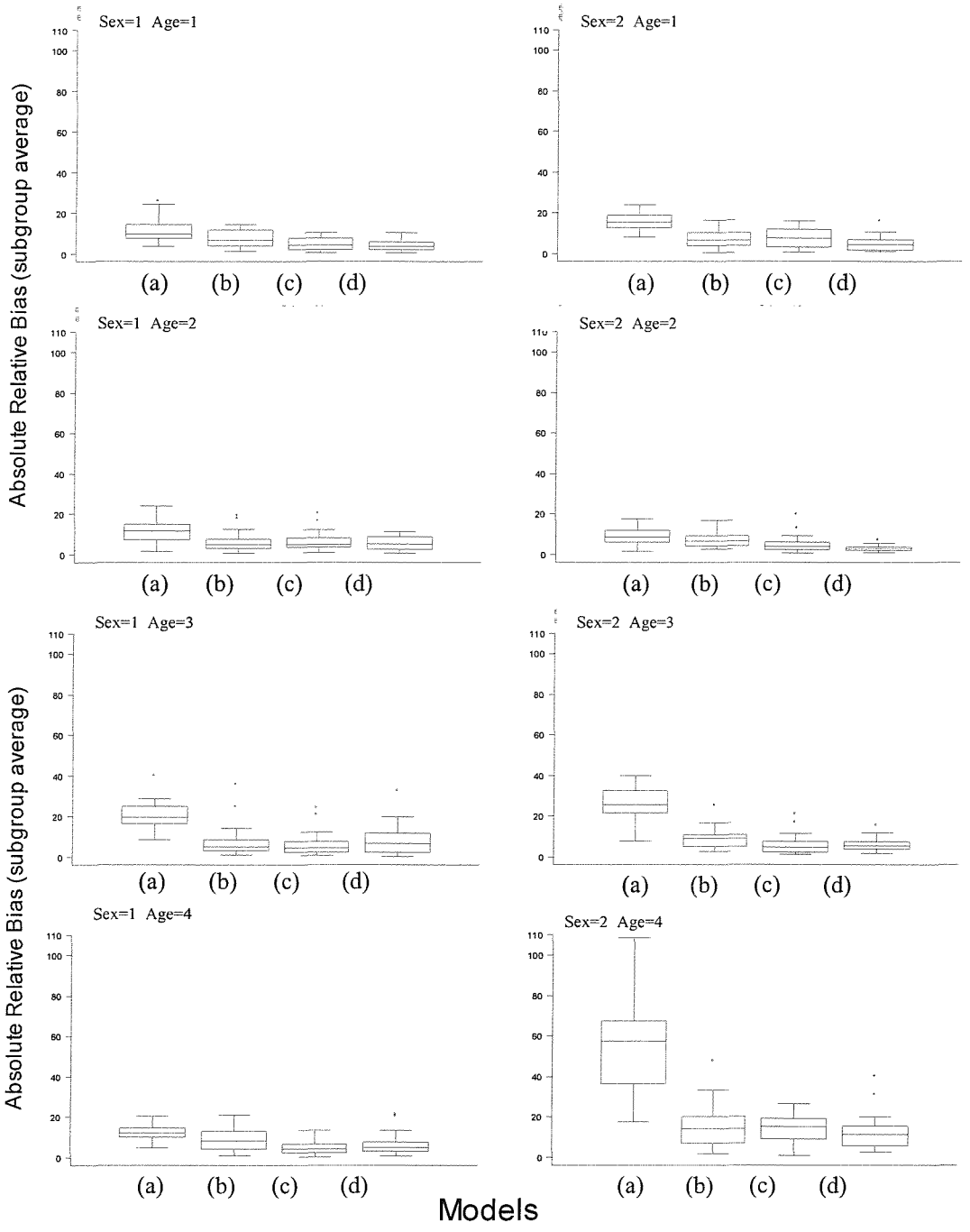

Figure 5.3 *Box plots of the distribution of the Subgroup aggregated ARB by Sex and Age, for conventional Logistic models (a), (b), (c) and (d).*

We know that the simpler the model the higher the precision with which its parameters and consequently the target estimates can be estimated. On the other hand, the simpler the model the higher the bias in these estimates. Therefore, the fact that there is no big difference in bias from model (b) to models (c) and (d) might mean that

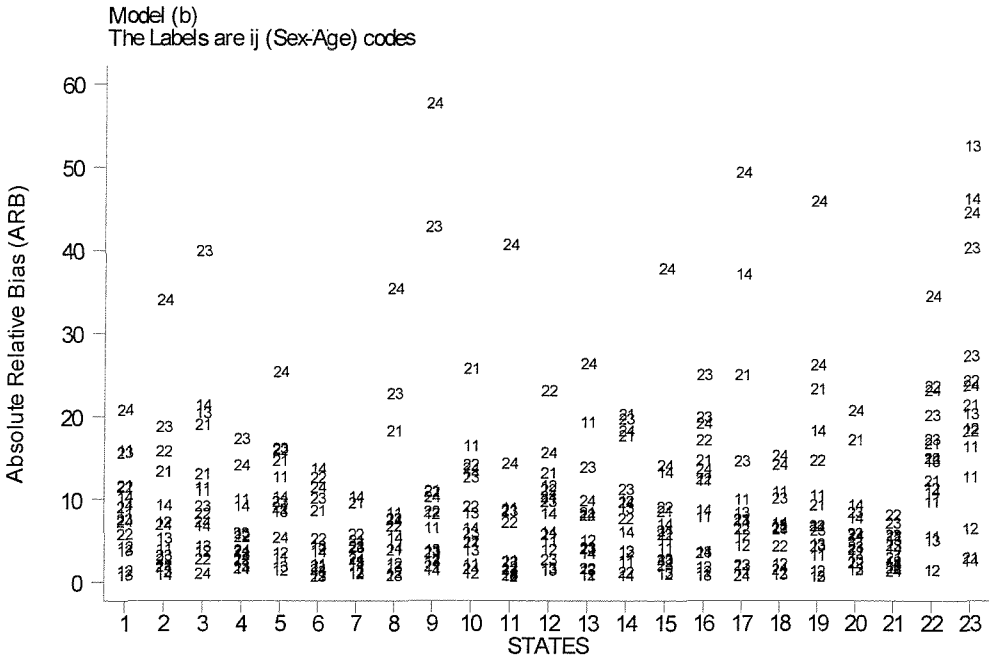

Figure 5.4 *ARB's for q=1 and q=2 i.e. Employees and Unemployed for Model (b) by States.*

the gain in precision with b-based models offsets its potential bias. This fact is important because it give us an insight into the potential superiority of one type of model among those considered so far. On the other hand, the a-based models show a big difference in ARB terms when compared to the b-based models so that a favourable bias-precision balance seems unlikely.

Figure 5.3 shows box plots of the distribution of the subgroup aggregates ARB (ARB_{ijk}) for the conventional Logistics models (a), (b), (c) and (d) by Sex and Age group. That figure shows that, in general, the pattern in Figure 5.2 is replicated for each Sex-Age group.

Figure 5.4 shows the ARB for Employees and Unemployed (*q=1,2*) for model (b) by States with the two digit labelling indicating Sex code and Age group code. We recall that the ARB figures are in fact absolute residuals from the model fitting process. From a visual inspection of this graph we can identify some subgroups for which the fit is fairly poor. They are mainly sub-groups related to Age 3 and Age 4, particularly when Sex equals 2. However, some of these points correspond to proportions lower

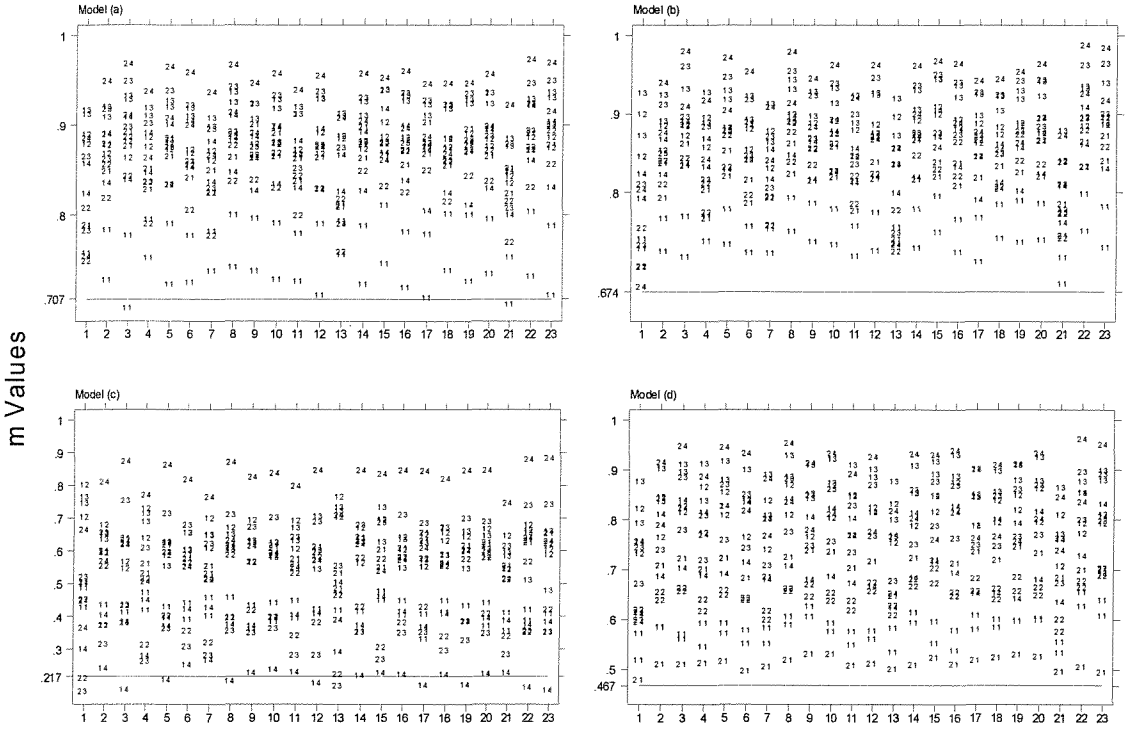

Figure 5.5 $m_{ijkq,ijkq}$ *values for q=1 and q=2 i.e. Employees and Unemployed for Modes (a), (b), (c) and (d) by States.*

than 0.02 for which an ARB of 60% might not be a problem. A similar pattern can be seen for models (a), (c) and (d).

However, as we have already discussed it in Chapter 4, ill-fitting points do not necessarily correspond to influential points. On the contrary, experience has shown that points that might influence the fitting process making its overall performance poorer are rarely ill-fitting points (Pregibon 1981). At this point, we are interested in the characteristics of potentially influential cells in order to learn more about the models and alternative variants that might improve the overall outcome.

In order to obtain a preliminary idea of the existence and characteristics of influential points we carry out a visual inspection of the $m_{ijkq,ijkq}$ values as was discussed in Section 4.2.5. Figure 5.5 shows four graphs corresponding to models (a), (b), (c) and (d) with the $m_{ijkq,ijkq}$ values for q=1,2 by States. In those graphs, the two digit labels

again indicate Sex code and Age group code. Each graph shows a horizontal line that indicates the rough cut-off point $\left[1-(2P/IJK(Q-1))\right]$ suggested by Hoaglin & Welch (1978) as a guide for determining potential influential points.

Only model (a) and model (c) register cells below the cut-off point. However, a pattern can be noticed in those graphs. For models (a), (b) and (d), the cells showing lower m-values are cells related to Age group 1. On the other hand, model (c) shows cells related to Age groups 3 and 4 with the lowest and highest m-values whilst Age groups 1 and 2 seem to be in between. It also seems important to mention the slightly different behaviour shown by States 1, 13 and 21. This is particularly noticeable for models (b) and (c). Those states are mainly urban and their socio-economic characteristics are certainly different from the other states, which may also affect the fitting process. An important fact about those states is that their sample size in the LFS is large and Sex-Age direct estimates are expected to be of a reasonable quality.

Taking into account those facts, we now define some variants of our models in order to explore their impact in the overall outcome. The description of the variants is as follows,

V1 -- No Age group 1
V2 -- No State 1
V3 -- No Age group 1 and State 1
V4 -- No States 1, 13 and 21

Note that those variants are equivalent to "adding" extra parameters to each model without having to include a further full interaction. For instance, fitting model (b) without Age group 1 is equivalent to having a new factor (say NewAge), with value 1 if Age is different from 1 and zero otherwise, interacting with the remaining factors in the model. That is, we are including an interaction Age-State but using the variable NewAge so that we add 22 extra parameters to the model instead of the 66 that would result from using the full Age-State interaction. Based on that fact, we observe that there is no need to fit variants 1 and 3 for model (c), for instance, as they do not add new information to the conventional Logistic model (c) and its variant 2 respectively.

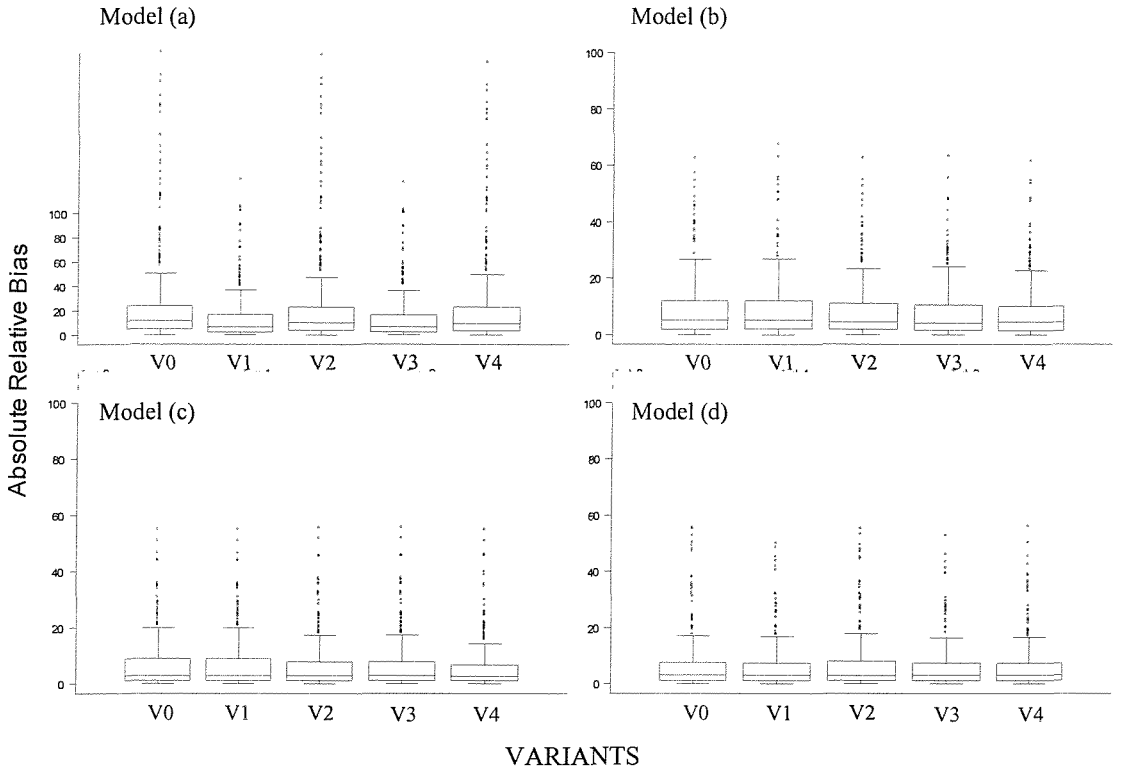

Figure 5.6 *Box plots of the subgroups ARB distribution (q=1,2), for variants V1, V2, V3 and V4 by Models (a), (b), (c) and (d).*

Figure 5.6 shows the box plots of the ARB for those variants along with their respective original models. We use "V0" in that graph to denote the respective original model without modifications. Due to comparative purposes those graphs do not contain information related to Age group 1 and states 1, 13 and 21.

For the a-based models we can see how deleting Age group 1 produce a noticeable impact in the overall performance pushing the box down. That impact is due to the fact that when we remove one Age group, we are actually "adding" extra Age-related interactions, including a Sex-Age interaction which has been shown to be an important one in the previous plots. However, the ARB levels still remain high. Deleting State 1 or States 1, 13 and 21 does not produce an appreciable overall impact for model (a). As for models (b), (c) and (d), there does not seem to be a clear advantage in using any of the variants; although slightly improvements can be seen in

some of them, it is unlikely that they will offset the increase in variability due to the extra parameters those variants suppose.

It is important to emphasise the interpretation of deleting categories from models (a), (b), (c) and (d). As we have already mentioned above, deleting categories is similar to adding interactions between a new partition of the variable subject to category removal and the rest of the variables. Analysing Figure 5.6 from that perspective, the relevance of the Sex-Age interaction is again evident. Once Sex-Age interaction is in the model, i.e. models (b), (c) and (d), adding extra interactions related to State does not have a major impact in the model.

Figure 5.7 shows Model (b) Unemployment ARB (q=2) against Unemployment proportions (in percentages figures) for subgroups by Sex-Age groups. Those graphs show the origin of the extreme ARB figures registered in previous charts; they are registered mainly in subgroups with proportions lower than 0.05, particularly in Sex=2 & Age=4, i.e. females aged 45 and more, where all proportions are lower than 0.02. These large ARB values for small proportions are expected given the "relative" nature of such a measure and they do not necessarily indicate bad estimates. The important fact here is that there are no extreme ARB values for moderate and large proportions.

This Census based comparative analysis has given us an idea of the potential usefulness of different possible models. The use of the SPREE method using a nine year old reference table does not seem to be a sensible approach. Traditional logistic analysis might offer an alternative to direct estimators, particularly using a model like model (b). However, a complete analysis of the appropriateness of these models as estimators of the subgroup proportions considered in this work needs to take into account the variance that the use of the LFS brings into the process; such an analysis as well as the study of SPREE methods using recent reference tables will be carried out next based on a simulation study.

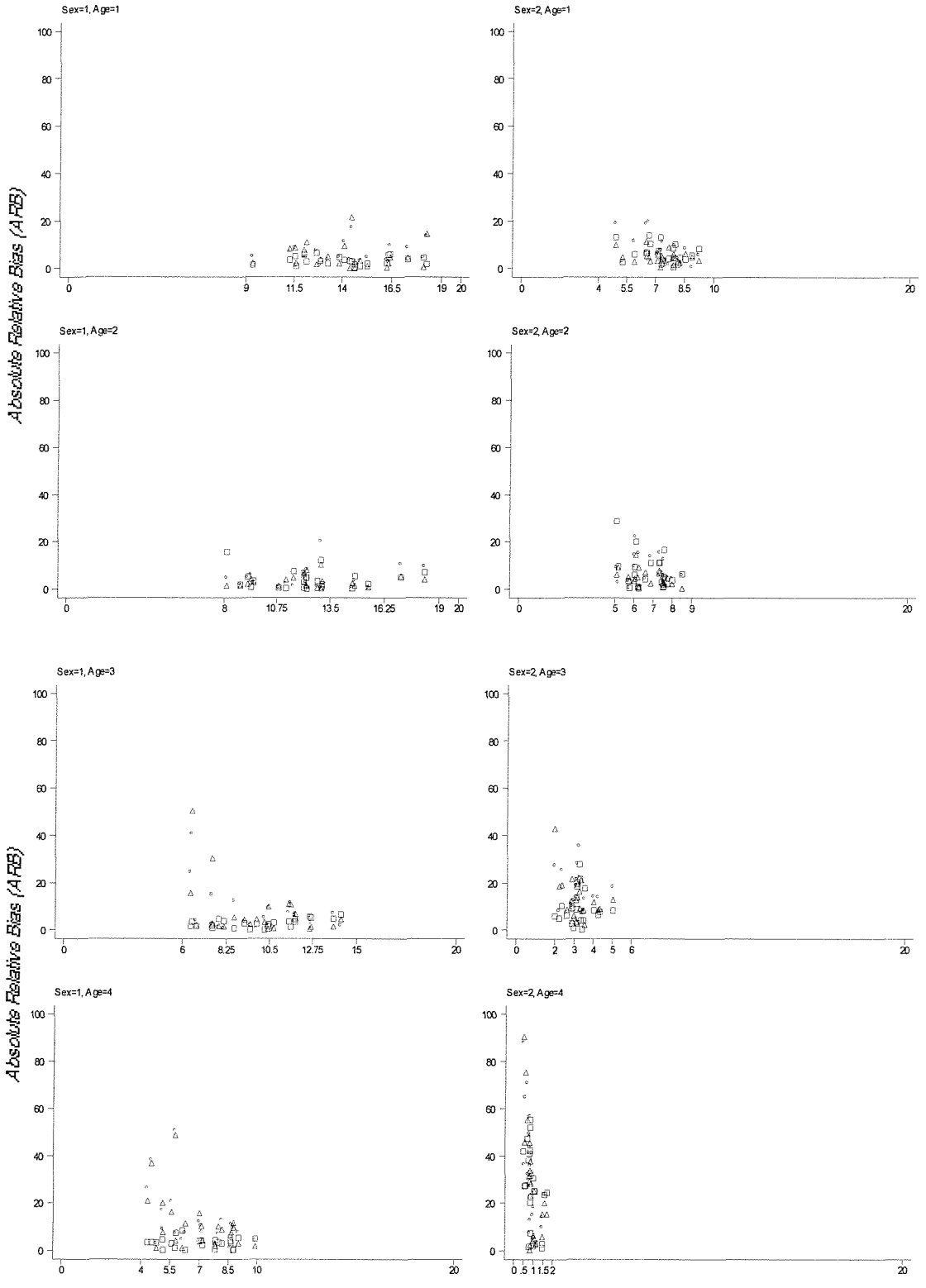

Figure 5.7 *Subgroups ARB Distribution (q=2) for conventional Logistic Models (b) against Parameter Proportion (Percentage figures).*

5.4. SIMULATION STUDY

In this section, we describe and analyse results from a simulation study based on the Venezuelan Census'90 data.

The primary goal of this simulation study is to asses the properties of the estimators and their variance estimators for the models proposed so far in this thesis. We have already mentioned the trade off between bias and variance between those models. From this simulation study we shall empirically learn about the characteristics of that trade off. Learning about the effect that each interaction term has on that process will give us an idea of what to expect from those models in different situations.

This study also includes SPREE models using the 1990 Census so that the characteristics of those procedures with an "adequate" reference table can be explored.

5.4.1. Description of the Simulation Study

5.4.1.a. Selection of the samples

The simulation study consisted of 1.000 samples selected from the Census'90 data following as close as possible the current sampling design of the Labour Force Survey (LFS). Technical details about the Census'90 and the LFS including details about their databases are given in Chapter 2. However, we now recall some key points that are relevant to the description of the simulation study.

The LFS sampling design is a stratified three-stage design. The LFS uses the Segments and the Sectors from the Census'90 as primary sampling units (PSU), creating special sub-divisions or sub-segments (of approx. 50 private addresses "PA" each) to be used as secondary sampling units (SSU). Finally, a sample of approximately five PA or tertiary sampling units (TSU) is selected within each sub-

segment in the sample. The Segments are stratified by geographical areas such as states or states divisions.

The Census'90 database used for the simulation study was the Expanded Database. This database comprises the information collected from the Census sample units using the Expanded Questionnaire. The Census'90 sample design was a stratified clusters design where approximately a 20% of clusters were sampled in the urban areas whilst 100% were included in rural areas. The Segments (of approx. 200 PA each) were used as strata in the urban areas and the "Sections" (partition of "Segments" of approx. 20 PA each) were the clusters.

The Expanded Database contains weights for each person and household in the sample. These weights are the adjusted weights after a complex post-stratification procedure involving ten different variables. Any attempt to withdraw samples from the Census using the Expanded Database has to take into account the Census sampling design. The original Census design weights were not available to us and their precise computation was not possible due to practical issues that arose during the Census execution. Each Segment was supposed to contain 10 Sections of approximately 20 PA each. However, some Segments ended up containing a larger number of PA than expected and so were divided into more than 10 Sections. As we have no access to the Basic Database, the actual number of Sections per Segment is unknown leaving us with no way of reproducing the original Census design weights.

To carry out the simulation we therefore considered the Census sample related to the urban areas as the actual Venezuelan urban data. For the rural areas, since they were completely enumerated in the Census, we selected a sample independently for each state of the same fraction as the one considered for urban areas, that is, a 20% sample within each state. This rural sample was then taken to be the actual Venezuelan rural data. Therefore in our simulation study, the combined 20% urban and rural sample is the target finite population and the parameters of interest are those that characterise this "synthetic" Venezuelan population.

Our aim was to select samples using a sampling design as similar as possible to the LFS sampling design. Therefore as in the LFS sampling design, the Census Segments were used as PSUs. We used states as the strata instead of sub-divisions of states. However, the information contained in the Expanded Database allowed us to sort the PSUs by geographical order within states introducing a stratification effect similar to the one present in the LFS sampling design which uses a systematic mechanism in the selection process. PSU or Segments were selected with probability proportional to the number of PA using a systematic mechanism. Table 5.3 shows the PSU sample size per State.

In the Urban areas, the sections comprising a segment were used as SSUs and their sub-sections as TSUs. One SSU per Segment sampled was selected with probability proportional to the number of PA. In the same way, one TSU per SSU sampled was selected also with probability proportional to the number of PA. Finally, all the PA within a selected TSU were considered as the final sample.

For the rural areas, the database does not contain any information that can be used as sub-divisions of Sectors. The LFS selects one SSU per PSU in rural areas, selecting ten consecutive PA from each SSU selected. For the simulation study, we selected ten consecutive PA with equal probability from each Sector in the sample. In a few cases the number of PA in a specific Sector was less than ten; In this case we included all its PA in the sample.

Table 5.3
Simulation Study PSU
Sample Size by State

STATE	PSU Sample	STATE	PSU Sample	STATE	PSU Sample	STATE	PSU Sample
1	261	7	81	13	212	19	31
2	66	8	11	14	34	20	26
3	24	9	50	15	16	21	315
4	74	10	34	16	32	22	15
5	28	11	75	17	52	23	20
6	115	12	44	18	58		

5.4.1.b. Selection Probabilities and Weights

Following the notation used in Chapter 2 but noting that our "Venezuela" is now the information in the database prepared for this simulation study, the probability of selecting the i*th* PSU within the h*th* stratum is $p1 = n_h(T_{hi} / T_h)$.

In the same way, let T_{hij} be the total of PA in the j*th* Census Section (SSU) of the i*th* Segment (PSU) in the h*th* stratum. The probability of selecting the j*th* Census Section within the i*th* urban PSU within the h*th* stratum is ${}_u p2 = (b_{hi}T_{hij} / B_{hi}t_{hi})$, where b_{hi} is the number of Sections selected for the census sample in the i*th* Segment in the h*th* stratum and B_{hi} is the total number of Sections in the same Segment and stratum.

Therefore the probability of selecting a PA in the i*ht* rural PSU within the h*th* stratum is ${}_r p2 = (c_{hi} / T_{hi})$, where c_{hi} is equal to ten, or T_{hi} if the Sector contains less than ten PA.

For the simulation study, the selection probability of any PA in the urban areas and in the rural areas respectively can be expressed as follows:

$$ {}_u p(PA_{hijk}) = n_h \left(\frac{T_{hi}}{T_h} \right) \left(\frac{b_{hi}T_{hij}}{B_{hi}t_{hi}} \right) \tag{5.17} $$

$$ {}_r p(PA_{hijk}) = n_h \left(\frac{T_{hi}}{T_h} \right) \left(\frac{c_{hi}}{T_{hi}} \right) = n_h \left(\frac{c_{hi}}{T_h} \right) \tag{5.18} $$

One problem is that we do not have any information about the values of B_{hi}. However, we know that the Sections were selected at random and that their sizes T_{hij} must be rather similar. Therefore, $(B_{hi}t_{hi}/b_{hi})$ must approximate T_{hi} and so we can use the following expression as a good approximation to ${}_u p(PA_{hijk})$:

$$ {}_u p(PA_{hijk}) \cong n_h \left(\frac{T_{hij}}{T_h} \right) $$

Noting that in urban areas T_{hij} is the sample size within a specific PSU, we have that in those areas $T_{hij}=c_{hi}$ and so we can write the common expression:

$$_u p(PA_{hijk}) = {}_r p(PA_{hijk}) = p(PA_{hijk}) = n_h\left(\frac{c_{hi}}{T_h}\right) \tag{5.19}$$

Therefore, for each sample in the simulation study, we have a set of weights w_{hi} attached to the units in the sample. Let the subscript g=1,...,1000 denote the g*th* sample of the simulation study. We shall denote the weights related to the g*th* sample as follows:

$$w_{hijk}(g) = w_{hi}(g) = \left(\frac{T_h}{n_h c_{hi}(g)}\right) = \left(\frac{T_s}{n_s c_{si}(g)}\right) \tag{5.20}$$

noting that h=s because the states are the strata in our simulation study.

5.4.1.c. Direct and Post-stratification Estimators of Parameters

Let us consider a variable Y related to people and let c=1,..,C denote the c*th* sub-group for which we require estimates, as denoted in Section 3.1.1. The Horwitz-Thompson estimator of the total Y and proportion P for the c*th* sub-group for the g*th* sample will be:

$$\hat{Y}_c(g) = \sum_{i=1}^{n_h} w_{hi} y_{hi,c}(g) \tag{5.21}$$

$$\hat{P}_c(g) = \frac{\hat{Y}_c(g)}{M_c} \tag{5.22}$$

where $y_{hi,c}(g)$ is the sample total for the c*th* sub-group within the *i*th PSU, *h*th stratum for the g*th* sample, and M_c is the census population total for the c*th* sub-group. The ratio Y/Z is estimated by:

$$\hat{R}_c(g) = \frac{\hat{Y}_c(g)}{\hat{Z}_c(g)} \tag{5.23}$$

Once a sample is drawn, a post-stratification adjustment is made at state levels using the Census sex-age counts at the same level of aggregation as it is currently done for the LFS (Section 3.2.4), that is A=22 sex-age post-strata. As in Section 3.2.4, the weights attached to the people in the sample are modified using this post-stratification adjustment. The resulting post-stratified weights are as follows:

$$_a w'_{hi}(g) = w_{hi}(g) \cdot \frac{_a M_s}{_a \hat{M}_s(g)} \tag{5.24}$$

where $_a M_s$ is the census population total for the a*th* post-stratum within the s*th* state and $_a \hat{M}_s(g)$ is its Horwitz-Thompson estimator of the form as (5.21) for the g*th* sample. The resulting estimators of the total Y, the proportion P and the ratio R=X/Z are then post-stratified (PS) estimators given by:

$$\hat{Y}_c^R(g) = \frac{\hat{Y}_c(g)}{\hat{M}_c(g)} \cdot M_c = \frac{\sum_{i=1}^{n_h} w_{hi}(g)\, y_{hi,c}(g)}{\sum_{i=1}^{n_h} w_{hi}(g)\, m_{hi,c}(g)} \cdot M_c \tag{5.25}$$

$$_y \hat{P}_c^R(g) = \frac{\hat{Y}_c^R(g)}{M_c} = \frac{\hat{Y}_c(g)}{\hat{M}_c(g)} \tag{5.26}$$

$$\hat{R}_c^R(g) = \frac{\hat{Y}_c^R(g)}{\hat{Z}_c^R(g)} = \frac{\hat{Y}_c(g)}{\hat{Z}_c(g)} = \frac{_y \hat{P}_c^R(g)}{_z \hat{P}_c(g)} \tag{5.27}$$

where $m_{hi,c}$ is the sample total of people for the c*th* sub-group within the *i*th PSU, *h*th stratum for the g*th* sample.

Let Y represents the total of the q*th* category of the labour force structure, that is $Y = M_{cq}$. Then, equations (5.21) and (5.25) are the direct and PS estimators respectively of the total of people in that category and the employment, unemployment, activity and inactivity rates ${}_qR_c$ are then estimated using (5.23) and (5.27).

5.4.1.d. Variance Estimator for the Direct and Post-stratification Estimators

Assuming that the PSUs were sampled with replacement within each stratum, we use the ultimate cluster technique (e.g. Kish 1965, Wolter 1985, Skinner et al. 1989, Sarndal et al. 1992) to estimate the variance and covariance of the Horwitz-Thompson estimator $\hat{Y}_c(g)$:

$$\hat{Var}\left(\hat{Y}_c(g)\right) = \frac{n_h}{(n_h - 1)} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi,c}(g) - \hat{\bar{Y}}_{h,c}(g)\right)^2 \tag{5.28}$$

$$\hat{Cov}\left(\hat{Y}_c(g), \hat{Z}_c(g)\right) = \frac{n_h}{(n_h - 1)} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi,c}(g) - \hat{\bar{Y}}_{h,c}(g)\right)\left(\hat{Z}_{hi,c}(g) - \hat{\bar{Z}}_{h,c}(g)\right) \tag{5.29}$$

The variances of $\hat{P}_c(g)$ and $\hat{R}_c(g)$ are:

$$\hat{Var}\left(\hat{P}_c(g)\right) \cong \frac{1}{M_c^2} \frac{n_h}{(n_h - 1)} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi,c}(g) - \hat{\bar{Y}}_{hi,c}(g)\right)^2 \tag{5.30}$$

$$\hat{Var}\left(\hat{R}_c(g)\right) \cong \frac{1}{\left(\hat{Z}_c(g)\right)^2} \left[\begin{array}{r} \hat{Var}\left(\hat{Y}_c(g)\right) - 2\hat{R}_c(g)\hat{Cov}\left(\hat{Y}_c(g), \hat{Z}_c(g)\right) \\ + \left(\hat{R}_c\right)^2 \hat{Var}\left(\hat{Z}_c(g)\right) \end{array}\right] \tag{5.31}$$

Applying the same arguments as in Section 2.2.5, an approximation to the variance and covariance of $\hat{Y}_c^R(g)$ is calculated by using the expressions:

$$\hat{Var}\left(\hat{Y}_c^R(g)\right) \cong \frac{n_h}{(n_h-1)} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi,c}^R(g) - \hat{\bar{Y}}_{hi,c}^R(g)\right)^2 \tag{5.32}$$

$$\hat{Cov}\left(\hat{Y}_c^R(g), \hat{Z}_c^R(g)\right) \cong \frac{n_h}{(n_h-1)} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi,c}^R(g) - \hat{\bar{Y}}_{hi,c}^R(g)\right)\left(\hat{Z}_{hi,c}^R(g) - \hat{\bar{Z}}_{hi,c}^R(g)\right) \tag{5.33}$$

Consequently, the expressions for the variances of $\hat{P}_c^R(g)$ and $\hat{R}_c^R(g)$ are:

$$\hat{Var}\left(\hat{P}_c^R(g)\right) \cong \frac{1}{M_c^2} \frac{n_h}{(n_h-1)} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi,c}^R(g) - \hat{\bar{Y}}_{hi,c}^R(g)\right)^2 \tag{5.34}$$

$$\begin{aligned} \hat{Var}\left(\hat{R}_c^R(g)\right) &\cong \frac{1}{\left(\hat{Z}_c^R(g)\right)^2} \left[\hat{Var}\left(\hat{Y}_c^R(g)\right) - 2\hat{R}_c^R(g)\, \hat{Cov}\left(\hat{Y}_c^R(g), \hat{Z}_c^R(g)\right) \right. \\ &\qquad \left. + \left(\hat{R}_c^R\right)^2 \hat{Var}\left(\hat{Z}_c^R(g)\right) \right] \\ &\cong \frac{1}{\left({}_z\hat{P}_c^R(g)\right)^2} \left[\hat{Var}\left({}_y\hat{P}_c^R(g)\right) - 2\hat{R}_c^R(g)\, \hat{Cov}\left({}_y\hat{P}_c^R(g), {}_z\hat{P}_c^R(g)\right) \right. \\ &\qquad \left. + \left(\hat{R}_c^R(g)\right)^2 \hat{Var}\left({}_z\hat{P}_c^R(g)\right) \right] \end{aligned} \tag{5.35}$$

5.4.1.e. Models considered in the Simulation

The models fitted to the data obtained in each of the 1000 samples are the SPREE models (a), (b), (c) and (d) and the Conventional Logistic models (a), (b), (c), (d) described in Section 5.3. However, there are two differences here. Firstly, we shall use the LFS-like samples to estimate the marginals used to update the required terms in the model. Secondly, the reference table used in the SPREE process will be the one defined by the Expanded Census'90 database from which the samples are drawn. This is the best possible scenario as any interaction term to be preserved from the reference

table is guaranteed to be equal to that of the target table. This is the equivalent to the "ideal" situation in which the preservation assumption flawlessly holds.

5.4.1.f. Estimators of Parameters

The 1000 sample estimates of the proportion of people in each of the three groups *q=1,2,3* that comprise the basic structure of the labour force were obtained for each model for each of the sub-groups *c=1,...,184,* that is,

$$\tilde{P}_{q/c} = \tilde{\pi}_{q/c} = \pi_{q/c}\left(\tilde{\theta}(g)\right) = \frac{e^{\mathbf{X}_{cq}\tilde{\theta}(g)}}{\sum_{q=1}^{3} e^{\mathbf{X}_{cq}\tilde{\theta}(g)}} = \frac{e^{\mathbf{X}_{cq}\tilde{\theta}(g)}}{1+\sum_{q=1}^{2} e^{\mathbf{X}_{cq}\tilde{\theta}(g)}} \tag{5.36}$$

Let $\tilde{\pi}_{12/c}(g)$ denote the sum $\tilde{\pi}_{1/c}(g)+\tilde{\pi}_{2/c}(g)$. The employment, unemployment, activity and non-activity rates ${}_qR_c$ are estimated by using the expressions,

$$\begin{aligned} {}_e\tilde{R}_c(g) &= \frac{\tilde{\pi}_{1/c}(g)}{\tilde{\pi}_{12/c}(g)} \quad ; \quad {}_u\tilde{R}_c(g) = \frac{\tilde{\pi}_{1/c}(g)}{\tilde{\pi}_{12/c}(g)} \\ {}_a\tilde{R}_c(g) &= 1-\tilde{\pi}_{3/c}(g) \quad ; \quad {}_n\tilde{R}_c(g) = \tilde{\pi}_{3/c}(g) \end{aligned} \tag{5.37}$$

5.4.1.g. Variance Estimators

Let $\boldsymbol{\pi}\left(\tilde{\boldsymbol{\theta}}(\mathbf{g})\right)$ be the IxJxKx(Q-1)=368 vector:

$$\boldsymbol{\pi}\left(\tilde{\boldsymbol{\theta}}(\mathbf{g})\right) = \begin{pmatrix} \boldsymbol{\pi}_1\left(\tilde{\boldsymbol{\theta}}(\mathbf{g})\right) \\ \vdots \\ \boldsymbol{\pi}_c\left(\tilde{\boldsymbol{\theta}}(\mathbf{g})\right) \\ \vdots \\ \boldsymbol{\pi}_C\left(\tilde{\boldsymbol{\theta}}(\mathbf{g})\right) \end{pmatrix} ; \qquad \text{with} \quad \boldsymbol{\pi}_c\left(\tilde{\boldsymbol{\theta}}(\mathbf{g})\right) = \begin{pmatrix} \pi_{1/ijk}\left(\tilde{\boldsymbol{\theta}}(\mathbf{g})\right) \\ \pi_{2/ijk}\left(\tilde{\boldsymbol{\theta}}(\mathbf{g})\right) \end{pmatrix}$$

We substitute (5.34) in its variance-covariance matrix form into the equation (4.46) in order to calculate the variance estimates of the matrix $\boldsymbol{\pi}\left(\tilde{\boldsymbol{\theta}}(\mathbf{g})\right)$, that is, the diagonal of the matrix:

$$\tilde{\mathbf{Cov}}\left(\boldsymbol{\pi}(\tilde{\boldsymbol{\theta}}(\mathbf{g}))\right) = \hat{\boldsymbol{\Delta}}(\mathbf{g})\mathbf{X}\left(\mathbf{X}^{t}\mathbf{D}\hat{\boldsymbol{\Delta}}(\mathbf{g})\mathbf{X}\right)^{-1}\left(\mathbf{X}^{t}\mathbf{D}\,\hat{\mathbf{Covar}}\left(\hat{\mathbf{P}}^{\mathbf{R}}(\mathbf{g})\right)\mathbf{D}\,\mathbf{X}\right)\left(\mathbf{X}^{t}\mathbf{D}\hat{\boldsymbol{\Delta}}(\mathbf{g})\mathbf{X}\right)^{-1}\mathbf{X}\hat{\boldsymbol{\Delta}}(\mathbf{g}) \tag{5.38}$$

where $\hat{\boldsymbol{\Delta}}(\mathbf{g}) = Blockdiag\left[\mathbf{Diag}\left(\boldsymbol{\pi}_{\mathbf{c}}(\mathbf{g})\right) - \boldsymbol{\pi}_{\mathbf{c}}(\mathbf{g})\boldsymbol{\pi}_{\mathbf{c}}(\mathbf{g})^{t}\right]$ and $\mathbf{X}$ and $\mathbf{D}$ are as in Section 4.5.2.

The estimates of the variances $\tilde{Var}\left({}_{a}\tilde{R}_{c}(g)\right) = \tilde{Var}\left({}_{n}\tilde{R}_{c}(g)\right) = \tilde{Var}\left(\tilde{\pi}_{3/c}(g)\right)$ are defined by the diagonal of (5.38). The estimates of the variances $\tilde{Var}\left({}_{e}\tilde{R}_{c}(g)\right)$ and $\hat{Var}\left({}_{u}\tilde{R}_{c}(g)\right)$ are as follows,

$$\tilde{Var}\left({}_{e}\tilde{R}_{c}(g)\right) \cong \frac{1}{\left(\tilde{\pi}_{12/c}(g)\right)^{2}}\left[\begin{array}{r}\tilde{Var}\left(\tilde{\pi}_{1/c}(g)\right) - 2\,{}_{e}\tilde{R}_{c}(g)\,\tilde{Covar}\left(\tilde{\pi}_{1/c}(g),\tilde{\pi}_{12/c}(g)\right)\\ +\left({}_{e}\tilde{R}_{c}(g)\right)^{2}\tilde{Var}\left(\tilde{\pi}_{12/c}(g)\right)\end{array}\right] \tag{5.39}$$

$$\tilde{Var}\left({}_{u}\tilde{R}_{c}(g)\right) \cong \frac{1}{\left(\tilde{\pi}_{12/c}(g)\right)^{2}}\left[\begin{array}{r}\tilde{Var}\left(\tilde{\pi}_{2/c}(g)\right) - 2\,{}_{u}\tilde{R}_{c}(g)\,\tilde{Covar}\left(\tilde{\pi}_{2/c}(g),\tilde{\pi}_{12/c}(g)\right)\\ +\left({}_{u}\tilde{R}_{c}(g)\right)^{2}\tilde{Var}\left(\tilde{\pi}_{12/c}(g)\right)\end{array}\right] \tag{5.40}$$

where:

$$\tilde{Var}\left(\tilde{\pi}_{12/c}(g)\right) = \tilde{Var}\left(\tilde{\pi}_{1/c}(g)\right) + \tilde{Var}\left(\tilde{\pi}_{2/c}(g)\right) + 2\left[\tilde{Covar}\left(\tilde{\pi}_{1/c}(g),\tilde{\pi}_{2/c}(g)\right)\right]$$

and,

$$\tilde{Covar}\left(\tilde{\pi}_{1/c}(g),\tilde{\pi}_{12/c}(g)\right) = \tilde{Var}\left(\tilde{\pi}_{1/c}(g)\right) + \tilde{Covar}\left(\tilde{\pi}_{1/c}(g),\tilde{\pi}_{2/c}(g)\right)$$

$$\tilde{Covar}\left(\tilde{\pi}_{2/c}(g),\tilde{\pi}_{12/c}(g)\right) = \tilde{Var}\left(\tilde{\pi}_{2/c}(g)\right) + \tilde{Covar}\left(\tilde{\pi}_{1/c}(g),\tilde{\pi}_{2/c}(g)\right)$$

5.5. PERFORMANCE INDICATORS

In this section we define the indicators that will be used to assess the properties of the different estimators considered in the simulation study i.e. the model estimators defined in Section 6.4.1.e plus the Direct and the PS estimators.

Let $\hat{\phi}(cq)$ be any of the estimators considered in this study for the c*th*-sub-population and q*th* LF category parameter $\phi(cq)$. This parameter can be either a proportion or one of the rates defined in the previous section. In the case of rates q will just denote which rate the parameter is referring to, i.e. Employment, Unemployment, Activity or Non-activity rates.

Let $\hat{\phi}(cq,g)$ denote a specific outcome of $\hat{\phi}(cq)$ generated at simulation g. Let $\hat{SE}_g\left(\hat{\phi}(cq)\right)$ and $\hat{CV}_g\left(\hat{\phi}(cq)\right)$ be the standard error estimate and the estimate of the coefficient of variation of the estimate $\hat{\phi}(cq)$ obtained from the gth simulation. An unbiased estimator for the expected value of $\hat{\phi}(cq)$, $E\left(\hat{\phi}(cq)\right)$ is given by:

$$\hat{E}\left(\hat{\phi}(cq)\right)=\sum_{g=1}^{1000}\frac{\hat{\phi}(cq,g)}{1000} \tag{5.41}$$

Therefore, an estimator of the bias of $\hat{\phi}(cq)$ is given by:

$$\hat{Bias}\left(\hat{\phi}(cq)\right)=\hat{E}\left(\hat{\phi}(cq)\right)-\phi(cq) \tag{5.42}$$

We need indicators that allow us to assess the performance of $\hat{\phi}(cq)$ as an estimator of the parameter $\phi(cq)$ and to compare it with other estimators of that parameter. The key properties of estimators we are interested in are bias, accuracy and confidence interval coverage. By assessing accuracy and bias we can learn about the usefulness of each estimator at estimating the target parameters and the gains obtained due to the

use of one estimator over another. On the other hand, we can use the information about coverage rates to assess the impact of not taking into account bias in the construction of confidence intervals.

Eight performance indicators for $\hat{\phi}(cq)$ are therefore defined as follows,

a) Absolute Relative Bias: $$ARB\left(\hat{\phi}(cq)\right)=\left|\frac{1}{1000}\cdot\sum_{g=1}^{1000}\left(\frac{\hat{\phi}(cq,g)}{\phi(cq)}-1\right)\right|$$

b) Relative Root MSE: $$RRMSE\left(\hat{\phi}(cq)\right)=\frac{\left[\frac{1}{1000}\cdot\sum_{g=1}^{1000}\left(\hat{\phi}(cq,g)-\phi(cq)\right)^2\right]^{1/2}}{\phi(cq)}$$

c) Coverage Rate: $$CR\left(\hat{\phi}(cq)\right)=\frac{1}{1000}\cdot\sum_{g=1}^{1000}i(\hat{\phi}(cq,g))\ ;$$

$$i(\hat{\phi}(cq,g))=\begin{cases}1 \text{ if } \hat{\phi}(cq)\in\left(\hat{\phi}(cq)\pm z(\alpha/2)\cdot\hat{SE}_g(\hat{\phi}(cq))\right)\\ 0 \text{ otherwise}\end{cases}$$

d) Average SE: $$SEA\left(\hat{\phi}(cq)\right)=\frac{1}{1000}\cdot\sum_{g=1}^{1000}\hat{SE}_g\left(\hat{\phi}(cq)\right)$$

e) Standard Error: $$SE\left(\hat{\phi}(cq)\right)=\left[\frac{1}{1000}\cdot\sum_{g=1}^{1000}\left(\hat{\phi}(cq,g)-\hat{E}\left(\hat{\phi}(cq)\right)\right)^2\right]^{1/2}$$

$$SE\left(\hat{\phi}(cq)\right)=\left[RRMSE^2\left(\hat{\phi}(cq)\right)-ARB^2\left(\hat{\phi}(cq)\right)\right]^{1/2}\cdot\phi(cq)$$

f) Coefficient of Variation: $$CV\left(\hat{\phi}(cq)\right)=SE\left(\hat{\phi}(cq)\right)/\phi(cq)$$

$$CV\left(\hat{\phi}(cq)\right)=\left[RRMSE^2\left(\hat{\phi}(cq)\right)-ARB^2\left(\hat{\phi}(cq)\right)\right]^{1/2}$$

g) Relative SE Bias: $$RSEB\left(\hat{\phi}(cq)\right)=\frac{1}{1000}\cdot\sum_{g=1}^{1000}\left(\frac{se_g\left(\hat{\phi}(cq)\right)-SE\left(\hat{\phi}(cq)\right)}{\phi(cq)}\right)$$

$$RSEB\left(\hat{\phi}(cq)\right)=\frac{SEA\left(\hat{\phi}(cq)\right)-SE\left(\hat{\phi}(cq)\right)}{\phi(cq)}$$

In the case of proportions, we are interested in the performance of the vector of binomial estimators $\hat{\boldsymbol{\varphi}}(\mathbf{c})=\left(\hat{\phi}(cq)\right)=\left(\hat{\phi}(c1),\hat{\phi}(c2),\hat{\phi}(c3)\right)$. Therefore the performance indicators to be analyzed will be the average of the three binomial indicators for each subgroup. For instance, the Absolute Relative Bias indicator we are interested in is given by,

$$ARB\left(\hat{\boldsymbol{\varphi}}(\mathbf{c})\right)=\frac{\sum_{q=1}^{Q}ARB\left(\hat{\phi}(cq)\right)}{Q}$$

In a similar way we obtain the following indicators $RRMSE\left(\hat{\boldsymbol{\varphi}}(\mathbf{c})\right)$, $SEA\left(\hat{\boldsymbol{\varphi}}(\mathbf{c})\right)$, $SE\left(\hat{\boldsymbol{\varphi}}(\mathbf{c})\right)$, $CV\left(\hat{\boldsymbol{\varphi}}(\mathbf{c})\right)$ and $RSEB\left(\hat{\boldsymbol{\varphi}}(\mathbf{c})\right)$.

Regarding the coverage indicator $CR\left(\hat{\boldsymbol{\varphi}}(\mathbf{c})\right)$, we use a different approach for proportions from the one given above. This is due to the fact that we are dealing with a multinomial variable and in this situation it is more informative to calculate an integral coverage rate, i.e. a rate indicating the percentage of samples for which the three confidence intervals simultaneously include their respective parameters. In this case we are interested in constructing simultaneous confidence intervals such that the combined coverage rate is about the $(1-\alpha)\%$ aimed. A common approach to construct those simultaneous confidence intervals is due to Goodman (1965). The approach consists to constructing all the Q confidence intervals -Q being the number of categories in the multinomial population- using $z(\alpha/2Q)$ instead of $z(\alpha/2)$ in the

formula for the confidence interval. The Coverage Rate indicator for proportion is therefore given by,

$$CR\left(\hat{\boldsymbol{\phi}}(\mathbf{c})\right) = \frac{1}{1000} \cdot \sum_{g=1}^{1000} i(\hat{\phi}(c,g)) \ ;$$

$$i(\hat{\phi}(c,g)) = \begin{cases} 1 \text{ if } \hat{\phi}(cq) \in \left(\hat{\phi}(cq) \pm z(\alpha/2Q) \cdot \hat{SE}_g(\hat{\phi}(cq))\right) \ \forall q \in c \\ 0 \text{ otherwise} \end{cases}$$

All these indicators are easily interpreted. The *ARB* indicator gives us a measure of the magnitude of the bias relative to the size of the parameter for a particular estimator.

The *RRMSE* can be interpreted in a similar way as the traditional coefficient of variation (*CV*), but noting that in this case the numerator is the mean squared error (*MSE*) instead of the variance. This is the standard indicator used to asses the accuracy of each estimator and thereby the gain or loss associated with using a particular model estimators. This indicator can be compared to the *CV* indicator in order to asses whether the coefficient of variation estimator $\hat{CV}_g$ is a useful measure of the accuracy of an estimator.

The *CR* indicator will allow us to assess the adequacy of confidence intervals for a particular estimator based on the standard error estimates, since this assumes unbiasness and normality of the estimator distribution. We will use $\alpha = 0.05$, so we are looking for values of this indicator that are significantly lower that 0.95 (or 95 in percentage figures).

The *CR* indicator can also alert us possible problems with the variance estimator of the unbiased Direct estimator and/or the almost unbiased Post-stratification estimator. The appropriateness of the standard error and coefficient of variation for estimating the real standard error and coefficient of variation can be assessed by looking at the *SEA, SE* and *RSEB* indicators. Large discrepancies between average and simulation

estimates of the standard error are indications of problems in precision measurement estimation from sample data. If the variance estimator formula is not adequate, we can check whether it is overestimating or underestimating the real variance by looking at the Relative SE Bias (*RSEB*). The *RSEB* indicator gives us a measure of the magnitude of the bias in the standard error estimator relative to the size of the parameter.

We also calculate some summary version of these indicators that can help provide a general picture of the characteristics and performance of the estimators. First of all, we will calculate the average of the seven performance indicators at different levels i.e. national averages, sex-age group averages and so on. To illustrate, the national average of the *ARB* indicator for proportions and rates estimators are respectively,

$$\bar{ARB}\left(\hat{\phi}\right)=\left(\frac{1}{184}\sum_{c=1}^{184}ARB\left(\hat{\phi}(cq)\right)\right)$$

and

$$\bar{ARB}\left(\hat{\phi}\right)=\left(\frac{1}{184}\sum_{c=1}^{184}ARB\left(\hat{\boldsymbol{\varphi}}(\mathbf{c})\right)\right)$$

The averages $\bar{RRMSE}\left(\hat{\phi}\right)$, $\bar{CR}\left(\hat{\phi}\right)$, $\bar{SEA}\left(\hat{\phi}\right)$, $\bar{SE}\left(\hat{\phi}\right)$, $\bar{CV}\left(\hat{\phi}\right)$ and $\bar{RSEB}\left(\hat{\phi}\right)$ are defined in the same way.

We also define some measures that will help us to understand the differences between estimators in terms of the percentage of subgroups satisfying a certain requirement regarding the level of a given indicator. These measures are given below for rates estimators. For proportion estimators we just have to substitute $\hat{\boldsymbol{\varphi}}(\mathbf{c})$ for $\hat{\phi}(cq)$ in any of these expressions.

a) Percentage of subgroups with a Relative Root MSE <= 0.15

$$\left[\%RRMSE\left(\hat{\phi}(cq)\right) < 0.15\right] = \left(\frac{1}{184}\sum_{c=1}^{184} i(cq)\right)\cdot 100;$$

$$i(cq) = \begin{cases} 1 \text{ if } \mathrm{RRMSE}\left(\hat{\phi}(cq)\right) < 0.15 \\ 0 \text{ otherwise} \end{cases}$$

b) Percentage of subgroups with a RRMSE Improvement > 0.15 with respect to the Post-Stratification estimators

$$\left[\%RIRRMSE\left(\hat{\phi}(cq)\right) > 0.15\right] = \left(\frac{1}{184}\sum_{c=1}^{184} i(cq)\right)\cdot 100$$

$$i(cq) = \begin{cases} 1 \text{ if } \mathrm{RC}\left(\hat{\phi}(cq), \hat{\phi}_{PS}(cq)\right) < -0.15 \\ 0 \text{ otherwise} \end{cases}$$

$$\mathrm{RC}\left(\hat{\phi}(cq), \hat{\phi}_{PS}(cq)\right) = \frac{\mathrm{RRMSE}\left(\hat{\phi}(cq)\right) - \mathrm{RRMSE}\left(\hat{\phi}_{PS}(cq)\right)}{\mathrm{RRMSE}\left(\hat{\phi}_{PS}(cq)\right)}$$

where $\hat{\phi}_{PS}(cq)$ denote a Post-Stratification estimator.

The %RRMSE indicator summarises the percentage of sub-populations for which the RRMSE yield by the simulation study is lower than a specific limit. We have chosen 0.15 as this limit on the basis that this figure is the "acceptable" limit for publication purposes. Therefore, the %RRMSE indicator will give us an idea of the percentage of "publishable" estimates when using each estimator considered in the study. However, it will not give us all the information about the gains or the improvement in RRMSE terms obtained by using a specific estimator with respect to any other estimator. For that purpose we have calculated the %RIRRMSE indicator.

The %RIRRMSE indicator gives us the percentage of sub-population groups for which the relative change in the RRMSE of a specific estimator with respect to the

RRMSE of the Post-stratification estimator $RC\left(\hat{\phi}(cq),\hat{\phi}_{PS}(cq)\right)$ is lower than a given number. Note that relative negative numbers in $RC\left(\hat{\phi}(cq),\hat{\phi}_{PS}(cq)\right)$ denote improvement in RRMSE terms of that estimator with respect to the PS estimator, whilst positive figures denote deterioration. We will use 0.15 as the limit.

These indicators will also be disaggregated by sex, age group, sex-age groups and state.

5.6. RESULTS

We now present and comment on the results obtained from the simulation study. In order to make a basic comparison to asses the appropriateness of the simulation process, Table 5.4 shows the estimated design based ARB from the simulation versus the "real" ARB calculated from the Census data for models (a), (b), (c) and (d) proportion estimators. Those figures are presented for Sex-age groups and along with the national average. The table does not show differences that might raise any concern.

Table 5.5 shows the national average of the performance indicators for the Direct, Post-stratification (PS), models (a), (b), (c) and (d) proportion estimators as well as the SPREE models (a), (b), (c) and (d) estimators. The most important aspect shown in this table is the quality of the SPREE estimators when the reference table is an

Table 5.4
PERFORMANCE INDICATOR ARB AND REAL ARB
FOR PROPORTION ESTIMATORS Model a, b, c and d
BY SEX AND AGE GROUPS and NATIONAL AVERAGE

Sex-Age Group		ARB				Sex-Age Group		ARB			
		(a)	(b)	(c)	(d)			(a)	(b)	(c)	(d)
Total	R	20.2	8.7	6.9	6.4						
	S	20.2	8.6	6.9	6.4						
1 1	R	11.8	7.6	4.9	4.0	2 1	R	15.9	7.0	7.9	5.0
	S	11.6	7.7	5.1	3.9		S	15.9	6.9	7.8	4.9
1 2	R	11.2	6.2	6.4	5.8	2 2	R	9.2	7.1	5.1	2.9
	S	11.3	6.2	6.5	5.7		S	9.1	7.1	5.1	3.2
1 3	R	20.3	7.5	6.0	8.2	2 3	R	25.9	9.4	6.0	6.3
	S	20.3	7.5	6.0	8.2		S	25.8	9.4	6.0	6.4
1 4	R	12.4	8.9	4.8	6.6	2 4	R	55.0	15.5	13.8	12.2
	S	12.4	8.9	5.0	6.6		S	55.3	15.2	13.5	12.0

R=Real ARB%
S=Estimated ARB% (simulation)

Table 5.5
PERFORMANCE INDICATORS FOR PROPORTION ESTIMATORS BY ESTIMATOR (Direct, PS, Model a, b, c and d and SPREE) NATIONAL AVERAGE

Performance Indicator	ESTIMATOR					
	Direct	PS	(a)	(b)	(c)	(d)
			Spree (a)	Spree (b)	Spree (c)	Spree (d)
CR	83.7	84.0	31.5	66.7	77.6	81.7
			92.2	92.0	88.0	92.9
ARB	0.8	0.9	20.2	8.6	6.9	6.4
			0.3	0.3	0.6	0.5
RRMSE	22.7	22.7	23.9	13.9	18.1	14.5
			8.8	9.2	15.2	11.8
SEA	3.7	3.7	1.8	1.8	2.9	2.4
			1.8	1.9	2.9	2.7
SE	3.5	3.5	1.7	1.7	2.7	2.1
			1.7	1.7	2.7	2.1
RSEB	-0.8	-0.7	0.9	0.8	0.7	1.0
			0.8	0.8	0.8	2.3
CV	22.6	22.6	10.2	9.6	15.5	12.1
			8.8	9.2	15.2	11.8

adequate one. We recall that for this simulation we construct the reference table from the 1990 Census data. Coverage rates for SPREE estimators are even superior to the Direct and PS estimators because they are not affected by the instability due to small sample sizes found in design based estimators. Note the dramatic reduction in bias caused by preserving appropriate interactions in contrast with the conventional Logistics models (a), (b), (c) and (d) and its obvious effect in reducing the RRMSE values.

As we expected, the standard error levels are rather similar for the SPREE and the conventional Logistic model estimators since the differences between them are the "constant" preserved interaction terms (see Chapter 4). The bias of the estimator of the standard error (RSEB) seems to be small and positive for the model estimators. For the design-based estimators, the national average shows a negative bias; we have

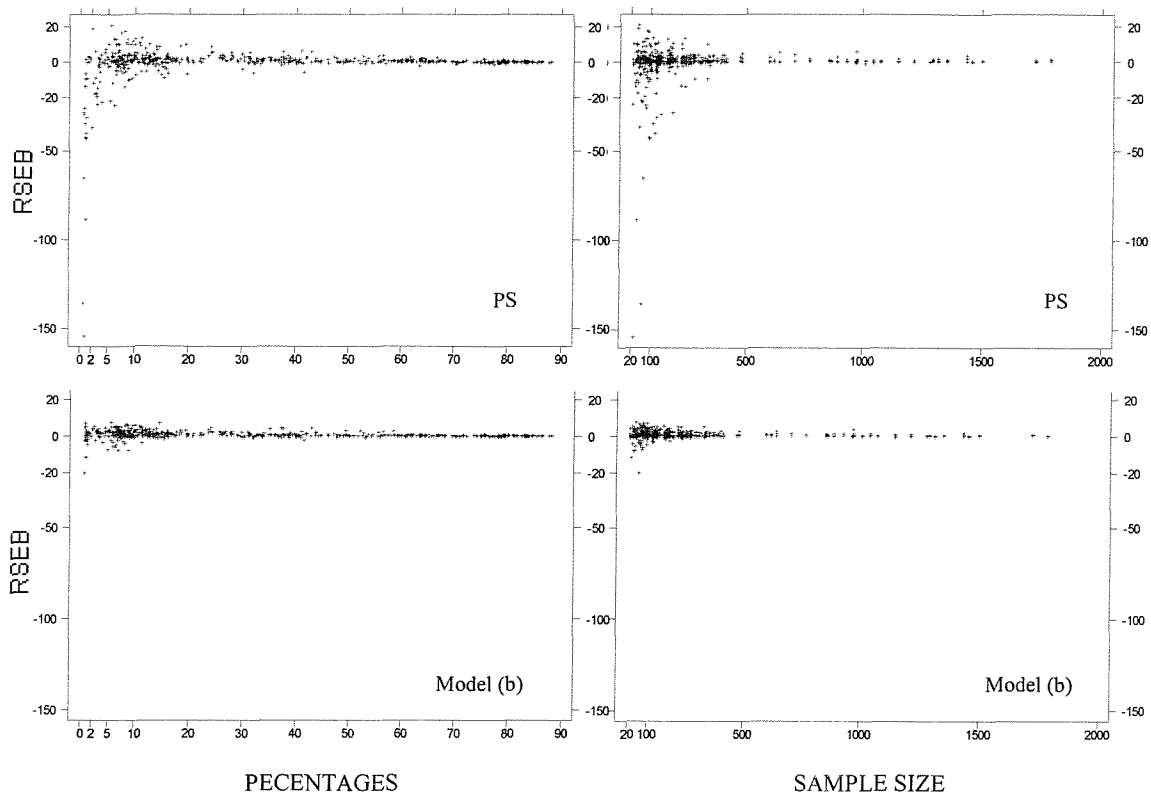

Figure 5.8 *Subgroups RSEB by Percentage Parameter and Sample Size for PS and Logit Model (b) Estimators.*

seen in Chapter 2 that the variance estimator of the design-based estimators tends to be conservative so a positive bias was expected. This national average bias is affected by two factors, the instability of the estimator of the SE in small sample size situations and the relative nature of the RSEB indicator itself, with the latter being the most important.

Figure 5.8 shows the subgroups RSEB against subgroups sample size and against the percentage parameters for the PS estimator and the Model (b) estimator. It can be seen that the bias of the SE estimator is rather stable and slightly positively biased for large sample sizes and large percentages. For combination of small sample size and small percentages the RSEB registers high values. All the extreme negative RSEB values shown in the graph correspond to subgroups with parameters lower than one percent. Also, the estimator of the SE is noticeable more stable in RSEB terms for model estimators.

Table 5.6
PERFORMANCE INDICATORS FOR PROPORTION ESTIMATORS FOR SEX-AGE GROUPS AND THE NATIONAL AVERAGE BY ESTIMATOR (Direct, PS, Model a, b, c and d)

Sex-Age Group	Coverage Rate (CR) %						Absolute Relative Bias (ARB) %						Relative Root MSE (RRMSE) %					
	Dir.	PS	(a)	(b)	(c)	(d)	Dir.	PS	(a)	(b)	(c)	(d)	Dir.	PS	(a)	(b)	(c)	(d)
Total	83.7	84.0	31.5	66.7	77.6	81.7	0.8	0.9	20.2	8.6	6.9	6.4	22.7	22.7	23.9	13.9	18.1	14.5
1 1	89.2	90.0	45.1	58.5	75.1	84.0	0.5	0.5	11.6	7.7	5.1	3.9	14.5	14.3	15.9	12.4	13.4	11.8
1 2	87.0	87.1	40.0	71.1	74.8	73.5	0.6	0.6	11.3	6.2	6.5	5.7	18.7	18.7	16.1	11.7	15.8	13.4
1 3	85.5	85.4	19.5	69.8	76.7	69.5	0.8	0.8	20.3	7.5	6.0	8.2	23.6	23.7	23.3	13.1	19.5	15.1
1 4	88.2	89.2	35.1	63.5	77.6	82.2	0.7	0.6	12.4	8.9	5.0	6.6	19.1	18.9	16.0	13.8	18.5	14.1
2 1	86.8	87.4	24.9	70.6	71.0	86.1	0.6	0.6	15.9	6.9	7.8	4.9	17.9	17.8	18.2	12.2	16.1	13.6
2 2	89.0	89.2	51.6	66.7	84.4	87.7	0.7	0.7	9.1	7.1	5.1	3.2	17.8	17.8	14.0	12.4	15.0	12.4
2 3	80.7	80.5	26.7	71.1	85.3	88.2	0.9	1.0	25.8	9.4	6.0	6.4	28.2	28.3	29.0	14.6	20.2	15.1
2 4	63.1	63.4	9.3	62.3	75.8	82.2	1.8	2.2	55.3	15.2	13.5	12.0	41.5	41.8	58.9	20.6	26.1	20.4

Although the superiority of the SPREE estimators with respect to the design-based estimators and the conventional Logistic model estimators is clear, we recall that this is true provide an appropriate reference table is being used. If this is not the case, we might consider the option offered by the conventional Logistic model estimators. On average, apart from Model (a), they all register a lower RRMSE than the design-based estimators. A disadvantage of these estimators is that their Coverage Rate is affected by their bias (see Table 5.5). Models (b) and (d) appear to be on average the best options among the four conventional Logistic models considered in the study; based on those average figures we would favour the Model (d) estimator as its RRMSE is not far from Model (b) RRMSE and its coverage rate (CR) is almost at the same level as the design-based estimators CR values. Note that the CV seems to be a better estimator of the RRMSE for Model (d) than for Model (b).

We now look at some disaggregated figures to see if the national averages in Table 5.5 reflect the behaviour of those estimators at different levels. Table 5.6 shows the Sex-age group average of the performance indicators for the Direct, Post-stratification (PS), models (a), (b), (c) and (d) proportion estimators. The Sex-age average figures seem to behave similarly to the national average figures. For group 2-4, i.e. females aged over 44, the figures shows particular poor performances for all the estimators.

Table 5.7
PERFORMANCE INDICATORS FOR PROPORTION ESTIMATORS FOR SEX-AGE GROUPS AND THE NATIONAL AVERAGE BY ESTIMATOR (Direct, PS, SPREE Model a, b, c and d)

Sex-Age Group	Coverage Rate (CR) %						Absolute Relative Bias (ARB) %						Relative Root MSE (RRMSE) %					
	Dir.	PS	(a)	(b)	(c)	(d)	Dir.	PS	(a)	(b)	(c)	(d)	Dir.	PS	(a)	(b)	(c)	(d)
Total	83.7	84.0	92.2	92.0	88.0	92.9	0.8	0.9	0.3	0.3	0.6	0.5	22.7	22.7	8.8	9.2	15.2	11.8
1 1	89.2	90.0	92.0	92.1	88.8	94.6	0.5	0.5	0.3	0.3	0.5	0.4	14.5	14.3	8.4	8.5	11.7	10.7
1 2	87.0	87.1	91.9	91.7	88.1	93.4	0.6	0.6	0.4	0.3	0.5	0.4	18.7	18.7	8.7	9.0	13.2	11.1
1 3	85.5	85.4	91.5	91.5	86.7	93.6	0.8	0.8	0.3	0.3	0.7	0.4	23.6	23.7	8.9	9.4	17.3	11.4
1 4	88.2	89.2	91.8	92.1	87.7	94.1	0.7	0.6	0.3	0.3	0.7	0.4	19.1	18.9	8.8	8.9	17.0	11.1
2 1	86.8	87.4	92.4	92.1	89.1	92.0	0.6	0.6	0.3	0.3	0.4	0.5	17.9	17.8	8.8	9.0	12.3	11.9
2 2	89.0	89.2	92.3	92.4	89.1	91.9	0.7	0.7	0.3	0.3	0.5	0.5	17.8	17.8	8.7	8.8	13.3	11.9
2 3	80.7	80.5	92.5	92.4	87.7	92.2	0.9	1.0	0.3	0.3	0.7	0.5	28.2	28.3	9.0	9.5	17.9	12.5
2 4	63.1	63.4	93.2	91.8	87.1	91.3	1.8	2.2	0.3	0.4	0.8	0.6	41.5	41.8	9.4	10.8	18.7	13.5

This is due to the fact that this group is the one with lower proportions and sample size levels. However, the important point to note in this case is that, apart from model (a), the model-based estimators have better RRMSE levels than the design-based estimators. Again, the model (d) estimator seems to be the best option.

Table 5.7 is the equivalent to Table 5.6 but for SPREE model estimators. Here again the behaviour is similar to the national average figures and the same comment regarding Sex-age group 2-4 applies. Since bias is not an important factor for these SPREE estimators, it is natural to choose as the best estimator the simplest model since its RRMSE levels are the lowest and its CR values the highest. This model is either the independence model (a) or the sex-age interaction model (b), depending on whether or not we need the sex-age group agreement to the national figures.

We arrive at the same conclusions for both the conventional Logistic estimators and the SPREE estimators after analysing the performance indicators by States. State level figures for conventional Logistic models are shown in Table 5.8.

Table 5.8
PERFORMANCE INDICATORS FOR PROPORTION ESTIMATORS FOR STATES AND THE NATIONAL AVERAGE BY ESTIMATOR (Direct, PS, Model a, b, c and d)

State Code	Coverage Rate (CR) %						Absolute Relative Bias (ARB) %						Relative Root MSE (RRMSE) %					
	Dir.	PS	(a)	(b)	(c)	(d)	Dir.	PS	(a)	(b)	(c)	(d)	Dir.	PS	(a)	(b)	(c)	(d)
Total	83.7	84.0	31.5	66.7	77.6	81.7	0.8	0.9	20.2	8.6	6.9	6.4	22.7	22.7	23.9	13.9	18.1	14.5
1	93.4	94.5	6.2	27.6	31.7	57.3	0.3	0.2	15.2	8.4	8.3	5.3	7.5	7.4	16.1	10.0	10.9	7.7
2	88.2	88.5	22.5	81.2	84.8	90.3	0.5	0.4	21.1	6.7	6.5	4.1	16.4	16.4	23.0	10.1	14.2	9.9
3	78.6	78.2	64.1	75.2	80.2	85.8	1.0	1.2	16.3	11.2	6.7	9.0	32.4	33.0	22.9	17.6	23.5	19.7
4	90.3	90.6	4.4	56.4	79.1	79.5	0.6	0.6	20.4	7.2	5.5	5.6	14.2	14.2	21.8	10.1	12.2	9.9
5	79.4	79.8	33.6	67.4	80.7	89.5	1.2	1.5	22.4	9.1	8.7	3.8	28.2	28.1	26.0	15.1	22.4	14.8
6	90.2	90.4	5.2	75.8	88.6	83.5	0.6	0.6	18.5	4.6	3.7	3.5	13.1	13.1	19.7	8.0	10.1	8.4
7	89.4	89.9	8.8	83.4	90.2	85.4	0.6	0.6	15.8	3.1	1.5	3.3	14.2	14.1	17.5	7.2	9.7	8.6
8	67.1	66.6	63.4	81.3	79.3	84.9	1.5	1.2	23.7	9.6	8.6	6.2	43.3	44.8	33.0	20.9	32.2	24.9
9	87.4	87.9	31.6	79.1	86.1	86.6	0.8	0.8	22.3	7.4	7.3	4.4	19.2	19.0	25.0	11.9	15.4	12.4
10	83.6	84.2	31.5	63.6	77.6	85.2	1.0	0.9	20.5	9.4	7.7	6.1	25.1	24.9	24.1	14.7	20.6	14.5
11	90.4	90.5	17.3	71.2	85.3	91.6	0.7	0.7	22.5	6.2	5.5	4.2	15.5	15.4	24.3	10.2	12.7	10.0
12	84.8	85.2	34.0	61.6	82.3	82.3	0.8	0.8	21.5	8.8	7.0	7.0	23.3	23.2	24.7	13.7	18.2	15.0
13	92.1	92.8	23.6	41.7	57.8	53.2	0.4	0.4	15.5	7.2	6.1	6.6	9.3	9.1	16.7	9.2	9.8	9.5
14	83.5	83.8	47.5	79.4	83.3	84.5	0.8	0.9	19.1	7.9	7.2	5.8	23.8	24.0	23.1	13.5	18.5	14.6
15	71.0	70.9	49.3	74.4	79.6	80.4	1.2	1.2	20.1	6.8	2.6	6.9	37.4	37.3	26.7	16.3	26.0	20.2
16	80.8	80.8	27.4	58.6	79.0	82.5	1.0	1.0	21.1	9.0	8.0	5.3	26.6	26.4	24.9	15.4	20.5	14.9
17	87.6	88.1	41.0	58.8	79.6	70.9	0.6	0.6	16.8	10.7	5.4	10.4	19.4	19.5	20.5	14.2	14.5	15.6
18	87.0	87.8	12.4	85.2	88.5	90.6	0.9	0.9	23.0	5.5	4.8	4.6	18.8	18.7	25.3	10.1	14.4	11.7
19	82.5	83.0	50.4	71.0	81.4	86.1	0.6	0.7	22.0	10.7	7.0	10.0	25.9	25.8	26.4	15.9	20.2	18.3
20	80.4	80.4	44.2	83.1	85.1	86.4	1.5	1.3	20.7	6.5	5.5	5.6	26.9	26.9	24.8	13.1	20.0	15.2
21	92.1	92.6	0.1	62.7	77.4	91.5	0.2	0.2	18.1	4.0	3.3	1.8	7.3	7.2	18.9	6.3	7.1	5.3
22	70.1	70.8	59.2	55.5	62.4	79.9	1.9	2.0	19.3	17.6	16.8	9.1	39.0	39.3	28.4	26.6	33.1	23.4
23	75.4	75.4	48.0	40.3	64.4	70.6	0.8	1.2	29.0	21.0	14.6	17.7	34.2	33.7	36.5	29.0	29.4	28.8

An interesting aspect shown by the tables analyzed so far is that there does not seem to be a big difference in RRMSE terms between the conventional Logistic estimators and the SPREE estimators based on models (c) and (d). This fact is important since it tells us that when the reference information available departs from an ideal case like the one we have created for our simulation study, SPREE estimators based on models (c) and (d) might rapidly become inferior to the conventional Logistic model (c) and (d) based estimators. However, the SPREE model (b) estimator might still be a good competitor.

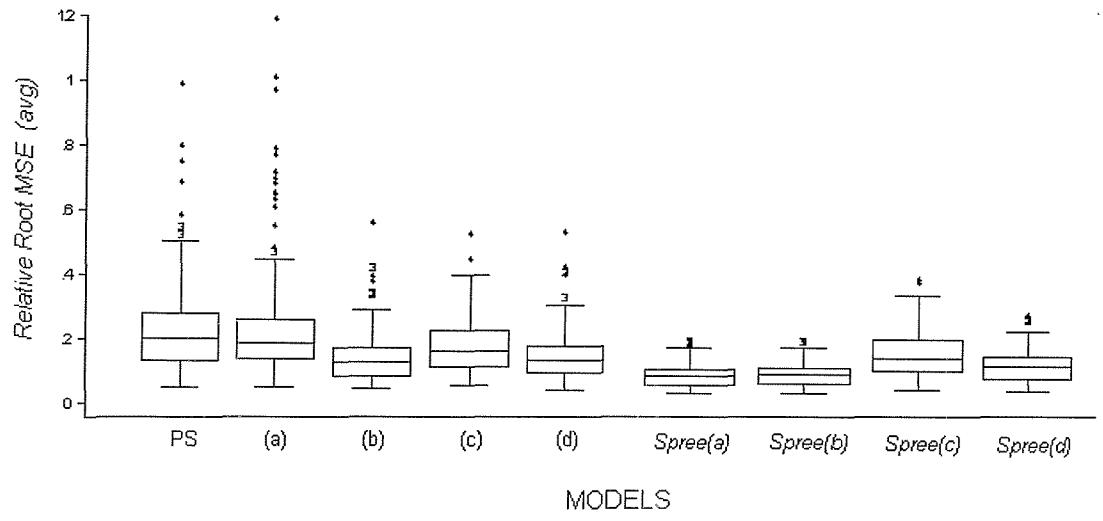

Figure 5.9 *Box Plots of the Subgroups RRMSE Distribution for the Direct, Post-Stratification (PS), Logistic and SPREE Models (a), (b), (c) and (d) Proportion Estimators.*

Figure 5.9 shows box plots for the subgroups RRMSE distribution for all the estimators in the simulation. Those plots confirm the comments made in this section regarding estimators and their RRMSE. Note, for instance, the similarity between the conventional Logistic models (c) and (d) estimators and the SPREE models (c) and (d) estimators.

Table 5.9 shows the percentage of subgroups with all three proportion estimates, i.e. Employees, Unemployed and Non-actives, registering a RRMSE lower than 0.15 (15%) at the same time. That gives us an idea of the proportion of subgroups for which is more likely to have stable rates estimates. The figures are presented by sex-age groups for the Direct, Post-stratification (PS), models (a), (b), (c) and (d) proportion estimators as well as the SPREE models (a), (b), (c) and (d) estimators. We can see again evidences of the gains obtained by using the model-based estimators. This time, the conventional Logistic Model (b) estimator seems to outperform the conventional Logistic Model (d) estimator by an appreciable margin, particularly for some sex-age subgroups (1-2, 1-3, 2-1). The SPREE estimators prove once more to

Table 5.9
PERCENTAGE OF SUBGROUPS WITH THREE PROPORTIONS RRMSE<0.15 FOR SEX-AGE GROUPS AND THE NATIONAL AVERAGE BY ESTIMATOR (Direct, PS, Model a, b, c and d and SPREE)

SEX-AGE GROUPS	ESTIMATOR					
	Direct	PS	(a)	(b)	(c)	(d)
			Spree (a)	Spree (b)	Spree (c)	Spree (d)
Total	20.7	20.1	18.5 78.8	52.2 75.5	32.1 37.5	42.4 57.1
1 1	43.5	43.5	43.5 82.6	73.9 82.6	60.9 60.9	73.9 78.3
1 2	26.1	21.7	26.1 82.6	65.2 82.6	39.1 52.2	47.8 78.3
1 3	13.0	13.0	4.3 78.3	65.2 78.3	34.8 39.1	43.5 73.9
1 4	17.4	17.4	13.0 78.3	56.5 78.3	21.7 21.7	52.2 73.9
2 1	30.4	30.4	4.3 78.3	56.5 78.3	43.5 47.8	39.1 43.5
2 2	21.7	21.7	52.2 78.3	60.9 78.3	34.8 43.5	43.5 43.5
2 3	13.0	13.0	4.3 78.3	26.1 73.9	17.4 21.7	26.1 34.8
2 4	0.0	0.0	0.0 100.0	0.0 100.0	0.0 100.0	0.0 100.0

be superior to their competitors, with the SPREE model (a) estimator registering the highest gains.

Table 5.10
PERFORMANCE INDICATORS FOR UNEMPLOYMENT RATE ESTIMATORS BY ESTIMATOR (Direct, PS, Model a, b, c and d and SPREE) NATIONAL AVERAGE

Performance Indicator	ESTIMATOR					
	Direct	PS	(a)	(b)	(c)	(d)
			Spree (a)	Spree (b)	Spree (c)	Spree (d)
CR	87.7	87.6	68.1 94.6	86.3 94.3	90.4 92.4	90.2 93.5
ARB	1.6	2.2	29.1 0.5	12.5 0.6	8.8 0.9	10.4 0.8
RRMSE	42.3	42.8	36.4 14.4	21.9 15.3	29.2 25.9	24.3 19.5
SEA	4.3	4.3	2.2 2.1	2.2 2.1	3.3 3.2	2.7 2.7
SE	4.7	4.8	2.1 2.0	2.1 2.1	3.3 3.3	2.7 2.7
RSEB	-7.1	-7.5	0.5 0.5	0.3 0.4	-0.5 -0.5	0.2 0.3
CV	42.3	42.7	17.5 14.4	16.4 15.3	26.8 25.9	20.8 19.5

We now present some tables and charts for estimated labour force rates. We shall concentrate mainly on Unemployment Rate (UR) indicators since estimation of these rates is basic indicator of the labour force.

Table 5.10 shows the national average of the performance indicators for the Direct, Post-stratification (PS), models (a), (b), (c) and (d) Rate estimators as well as the SPREE models (a), (b), (c) and (d) Rate estimators. As in Table 5.5, the most important result shown in this table is the quality of the SPREE estimators when the reference table is an adequate one. In general, the performance indicators for the Unemployment Rates show a similar behaviour to those of proportions. The SPREE estimators show lower RRMSE figures and higher CR values than their competitors

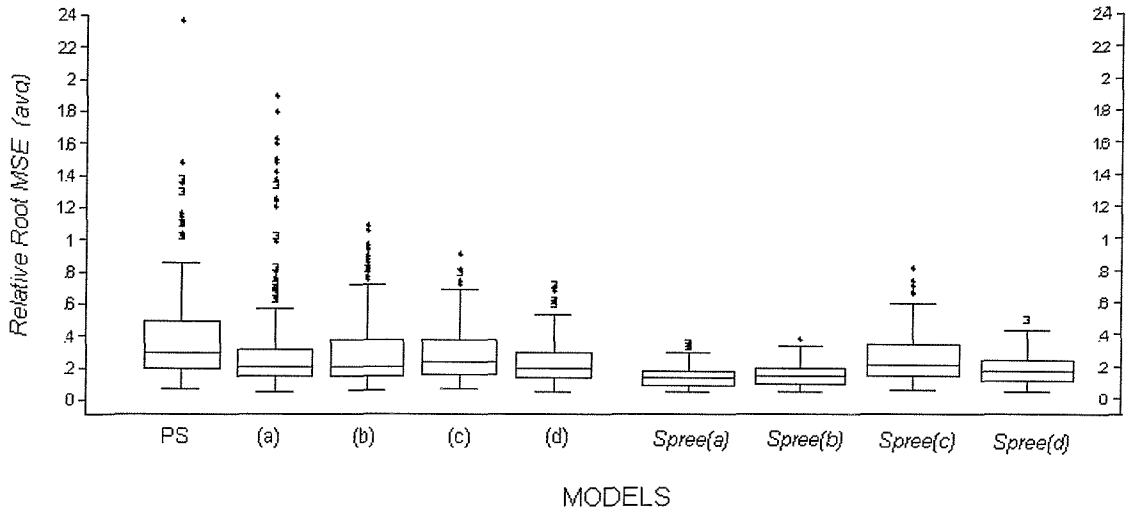

Figure 5.10 *Box Plots of the Subgroups RRMSE Distribution for the Direct, Post-Stratification (PS), Logistic and SPREE Models (a), (b), (c) and (d) Unemployment Rate Estimators.*

thanks to the lower ARB registered. As expected, the simpler the model the better the RRMSE figures for the SPREE estimators. As for the Conventional Logistic model estimators, the Model (b) estimator is the one with the lower RRMSE; in this case, however, the CR value for the Conventional Logistic Model (b) is similar to that of the design-based estimators and these values are even higher for the Conventional Logistic models (c) and (d). In this sense, it seems reasonable to choose the conventional Logistic model (b) estimator among the conventional Logistic estimators as its RRMSE is the lowest and its CV is of an acceptable level. However, Figure 5.10 shows how the conventional Logistic model (b) estimator, although having a lower national average for RRMSE than the conventional Logistic model (d) estimator, has more extreme RRMSE values than the model (b) estimator.

The RSEB figures show again a negative bias for the design-based estimators. The explanation in this case is the same as for the case of proportions discussed above. Subgroups with low unemployment rates and small sample sizes show a higher bias; two subgroups with a particular low UR push the national RSEB average downwards.

Table 5.11
PERFORMANCE INDICATORS FOR UNEMPLOYMENT RATE ESTIMATORS FOR SEX-AGE GROUPS AND THE NATIONAL AVERAGE BY ESTIMATOR (Direct, PS, Logistic Model a, b, c and d)

Sex-Age Group	Coverage Rate (CR) %						Absolute Relative Bias (ARB) %						Relative Root MSE (RRMSE) %					
	Dir.	PS	(a)	(b)	(c)	(d)	Dir.	PS	(a)	(b)	(c)	(d)	Dir.	PS	(a)	(b)	(c)	(d)
Total	87.7	87.6	68.1	86.3	90.4	90.2	1.6	2.2	29.1	12.5	8.8	10.4	42.3	42.8	36.4	21.9	29.2	24.3
1 1	93.3	93.4	87.7	87.7	92.9	91.4	0.6	0.7	8.5	8.1	4.4	5.1	20.5	20.6	15.9	15.4	18.6	15.5
1 2	93.0	92.9	91.4	91.1	91.7	93.0	0.6	0.8	5.5	5.2	5.2	4.0	24.1	24.0	15.5	15.3	20.9	16.5
1 3	91.7	91.6	73.7	92.6	91.2	92.5	0.7	0.8	12.8	6.4	3.6	7.0	32.3	32.7	19.7	17.2	28.8	19.4
1 4	91.3	91.3	66.0	87.4	91.0	88.0	1.3	1.2	14.8	11.9	3.6	12.6	35.0	35.8	22.6	22.1	33.8	24.2
2 1	91.0	91.2	82.3	85.6	87.8	89.1	1.2	1.3	9.0	8.1	9.1	6.9	26.6	26.3	15.6	15.2	19.6	19.0
2 2	92.0	91.9	81.1	84.2	89.8	91.0	1.0	1.5	11.0	10.8	9.3	4.6	30.5	30.7	18.6	19.1	23.5	21.7
2 3	83.3	83.1	49.1	83.8	90.6	89.8	2.7	2.9	46.1	17.5	12.5	15.4	61.8	62.1	52.3	27.3	36.7	32.2
2 4	66.2	65.4	13.3	78.2	87.8	86.7	4.9	8.4	125	32.1	22.7	28.0	108	110	131	43.2	51.5	46.1

The national average pattern of the performance indicators is similar to that of the disaggregated figures by sex-age and states, so, our conclusions are the same. As an example, Table 5.11 shows the Sex-age CR, ARB and RRMSE average indicators for the Direct, Post-stratification (PS), models (a), (b), (c) and (d) Rate estimators as well as the SPREE models (a), (b), (c) and (d) Rate estimators.

Table 5.12 shows the percentage of subgroups with an Unemployment Rate estimator RRMSE lower than 0.15 (15%) and a Relative Improvement of the RRMSE with respect to the PS estimator (RIRRMSE) higher than 15%. We have already seen that the model-based estimators offer significant improvements when compared to the design-based estimators in this regard. This fact is confirmed when looking at the RIRRMSE values, which show the model estimators producing relative reductions of the RRMSE values with respect to the design-based estimators for a high percentage of subgroups. On the other hand, only the 13.6% and the 13.0% of the subgroups have an RRMSE lower than 0.15 for the Direct and the PS estimators respectively. This figure rises to 39.7% for the Conventional Logistic model (b) and 53.3% for the SPREE model (b) estimator. Table 5.13 shows the same figures disaggregated by Sex-age groups. Although this again represents an improvement over to the design-based

Table 5.12
PERCENTAGE OF PERFORMANCE INDICATORS
FOR UNEMPLOYMENT RATE ESTIMATORS SATISFYING A GIVEN CONDITION
BY ESTIMATOR (Direct, PS, Models a, b, c and d and SPREE)

Performance Indicator	ESTIMATOR					
	Direct	PS	(a)	(b)	(c)	(d)
%RRMSE <=15	13.6	13.0	29.3 58.7	39.7 53.3	22.3 25.5	31.0 36.4
%RIRRMSE > 15%	0.0	---	66.3 100.0	88.6 100.0	52.2 64.7	90.8 98.9

estimators, it is also clear that none of these estimators offers publishable RRMSE level for almost half of the subgroups.

Based on this analysis, it therefore seems sensible to favour the use of the Conventional Logistic model (b) or (d) estimators in this simulation. The difference in RRMSE values between the Conventional Logistic and SPREE estimators for models (b) and (d) do not seem to be too large and, bearing in mind that in this study we are using an "ideal" reference table that is unlikely to be available in practice, the former estimators would appear to be a safe choice. It is also clear that SPREE estimators offer the best choice if a good reference table is available.

Table 5.13
PERCENTAGE OF PERFORMANCE INDICATORS
FOR UNEMPLOYMENT RATE ESTIMATORS SATISFYING A GIVEN CONDITION
BY ESTIMATOR (Direct, PS, Models a, b, c and d)

Sex-Age Group	% Relative Root MSE (%RRMSE) <=15						% Relative RRMSE Improvement (%RIRRMSE) > 15%				
	Direct	PS	(a)	(b)	(c)	(d)	Direct	(a)	(b)	(c)	(d)
Total	13.6	13.0	29.3	39.7	22.3	31.0	0.0	66.3	88.6	52.2	90.8
1 1	39.1	39.1	47.8	52.2	43.5	47.8	0.0	56.5	78.3	17.4	87.0
1 2	17.4	17.4	52.2	52.2	30.4	47.8	0.0	87.0	95.7	39.1	95.7
1 3	8.7	8.7	21.7	43.5	13.0	39.1	0.0	78.3	100.0	17.4	95.7
1 4	13.0	8.7	4.3	39.1	13.0	34.8	0.0	60.9	91.3	8.7	91.3
2 1	17.4	17.4	56.5	60.9	43.5	34.8	0.0	87.0	91.3	82.6	87.0
2 2	13.0	13.0	47.8	43.5	26.1	34.8	0.0	87.0	73.9	69.6	82.6
2 3	0.0	0.0	4.3	21.7	8.7	8.7	0.0	47.8	91.3	95.7	95.7
2 4	0.0	0.0	0.0	4.3	0.0	0.0	0.0	26.1	87.0	87.0	91.3

5.7. TIME AS AN EXTRA DIMENSION

So far in this thesis we have considered estimators that borrow strength from the two dimensions defining our "small areas" or sub-groups, i.e. space (horizontal) and characteristics of units (vertical). We have seen that SPREE works well when a good reference table is used (see Section 5.5). Therefore, we expect in practical situations a decrease in SPREE performance as the gap between the reference information and the target period increases. This suggests that incorporating information about the effect that "time" plays into the estimation process should be helpful.

Although the SPREE method uses information from a previous point in time, it does not "borrow" strength from that third dimension; instead such information is used without any intermediate processing so that the estimation process of borrowing information both horizontally and vertically takes place conditioned on it. However, since the LFS is a panel survey taking place twice a year (see Chapter 2), it seems sensible to explore the possibility of borrowing strength in time as an extra dimension. Recent work in borrowing strength over time has been investigated in the context of estimators based on specific area-level random effects models (see Section 1.5 for references); none of them apply to our specific case or to SPREE estimation. We now discuss some issues that need to be considered if we wish to extend SPREE by borrowing strength over time.

Consider a set of T tables containing the unknown counts M_{ct}, where $t = 1,...,T$ denotes a sequence of time periods and c denotes the lexicographical order of the cells $ijkq$. Let us suppose that LFS direct estimates for that set of tables are available although they are considered unreliable. Consider also a table containing the Census counts M_{c0}, that is, for time t=0. At time one (t=1), it seems reasonable to expect good estimates from SPREE using the Census as reference information. That expectation fades as we move further from t=0. At time T, SPREE might even be worse than the standard LFS direct estimates at t=T.

A naïve first approach would consist in using the SPREE-estimated table for t=1 as the reference table for SPREE at t=2 and so on until we get to the target time period t=T. However, we note that this process actually "preserves" interaction terms from the original Census table with the remaining terms updated using the sample at each time period. Consequently, since at each time period we preserve the same interaction terms from t=*0*, the final estimated table at t=T is the same as we would have obtained from applying SPREE directly using the time T sample and the t=*0* Census as the reference information, i.e. this naive approach and the conventional SPREE procedure used throughout this thesis are equivalent.

Different situations can arise depending on whether we choose to preserve different interaction terms at each time period. For instance, we might preserve at t=2 some terms that were updated at t=1, so that the outcome is a table with some structural terms preserved from the Census, some structural terms preserved from the previous time period estimated table and the remaining terms updated. Even in these situations this approach does not seem to offer a significant advantage over the simple SPREE methods considered in this thesis.

Consider now another set of tables containing the "known" set of counts $M_{tc}^{'}$, $t=(1-g),(2-g),...,(T-g)$ corresponding to T previous time periods; here g is a constant such that $t=(T-g)$ represents a point in time before the point in time represented by $t=1$ (see Figure 5.11).

We could treat each set of tables, i.e. $t=(1-g),...,(T-g)$ and t=*1,...,T*, as single five-dimensional tables with counts $M_{c}^{'}$ and M_{c} respectively, where c now denotes the lexicographical order of the cells *ijkqt*. In this situation, the former table provides us with valuable information regarding the structural changes of the "marginal" four-dimensional tables over time. Note, however, that the LFS is a panel survey with a specific rotation system and thus observations in the sample from different time periods can not be regarded as independent. This fact affects the assumption of independence required for the formulation of the model-based estimation procedures described in Chapter 4. The effect that this dependence has on the estimation process

as well as the potential extension to SPREE to allow a longitudinal log linear modelling approach remain to be investigated.

It should be pointed out that this approach requires an expanded reference table which, in our case, seems rather unrealistic. An important point is that this method assumes the preserved "time effects" to be the same for the target expanded table as for the expanded reference table. In our situation, this means that the gender, age group and state dynamics in time of the labour force follows the same pattern for the target period as it does for the reference period. This is a strong assumption that needs to be carefully checked.

Another approach to borrowing strength in time consists in testing for structures preservation over time. This can be done by applying SPREE independently to each of the T periods of time using the same reference information. If we find that the estimated parameters of the fitted models are not significantly different, then we might pool the datasets and apply SPREE to the pooled dataset. The resulting estimates should be of higher precision than the ones obtained from applying the four dimensional SPREE using just one dataset, as no extra parameters are being added to

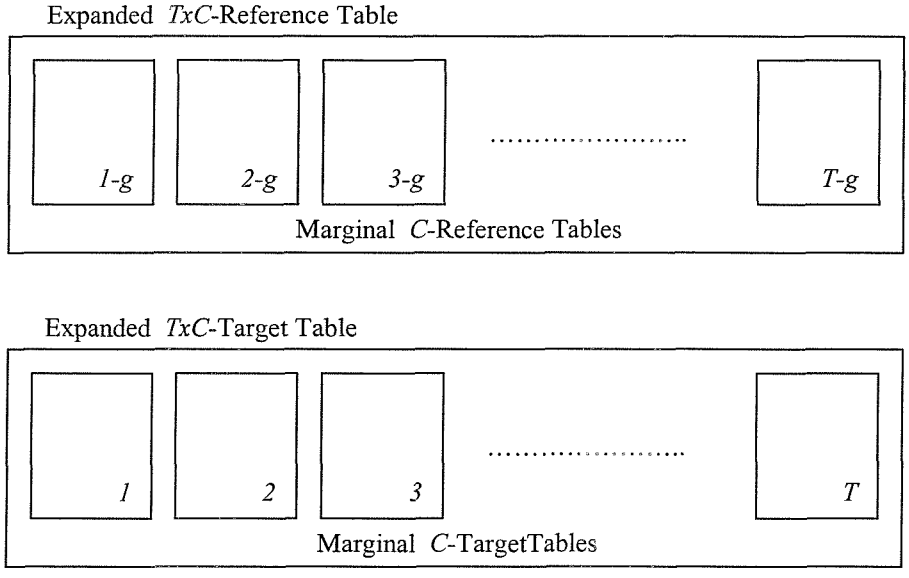

Figure 5.11 Representation of the Tables Involved in the SPREE Process with "Time" as an Extra Dimension.

the process and the sample size has increased. The reduction in variance will depend on the level of dependence between samples. Note that this approach is intended to improve the estimation of the terms being "updated" during the SPREE process and it does not address the issue of inadequate reference information. Note also that a similar procedure can be applied to conventional Logistic estimators where no reference information is used.

One way of testing testing preservation of structure consists in fitting separate SPREE-equivalent logistic models for each period of time and then testing the equality of the model parameters. Rao, Kumar and Roberts (1989) describe a similar procedure based on Rao-Scott corrections to Chi-Squared and Likelihood Ratio tests for an unsaturated logistic model for binary response and two periods of time. Its extension to SPREE equivalent logistic models for multinomial response is straightforward. Finally, we note that full preservation of structure becomes less likely as the gap between the earliest and the latest periods of time increases; therefore the approach described above seems to be more plausible when data are available for two consecutive time periods.

CHAPTER 6

CONCLUSIONS

This thesis has focused on the problem of producing reliable estimates of Employment, Unemployment and Activity rates by Sex-age groups for Venezuelan States using the Population Census as auxiliary information. From research we conducted in 15 national statistics offices in the Latin America region we observed that this is a common problem.

The SPREE approach to Small Area Estimation (Purcell and Kish 1980) is suited to dealing with this sort of small area estimation (SAE) problem. Although the use of SPREE methods in the SAE context has been treated in the literature, its use for the estimation of product multinomial variables as well as a general methodology for variance estimation were largely unexplored.

The methodology proposed in this thesis is based on the availability of two datasets, the Population Census database and the Labour Force Survey (LFS) database. An issue that must be taken into account when applying this methodology is that of the compatibility between those two sources as far as the conceptual definition of the labour force variables is concerned. A difference in such conceptual definition means that those two sources are in fact measuring different variables, which represents an obvious complicating factor in the analysis. Even if after examining such

compatibility we find there is a perfect conceptual match, we still have to considere how these variables are actually measured. Thus even though the two sources can share variables with the same conceptual definition, yet different methods of measurement can also complicate the analysis. This is the case in Venezuela where the labour force status for the Census is obtained by using one direct question in contrast to the algorithm based on a set of questions (Summary Code) used in the LFS. Thus the presence of such differences have to be examined carefully in order to decide if an adjustment strategy is necessary before applying the methodology describe in this thesis.

A detailed theoretical description of the Venezuelan LFS parameter and variance estimators did not exist. Therefore, such a theory was developed in this document from a model assisted perspective.

The use of SPREE in the context of product multinomial variables was then described. In doing so, we established a link between SPREE methods, Log-linear models and Logistic models allowing for the integration of complex sampling designs via a Pseudo-Likelihood approach to estimation. The main attractiveness of such a link is that it offers the possibility of implementing the SPREE method from a GLM perspective. This link has served as the base for the development of all the theory regarding SPREE and related SAE procedures described in this document..

There are some potential barriers to the convenient application of SPREE methods. To start, we note that SPREE involves the application of the Iterative Proportional Fitting (IPF) algorithm (Deming and Stephan 1940). Practical implementation of such a procedure then typically requires the development of "domestic" computational algorithms. Another issue related to the use of SPREE methods is that, apart from special situations, the computation of variance estimates is not obvious. In order to deal with this problem, we showed the equivalence of the Log-linear and Logistic version of SPREE to the well known "Exposure" technique from regression theory. This equivalence allows us to easily implement SPREE using standard commercial statistical software. Under this approach, the computation of parameter and variance estimates taking into account the sampling design is straightforward, as is the

computation of goodness of fit measures and related diagnostics. Consequently, a critical evaluation of the SPREE estimation procedure becomes feasible. On the whole, this "exposure" approach to SPREE facilitates and makes more flexible and accessible its application and practical implementation.

The "exposure" approach to SPREE was then used in an empirical analysis of the Venezuelan labour force. A first analysis was conducted using the Venezuelan 1990 Census as the information source for the target period and the Venezuelan 1981 Census as the reference source of information. This analysis showed that SPREE methods where not suitable given a nine year gap between the reference and target time period. We also compared the SPREE and the Unsaturated SPREE methods with conventional Logistic model-based estimation for this situation. The latter approach does not depend on the reference source of information, and performed creditably in our analysis.

We then used a simulation study to explore the application of SPREE using an "ideal" reference source. In this study the properties of a number of SPREE and conventional Logistic model-based estimators were analysed. The simulations were based on the Venezuelan 1990 population Census and were designed to replicate as closely as possible the LFS sampling design. In this case the gap between the reference and the target period used in the study was zero.

This study showed that the presence of a Sex-age interaction term in SAE models was significant for this data. In addition, the clear superiority of the SPREE method over design-based estimators and conventional Logistic model-based estimators was evident. However, it is important to point out that this superiority is based on the availability of "ideal" reference information; an unrealistic assumption for inter-censal periods. How fast the gains from this "ideal" situation fade as the gap between the reference information and the target period increases is a topic for future research.

The definition of an adequate gap size for the use of SPREE depends on both the variables under study and the dynamic of the process governing them. Between 1981 and 1990 Venezuela went through a delicate economical transition period. This fact

might have had an important impact on the structure of the labour force in the country, and hence directly affect the performance of SPREE. An idea of the magnitude of this sort of changes in the labour force structure can be obtained with the help of experts in the field. Examination of the data at the national level can also be considered; however, it is important to bear in mind that "local" changes can also have a significant impact on interaction structures.

Although SPREE methods proved superior to conventional Logistic model-based estimators in the simulation study, the difference between them was not as large as we might have expected; this is true particularly for models that are calibrated to national Sex-age group totals, i.e. models different from the independence model. This fact led us to recommend conventional logistic model-based estimators as favourable alternatives to SPREE-based estimators, particularly in situations when there is a reasonable doubt about the quality of the available reference information.

The logistic model containing only sex-age interaction (model b) and the logistic model containing sex-age interaction and sex-state interactions (model d) showed the best performance among the different model structures considered in this study. Model b seemed to be the best overall option.

Although the use of SPREE and conventional Logistic model-based estimators produce an important improvement over the design-based estimators, we found that they do not as yet offer "publishable" precision levels for almost half of the target subgroups. This fact led us to briefly discuss alternative methods involving the use of "time" as an extra dimension. We expect this approach to offer higher levels of precision than approaches that ignore changes over time particularly for the case of the conventional Logistic model-based estimators. Such an analysis represents future research.

Throughout the empirical analysis in this thesis, we have assumed that the process behind the generation of the finite population follows a product multinomial distribution with $E(M_{ijkq}) = M_{ijk\bullet}\pi_{q/ijk}$ and $V(M_{ijkq}) = M_{ijk\bullet}\pi_{q/ijk}(1-\pi_{q/ijk})$. A deeper examination of the structure of the population might suggest a better specification of

this distributional assumption, leading to an improvement in the estimation process. This is another topic for further work.

Synthetic estimators are more convenient, in terms of theoretical and practical simplicity, than specific area-level random effects model-based methods. This fact is more evident in national statistics offices with scarce highly specialized human resources. However, alternative options for logistic methods involving specific area-level random effects models need to be studied and compared to the simpler synthetic competitors. Important work in this direction has been conducted (e.g. Farrel et al. 1989, Malec et al. 1997, Gosh et al. 1998, Jian et al. 1999). The adaptation and extension of those works to our particular situation and an appropriate comparative analysis is an important task that remains to be undertaken.

Finally, it is important to point out that all our comments so far have been related to the SAE problem from an "estimation" perspective. However, it is just as important to conduct research aimed at exploring other approaches as institutional related strategies and survey design. It is vital that national statistics offices in Latin-America start playing a leading role in the development of the national statistics systems in their respective countries. The lack of auxiliary information is a task of prime importance and its negative impact in the national statistics system cannot be ignored. As regards to survey design, different issues like cluster structure, strata conformation and sample sizes should be considered and analysed in order to minimize the need for indirect estimators.

APPENDIX

Table 7.1

Venezuela

1990 Census Proportions and Absolute Relative Residuals (ARB)
for Thirteen Different Models

Categories					Census Prop.	MODELS												
										SPREE				Unsaturated SPREE				
SG	STT	Sex	Age	LF		(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
1	1	1	1	1	49.9	20.7	13.0	9.8	2.6	6.3	7.9	9.3	3.0	3.5	3.1	1.7	10.4	16.6
	1	1	1	2	12.7	18.8	7.3	6.5	1.7	14.2	9.1	6.8	16.0	15.0	15.7	15.5	9.9	9.8
	1	1	1	3	37.4	33.9	19.8	15.3	2.9	13.2	7.5	10.1	9.4	0.4	9.5	3.0	10.5	18.8
2	1	1	2	1	83.0	2.3	0.0	1.9	2.5	1.7	2.1	3.9	0.1	1.8	1.6	2.2	2.1	1.6
	1	1	2	2	9.4	5.3	5.5	0.7	4.2	34.0	5.8	9.7	6.9	22.3	35.6	23.2	8.6	1.1
	1	1	2	3	7.6	18.5	6.7	22.2	21.9	23.1	30.4	30.9	9.2	8.2	26.4	5.2	33.6	16.1
3	1	1	3	1	86.0	1.4	0.2	0.6	1.8	7.5	1.6	3.1	0.2	7.6	7.4	7.4	1.2	1.8
	1	1	3	2	8.1	15.9	0.5	3.7	1.4	26.0	6.1	1.7	5.2	24.8	25.3	24.0	4.4	2.4
	1	1	3	3	5.9	42.9	2.5	13.4	27.5	74.3	14.6	42.4	10.2	76.4	73.3	75.3	11.3	22.8
4	1	1	4	1	67.0	8.5	9.6	6.3	2.7	30.6	17.6	8.7	11.4	26.2	31.0	25.7	18.5	10.9
	1	1	4	2	6.1	13.3	4.4	8.5	1.2	27.5	25.5	7.3	27.3	0.0	21.8	0.6	18.0	2.9
	1	1	4	3	26.9	18.1	25.1	17.6	6.5	70.2	38.0	20.0	22.1	65.4	72.3	64.2	42.1	28.0
5	1	2	1	1	31.7	26.9	9.9	14.3	1.0	24.2	12.9	10.7	3.9	24.9	13.5	26.8	21.4	9.6
	1	2	1	2	9.3	26.9	5.5	8.2	3.3	14.7	28.7	25.7	23.6	7.7	10.5	8.2	25.6	20.6
	1	2	1	3	59.0	18.7	6.2	9.0	0.0	10.7	11.5	9.8	5.8	12.2	5.6	13.1	15.5	8.4
6	1	2	2	1	54.1	5.4	10.1	2.8	3.8	8.0	4.3	2.3	1.8	1.2	9.1	2.1	3.5	8.6
	1	2	2	2	7.5	2.7	3.2	0.8	2.7	38.4	4.8	8.4	7.6	53.0	43.6	53.3	2.8	18.6
	1	2	2	3	38.4	8.1	14.9	4.1	4.9	3.8	5.1	5.0	4.0	8.6	4.3	7.4	5.4	15.7
7	1	2	3	1	55.4	1.0	4.9	0.8	1.5	23.1	14.3	0.7	7.8	14.6	20.8	13.7	12.1	3.8
	1	2	3	2	4.2	21.7	13.5	6.3	8.4	9.8	33.6	18.8	30.8	21.6	12.8	21.0	30.7	24.9
	1	2	3	3	40.4	0.8	8.2	1.7	1.1	30.7	23.2	1.0	14.0	17.8	27.1	16.6	19.8	7.8
8	1	2	4	1	30.2	0.0	4.0	11.2	5.7	43.1	6.9	12.2	16.7	52.7	42.5	52.1	6.6	5.7
	1	2	4	2	1.7	51.9	22.9	24.6	15.4	28.4	16.4	11.8	11.5	46.3	49.9	46.2	41.7	34.1
	1	2	4	3	68.1	1.3	2.3	5.6	2.1	19.8	2.7	5.7	7.1	24.5	20.1	24.2	1.9	3.4
9	2	1	1	1	39.2	12.6	1.7	0.7	0.5	9.6	11.7	12.0	11.0	9.6	9.8	10.6	11.1	11.4
	2	1	1	2	16.3	11.2	3.3	2.2	0.4	18.2	6.5	12.0	1.0	19.0	16.1	21.6	4.8	10.7
	2	1	1	3	44.5	15.2	0.3	1.4	0.3	15.1	7.9	6.1	9.4	15.4	14.5	17.2	8.1	6.1
10	2	1	2	1	71.3	1.9	2.1	2.5	0.8	0.6	6.3	4.8	5.3	1.3	0.1	1.1	7.1	5.9
	2	1	2	2	14.8	5.1	3.3	5.3	1.5	13.2	7.5	1.1	2.2	20.0	11.2	17.6	9.8	4.4
	2	1	2	3	14.0	15.0	7.4	7.4	5.7	10.7	24.1	25.5	24.6	14.6	12.4	12.8	26.0	25.6
11	2	1	3	1	75.7	1.4	2.1	2.4	1.0	14.1	4.4	6.1	3.6	15.0	13.8	15.2	4.3	5.7
	2	1	3	2	12.6	21.4	5.5	5.7	0.6	11.6	9.3	1.2	15.4	18.3	9.6	19.9	10.5	2.3
	2	1	3	3	11.7	32.1	7.9	9.5	6.1	79.1	38.4	37.9	39.6	77.7	79.2	77.0	39.0	34.8
12	2	1	4	1	58.5	2.3	0.3	1.3	1.7	27.3	6.3	6.0	5.7	26.8	27.8	26.5	7.2	5.5
	2	1	4	2	9.0	23.7	7.5	5.3	3.2	0.8	4.8	7.9	1.0	6.7	1.2	4.2	5.6	1.4
	2	1	4	3	32.5	10.8	2.7	0.9	4.0	48.9	10.0	13.0	10.6	50.1	49.7	49.0	11.5	9.5
13	2	2	1	1	15.7	18.2	4.5	1.6	8.9	53.9	7.8	7.0	10.6	59.2	57.6	60.1	11.7	12.0
	2	2	1	2	7.9	22.2	3.9	4.5	5.9	32.3	11.1	3.9	3.2	21.5	25.8	28.0	4.9	6.5
	2	2	1	3	76.3	6.1	1.3	0.8	1.2	14.5	2.8	1.8	1.8	14.4	14.6	15.3	2.9	3.2
14	2	2	2	1	33.2	3.1	4.1	5.2	0.4	16.0	4.9	4.1	7.1	14.4	12.6	14.8	1.3	3.1
	2	2	2	2	6.9	16.8	13.6	10.9	2.3	48.0	9.2	2.0	21.8	36.3	42.5	32.6	1.1	15.4
	2	2	2	3	59.9	3.6	0.7	1.6	0.0	3.4	1.6	2.5	1.4	3.8	2.1	4.5	0.6	3.5
15	2	2	3	1	37.0	1.1	5.0	4.8	2.1	14.0	3.9	2.0	5.6	15.5	14.3	15.2	3.5	2.1
	2	2	3	2	3.3	76.0	22.0	21.5	9.2	115.6	42.1	25.0	22.4	85.3	111.4	95.9	38.8	28.7
	2	2	3	3	59.7	4.9	1.9	1.8	0.8	2.3	4.7	2.6	4.7	4.9	2.7	4.1	4.3	2.9
16	2	2	4	1	18.1	4.2	9.3	4.0	7.1	58.1	5.2	1.5	5.1	57.3	58.2	57.1	6.6	3.3
	2	2	4	2	0.8	195.8	47.8	55.3	31.8	12.1	24.2	44.3	5.5	7.3	1.9	14.0	35.1	56.7
	2	2	4	3	81.1	1.1	1.6	0.3	1.3	13.1	0.9	0.8	1.1	12.7	13.0	12.6	1.1	0.2

Table 7.1

Venezuela

1990 Census Proportions and Absolute Relative Residuals (ARB)
for Thirteen Different Models

Categories					Census Prop.	MODELS													
										SPREE				Unsaturated SPREE					
SG	STT	Sex	Age	LF		(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
17	3	1	1	1	47.0	1.5	12.2	9.7	4.8	27.2	8.8	1.8	6.5	30.2	32.8	29.8	14.0	9.7
	3	1	1	2	11.3	8.7	6.5	3.7	8.3	8.7	4.9	11.2	1.9	22.9	17.0	22.0	3.1	8.9
	3	1	1	3	41.8	0.6	15.5	9.9	7.6	32.9	8.6	5.1	7.8	40.2	41.5	39.5	16.5	13.3
18	3	1	2	1	74.8	2.5	1.8	0.8	4.7	1.4	4.6	0.7	5.8	3.7	2.0	3.5	4.6	3.1
	3	1	2	2	9.2	2.5	4.7	5.2	2.2	35.1	8.7	19.8	0.4	40.4	28.7	41.3	2.9	13.8
	3	1	2	3	16.0	10.1	11.1	0.9	21.0	13.8	26.7	14.9	27.3	5.8	7.4	7.3	23.1	22.5
19	3	1	3	1	79.9	4.0	0.4	0.0	2.8	10.6	0.0	0.6	1.3	11.6	11.3	11.5	1.8	2.3
	3	1	3	2	6.4	2.4	24.4	1.8	15.7	26.0	52.7	17.7	40.7	17.1	21.0	17.7	46.0	42.4
	3	1	3	3	13.6	22.2	13.8	0.6	24.0	74.4	25.0	4.8	26.6	75.8	76.5	76.0	32.1	33.4
20	3	1	4	1	64.4	10.8	8.4	3.5	3.1	17.1	3.5	3.5	2.3	13.9	16.3	14.1	3.7	6.6
	3	1	4	2	4.3	3.2	26.1	3.6	21.0	47.6	38.7	18.2	28.1	56.9	45.4	57.9	35.5	45.7
	3	1	4	3	31.2	21.9	13.8	6.8	3.4	41.8	1.8	4.8	0.8	36.6	40.0	37.3	2.8	7.3
21	3	2	1	1	12.7	0.4	28.4	36.6	9.5	20.8	20.3	1.8	24.8	75.5	66.5	74.8	16.3	20.9
	3	2	1	2	6.5	39.6	18.5	6.6	11.2	25.6	41.7	31.6	27.4	5.7	0.8	7.7	18.0	24.6
	3	2	1	3	80.8	3.3	3.0	5.2	0.6	1.2	6.5	2.8	6.1	11.4	10.5	11.1	1.1	1.3
22	3	2	2	1	27.6	24.0	15.1	2.3	1.3	28.8	16.7	2.3	11.6	22.8	24.2	21.9	13.4	11.7
	3	2	2	2	5.2	0.4	2.5	9.5	9.1	56.8	25.0	23.2	6.1	46.1	50.7	47.2	14.8	8.2
	3	2	2	3	67.3	9.9	6.0	0.2	1.2	7.5	4.9	0.8	4.3	5.8	6.0	5.3	4.4	4.2
23	3	2	3	1	31.1	19.1	10.9	0.1	2.1	2.1	17.3	2.3	12.5	11.1	11.2	11.8	8.5	7.8
	3	2	3	2	2.0	84.7	27.1	5.9	43.0	119.7	42.4	7.3	79.8	91.8	99.4	87.8	31.1	24.5
	3	2	3	3	66.9	11.4	5.9	0.1	0.3	2.6	9.3	1.3	8.2	2.4	2.2	2.8	4.9	4.4
24	3	2	4	1	15.2	10.5	4.0	16.2	12.0	52.1	8.1	10.6	1.6	58.0	56.0	58.7	1.0	3.3
	3	2	4	2	0.8	86.3	7.4	20.4	0.1	19.4	12.9	6.0	39.9	36.4	38.4	38.0	14.0	12.0
	3	2	4	3	84.0	2.7	0.7	2.7	2.2	9.6	1.6	1.8	0.7	10.8	10.5	11.0	0.0	0.7
25	4	1	1	1	48.3	4.0	4.5	1.6	0.2	10.1	7.6	4.1	8.4	17.0	11.9	15.1	5.8	2.4
	4	1	1	2	14.1	26.0	11.1	3.7	9.5	19.1	10.8	10.0	10.8	12.3	17.2	12.9	8.4	3.5
	4	1	1	3	37.6	14.8	1.7	0.6	3.3	20.1	5.7	1.5	6.7	26.5	21.7	24.3	4.3	1.8
26	4	1	2	1	79.8	6.7	3.6	3.6	2.1	6.5	1.0	0.6	0.9	7.0	8.1	6.1	2.2	0.7
	4	1	2	2	12.3	4.5	5.4	4.7	1.3	26.8	2.4	4.5	1.9	26.1	31.6	23.4	6.3	0.5
	4	1	2	3	7.9	60.8	28.3	28.6	19.3	24.2	6.0	0.7	5.6	30.0	32.6	25.1	12.1	7.4
27	4	1	3	1	82.2	4.1	1.6	2.8	0.3	9.3	1.8	0.9	2.1	8.4	8.5	8.7	0.7	1.3
	4	1	3	2	11.5	21.9	7.1	3.7	11.0	28.7	10.5	1.4	11.1	24.3	25.4	26.3	7.6	8.3
	4	1	3	3	6.3	93.7	33.8	30.2	23.7	68.7	4.9	8.7	6.5	65.2	65.0	66.3	4.4	1.5
28	4	1	4	1	64.0	0.7	0.8	1.2	3.9	17.8	0.8	0.0	0.6	21.4	18.2	22.4	0.2	4.8
	4	1	4	2	8.7	24.7	9.2	0.2	11.7	5.9	2.1	1.8	1.7	4.4	5.9	6.5	2.2	8.7
	4	1	4	3	27.2	9.7	1.1	2.8	5.3	43.7	1.3	0.6	0.7	48.9	44.6	50.6	1.1	8.4
29	4	2	1	1	20.8	21.4	0.9	3.4	7.0	42.5	0.9	4.8	0.2	39.0	47.8	35.8	6.7	5.2
	4	2	1	2	8.9	23.8	0.2	5.5	4.9	5.8	10.4	11.7	9.6	4.0	1.1	7.5	15.9	10.3
	4	2	1	3	70.3	9.3	0.2	0.3	1.4	13.3	1.0	2.9	1.1	12.0	14.0	11.5	0.0	2.8
30	4	2	2	1	34.9	16.4	9.2	7.7	4.2	14.1	2.0	6.3	1.6	22.5	18.8	20.0	6.8	7.4
	4	2	2	2	8.5	3.5	5.5	6.4	0.2	48.7	11.6	13.8	11.1	52.0	50.6	50.0	15.0	14.2
	4	2	2	3	56.6	9.6	4.8	3.8	2.6	1.4	0.5	1.8	0.7	6.1	4.0	4.8	1.9	2.4
31	4	2	3	1	39.3	12.3	6.5	5.8	1.8	11.1	5.3	3.8	4.8	9.2	10.6	11.3	6.3	4.5
	4	2	3	2	5.0	17.1	18.2	8.2	13.0	18.0	23.9	15.9	23.2	14.7	11.8	19.2	27.7	22.9
	4	2	3	3	55.7	10.2	3.0	3.3	0.1	6.2	1.6	1.3	1.3	5.2	6.4	6.2	1.9	1.1
32	4	2	4	1	19.7	10.9	5.6	3.7	1.6	54.6	0.3	6.3	2.3	53.0	58.3	54.6	8.0	0.8
	4	2	4	2	1.5	80.3	9.6	1.3	6.0	34.0	9.7	9.0	10.1	40.1	38.5	38.3	16.5	17.3
	4	2	4	3	78.9	4.2	1.2	0.9	0.5	14.3	0.3	1.4	0.8	14.0	15.3	14.3	2.3	0.5

Table 7.1

Venezuela

1990 Census Proportions and Absolute Relative Residuals (ARB)
for Thirteen Different Models

Categories					Census Prop.	MODELS												
										SPREE				Unsaturated SPREE				
SG	STT	Sex	Age	LF		(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
33	5	1	1	1	47.6	3.2	13.1	7.8	4.5	16.5	2.9	4.4	8.9	25.5	20.9	23.7	1.4	4.0
	5	1	1	2	12.1	21.3	4.6	3.0	11.0	10.9	1.6	8.0	8.4	13.4	10.2	15.9	2.5	4.0
	5	1	1	3	40.2	2.7	14.1	10.2	2.0	22.7	3.9	2.8	13.0	34.2	27.9	32.8	0.9	6.0
34	5	1	2	1	75.9	4.8	0.8	2.8	2.0	1.8	4.3	3.6	5.6	4.4	5.1	3.1	1.7	3.2
	5	1	2	2	12.1	7.2	5.7	0.5	6.0	7.1	12.9	3.2	10.9	19.7	11.7	15.0	9.8	7.7
	5	1	2	3	12.0	37.4	11.1	18.2	6.7	3.9	14.0	19.7	24.5	7.6	20.3	4.6	1.1	12.4
35	5	1	3	1	78.5	2.7	1.2	1.2	3.6	13.7	5.6	2.0	6.5	13.1	13.3	13.7	5.2	5.2
	5	1	3	2	10.5	24.7	9.2	2.5	10.0	18.7	2.1	2.1	3.7	20.2	16.4	23.9	3.0	7.0
	5	1	3	3	11.0	42.9	0.5	10.8	16.1	80.1	41.8	16.3	49.9	74.5	79.2	75.2	40.0	30.2
36	5	1	4	1	64.3	9.6	7.6	6.2	1.1	13.4	7.0	1.1	2.6	16.6	12.3	17.8	8.0	1.4
	5	1	4	2	7.2	23.8	7.7	2.4	4.4	4.8	1.1	10.5	4.5	8.3	3.9	3.6	3.7	2.0
	5	1	4	3	28.5	27.8	19.2	14.5	3.6	31.6	16.1	0.3	4.9	39.5	28.7	41.1	19.1	3.6
37	5	2	1	1	14.6	10.5	14.5	24.8	3.2	57.9	10.9	12.0	0.2	68.8	79.7	62.9	30.2	13.8
	5	2	1	2	5.9	16.6	11.4	6.0	2.8	42.8	20.4	8.8	7.1	28.9	35.2	40.4	14.7	16.0
	5	2	1	3	79.5	3.2	3.5	5.0	0.4	13.8	3.5	2.9	0.5	14.8	17.3	14.6	6.6	3.7
38	5	2	2	1	28.1	24.5	15.5	7.5	1.1	9.5	0.1	3.2	8.7	32.6	21.9	28.6	12.0	16.9
	5	2	2	2	6.2	3.0	0.3	1.0	5.0	55.8	22.0	8.0	30.1	44.9	52.2	39.4	16.5	4.4
	5	2	2	3	65.7	10.8	6.6	3.3	0.0	1.2	2.1	0.6	6.6	9.7	4.4	8.5	3.6	7.6
39	5	2	3	1	31.4	20.9	12.8	3.1	1.3	13.8	36.8	5.2	27.1	2.5	8.9	5.7	32.2	14.5
	5	2	3	2	2.9	58.4	8.9	9.3	3.8	112.2	39.5	11.9	27.7	64.5	93.1	80.2	26.8	18.8
	5	2	3	3	65.7	12.6	6.5	1.9	0.5	11.5	19.4	3.0	14.2	1.7	8.3	0.8	16.6	7.8
40	5	2	4	1	13.0	34.9	27.0	34.6	5.4	51.3	13.9	42.0	0.0	43.0	53.9	45.5	8.1	25.4
	5	2	4	2	0.8	127.2	12.7	22.6	2.3	29.9	2.7	28.5	11.1	20.2	34.5	13.2	6.0	21.2
	5	2	4	3	86.2	6.5	4.2	5.4	0.8	8.0	2.1	6.6	0.1	6.7	8.4	7.0	1.2	4.0
41	6	1	1	1	43.1	12.5	2.1	0.6	1.5	10.2	10.2	9.2	13.0	8.4	8.7	11.4	11.0	9.7
	6	1	1	2	14.4	10.7	3.6	3.1	0.2	24.6	14.7	22.2	5.8	22.8	18.4	29.1	8.2	19.8
	6	1	1	3	42.5	16.3	0.9	0.4	1.6	18.7	5.3	1.8	11.2	16.2	15.1	21.4	8.4	3.1
42	6	1	2	1	75.9	3.0	0.6	0.3	0.0	1.0	4.7	3.7	4.7	1.5	0.4	1.5	5.4	4.3
	6	1	2	2	12.3	3.6	2.5	4.6	1.6	0.9	20.1	11.0	13.8	16.1	1.3	10.2	18.5	11.5
	6	1	2	3	11.8	22.9	1.2	2.7	1.4	7.6	9.2	12.1	15.9	7.3	0.9	0.7	15.3	16.0
43	6	1	3	1	79.4	1.8	1.4	1.6	0.9	13.2	5.9	5.8	5.6	13.0	12.4	13.2	4.9	5.1
	6	1	3	2	9.9	16.2	0.3	2.6	4.6	15.4	6.6	0.9	14.3	18.8	9.4	22.7	13.3	4.8
	6	1	3	3	10.8	28.3	10.5	9.5	10.7	83.3	49.7	42.2	54.2	78.5	82.8	76.6	48.3	32.9
44	6	1	4	1	63.7	4.2	2.4	0.7	2.6	20.2	0.7	4.6	2.5	21.9	21.8	21.0	3.2	1.9
	6	1	4	2	8.2	27.7	12.7	3.0	8.9	13.8	17.2	1.8	9.2	6.4	12.2	11.8	14.9	15.8
	6	1	4	3	28.1	17.4	9.0	0.6	8.5	41.7	3.3	11.0	2.9	47.8	45.7	44.2	3.0	0.2
45	6	2	1	1	18.6	21.0	0.3	1.4	1.9	55.9	11.1	8.3	9.0	51.3	49.6	54.2	6.5	9.2
	6	2	1	2	7.1	20.2	5.7	6.0	3.3	40.9	20.0	7.9	0.8	15.8	22.5	31.9	3.2	11.0
	6	2	1	3	74.3	7.2	0.6	0.2	0.2	17.9	4.7	2.8	2.2	14.4	14.6	16.7	1.9	3.4
46	6	2	2	1	34.3	10.6	3.1	0.7	4.8	18.1	7.1	7.3	5.4	21.6	15.8	23.7	4.8	11.9
	6	2	2	2	6.1	17.6	14.3	9.4	4.4	46.1	5.9	8.4	22.5	40.0	50.5	31.5	14.0	18.1
	6	2	2	3	59.7	7.9	3.2	0.5	3.2	5.7	3.5	5.1	0.8	8.3	3.9	10.4	1.3	8.7
47	6	2	3	1	39.5	4.1	2.1	3.5	1.1	6.2	11.3	0.8	9.4	12.2	2.5	10.5	15.1	6.2
	6	2	3	2	3.1	64.6	13.9	8.8	3.5	111.1	36.2	15.6	12.6	65.1	106.6	88.7	33.2	22.2
	6	2	3	3	57.4	6.3	0.7	1.9	0.6	1.8	9.8	1.4	7.2	4.9	4.0	2.5	12.2	5.5
48	6	2	4	1	21.1	7.2	12.2	2.1	12.3	58.7	8.3	10.5	12.6	58.2	60.6	56.9	13.2	4.7
	6	2	4	2	1.0	111.9	5.4	25.1	4.9	37.4	13.4	18.0	30.6	26.8	34.0	14.8	10.0	14.7
	6	2	4	3	77.8	0.4	3.3	0.2	3.4	16.4	2.4	3.1	3.8	16.2	16.9	15.7	3.7	1.1

Table 7.1

Venezuela

1990 Census Proportions and Absolute Relative Residuals (ARB)
for Thirteen Different Models

Categories					Census Prop.	MODELS												
										SPREE				Unsaturated SPREE				
SG	STT	Sex	Age	LF		(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
49	7	1	1	1	45.1	8.0	1.4	1.9	1.2	10.1	8.9	10.4	8.4	12.2	10.3	11.0	8.4	7.9
	7	1	1	2	14.9	17.3	2.9	1.1	3.4	15.8	7.0	9.0	4.5	15.9	16.5	15.4	7.2	5.9
	7	1	1	3	39.9	15.5	0.5	1.7	0.2	17.3	7.5	8.4	7.8	19.7	17.8	18.2	6.9	6.7
50	7	1	2	1	76.2	3.6	0.2	0.4	0.2	1.7	4.3	4.0	3.8	3.1	2.6	2.7	3.5	3.2
	7	1	2	2	13.0	1.4	0.3	0.8	0.2	17.1	6.4	7.1	3.8	21.1	19.9	20.7	3.6	2.6
	7	1	2	3	10.8	26.9	1.6	3.5	1.5	8.7	22.7	19.5	22.0	3.4	5.5	6.1	20.4	19.6
51	7	1	3	1	79.9	2.5	0.3	0.4	0.3	11.3	3.3	3.2	2.9	11.2	11.0	11.3	2.8	3.5
	7	1	3	2	10.7	16.0	0.2	3.2	0.7	17.8	1.6	4.6	4.3	16.9	16.3	17.5	3.3	2.3
	7	1	3	3	9.3	39.8	3.1	0.2	3.4	76.6	29.9	32.5	30.1	76.3	75.7	76.8	27.7	32.6
52	7	1	4	1	62.2	0.5	1.2	0.1	1.4	22.9	4.5	3.9	4.0	24.5	22.8	25.0	4.6	6.8
	7	1	4	2	8.8	25.7	10.4	0.5	9.8	1.8	1.9	2.9	0.7	2.6	2.8	3.0	0.5	5.6
	7	1	4	3	29.1	8.7	0.5	0.3	0.0	49.5	9.1	9.2	8.7	51.5	49.7	52.7	9.8	12.9
53	7	2	1	1	21.2	27.7	8.5	3.8	8.6	37.0	2.6	0.7	1.2	34.7	36.8	33.3	1.7	6.5
	7	2	1	2	8.2	21.2	4.1	1.8	3.1	13.5	4.6	5.9	9.3	14.6	14.0	14.8	3.7	4.0
	7	2	1	3	70.6	10.8	2.1	0.9	2.2	12.7	1.3	0.9	1.4	12.1	12.7	11.7	0.9	2.4
54	7	2	2	1	35.2	10.5	3.4	0.8	3.7	17.6	5.8	3.0	7.2	20.4	20.9	19.1	9.0	7.0
	7	2	2	2	7.6	5.4	2.9	1.2	1.9	41.6	0.1	1.1	5.6	45.1	45.5	44.8	6.2	5.1
	7	2	2	3	57.3	7.2	2.5	0.6	2.5	5.3	3.6	1.7	3.7	6.6	6.8	5.8	4.7	3.6
55	7	2	3	1	38.2	10.5	4.5	0.9	4.5	10.2	7.1	6.3	8.2	7.7	10.5	8.8	6.8	7.8
	7	2	3	2	4.3	32.1	8.0	7.9	9.1	37.2	11.9	9.3	16.8	35.1	33.0	35.6	13.8	12.2
	7	2	3	3	57.5	9.4	2.4	0.0	2.3	4.0	3.8	3.5	4.2	2.5	4.5	3.2	3.5	4.3
56	7	2	4	1	19.7	4.2	1.0	0.2	2.2	56.3	3.1	0.7	3.3	53.6	56.7	54.6	4.7	0.2
	7	2	4	2	1.5	72.4	13.7	2.7	15.4	33.8	9.5	5.1	15.4	39.2	38.3	39.4	16.5	18.4
	7	2	4	3	78.9	2.4	0.5	0.1	0.8	14.7	0.9	0.1	1.1	14.1	14.8	14.3	1.5	0.4
57	8	1	1	1	43.7	3.3	6.8	6.4	1.3	18.5	1.2	4.0	3.2	19.6	19.6	17.7	0.0	2.8
	8	1	1	2	16.4	4.6	9.5	6.0	4.9	11.9	0.1	9.9	1.5	26.9	15.7	26.1	4.0	15.9
	8	1	1	3	39.9	5.5	11.4	9.5	3.4	25.2	1.3	0.4	4.2	32.5	27.9	30.2	1.6	3.4
58	8	1	2	1	75.2	5.7	1.9	2.2	0.2	3.0	3.5	1.0	4.0	5.5	3.4	4.7	3.3	2.0
	8	1	2	2	13.1	0.8	2.2	1.8	3.3	6.5	14.2	4.7	14.0	25.6	4.2	25.0	16.1	1.1
	8	1	2	3	11.7	35.4	9.4	12.1	2.1	12.0	6.8	11.5	9.8	6.5	17.1	2.3	3.1	13.8
59	8	1	3	1	77.2	2.2	1.3	0.5	2.7	13.1	4.0	4.9	4.5	13.7	13.0	13.9	3.9	5.0
	8	1	3	2	11.8	21.5	5.9	4.5	5.1	13.1	6.6	5.4	6.5	19.5	13.8	20.3	5.0	2.6
	8	1	3	3	11.0	38.8	2.7	1.7	13.4	77.7	35.3	28.9	38.3	75.1	76.6	76.0	32.5	32.5
60	8	1	4	1	62.2	6.9	5.1	2.6	0.9	22.9	4.0	4.4	5.2	19.4	20.4	20.5	0.7	1.1
	8	1	4	2	7.0	7.4	11.8	4.0	15.9	6.2	2.0	11.6	2.9	26.8	8.0	26.1	2.2	19.8
	8	1	4	3	30.8	15.6	7.5	6.2	1.8	47.7	8.6	11.5	11.2	45.3	43.0	47.2	2.0	6.8
61	8	2	1	1	14.4	8.7	16.3	18.6	5.7	60.8	13.6	16.3	10.3	70.6	70.5	66.6	22.6	16.5
	8	2	1	2	7.3	19.1	7.5	13.1	0.6	29.9	8.6	2.1	6.1	20.2	39.6	20.2	17.5	0.4
	8	2	1	3	78.3	3.4	3.7	4.6	1.0	14.0	3.3	2.8	2.5	14.9	16.7	14.1	5.8	3.0
62	8	2	2	1	29.7	17.3	8.9	5.5	1.3	19.4	9.6	9.7	7.1	25.4	19.3	22.5	9.1	11.1
	8	2	2	2	7.4	2.1	1.2	3.2	6.9	62.9	34.9	16.8	36.4	46.9	58.8	46.6	28.5	7.9
	8	2	2	3	62.8	8.4	4.1	2.2	0.2	1.7	0.4	2.6	0.9	6.5	2.2	5.1	0.9	4.3
63	8	2	3	1	34.5	10.3	3.3	1.1	4.2	8.1	12.0	3.9	9.6	11.1	9.8	13.3	10.1	5.2
	8	2	3	2	3.1	75.7	21.1	18.4	13.9	102.4	33.0	9.4	30.7	76.6	93.9	77.1	28.0	17.0
	8	2	3	3	62.4	9.5	2.9	0.3	1.6	0.6	8.3	2.7	6.9	2.3	0.7	3.5	7.0	3.7
64	8	2	4	1	15.6	13.3	6.9	11.9	4.5	61.4	9.3	2.3	13.5	52.2	54.6	54.0	4.0	5.6
	8	2	4	2	0.7	244.0	70.7	47.5	55.3	5.8	36.8	53.7	31.3	17.0	4.5	16.2	33.9	62.2
	8	2	4	3	83.7	4.4	1.8	2.6	0.4	11.5	1.4	0.0	2.3	9.6	10.2	9.9	1.0	1.5

Table 7.1

Venezuela

1990 Census Proportions and Absolute Relative Residuals (ARB)
for Thirteen Different Models

Categories					Census Prop.	MODELS													
SG	STT	Sex	Age	LF		(a)	(b)	(c)	(d)	SPREE				Unsaturated SPREE					
										(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
65	9	1	1	1	36.1	10.3	1.1	0.3	3.3	14.3	7.2	11.7	5.5	15.6	15.2	15.4	6.6	7.2
	9	1	1	2	18.1	10.6	4.3	4.5	0.5	3.1	10.9	0.2	8.9	21.1	10.5	17.2	3.1	4.5
	9	1	1	3	45.8	12.3	2.5	2.0	2.4	12.5	10.0	9.3	7.9	20.7	16.1	18.9	6.4	3.9
66	9	1	2	1	66.1	0.2	4.3	4.2	2.6	0.5	8.0	6.0	7.3	1.1	0.7	1.7	9.3	7.0
	9	1	2	2	18.3	11.5	9.3	6.9	3.9	6.6	12.4	2.9	13.1	13.3	0.0	17.8	18.2	3.1
	9	1	2	3	15.6	14.4	7.4	9.6	6.5	5.5	19.4	28.9	15.6	11.2	3.0	13.9	18.2	26.0
67	9	1	3	1	73.1	2.7	1.2	2.3	0.2	13.8	2.4	4.1	2.2	15.4	15.0	14.8	3.8	3.8
	9	1	3	2	14.2	18.6	1.7	6.6	4.5	2.4	17.4	5.2	16.6	11.7	9.2	8.0	8.7	10.1
	9	1	3	3	12.7	35.9	4.7	5.9	3.7	76.7	33.3	29.3	30.9	75.2	76.0	75.7	31.7	32.9
68	9	1	4	1	53.3	0.8	1.3	0.3	0.1	32.7	9.0	4.6	7.3	30.5	31.8	30.4	7.7	6.7
	9	1	4	2	9.9	21.0	4.0	4.9	1.8	5.3	1.7	5.5	3.4	15.2	1.2	20.2	6.4	11.3
	9	1	4	3	36.8	6.7	0.9	1.8	0.3	48.8	12.5	8.2	9.6	48.3	46.3	49.4	9.5	12.7
69	9	2	1	1	14.3	22.6	0.7	0.9	4.3	35.9	6.1	2.6	3.1	46.8	47.7	49.1	4.9	3.0
	9	2	1	2	8.0	18.5	9.5	10.1	4.5	18.0	3.3	12.8	1.3	27.5	45.9	13.6	21.3	7.0
	9	2	1	3	77.6	6.1	0.8	1.2	0.3	8.5	1.5	1.8	0.4	11.5	13.5	10.5	3.1	0.2
70	9	2	2	1	30.9	1.1	8.4	8.8	4.0	7.1	2.6	2.7	0.0	6.5	1.0	7.0	8.8	3.7
	9	2	2	2	7.6	15.9	12.4	16.5	2.7	58.9	28.5	7.8	25.9	37.0	48.0	44.2	10.7	4.4
	9	2	2	3	61.6	1.4	2.7	2.4	2.4	3.7	4.8	0.4	3.2	1.3	5.4	1.9	5.7	2.4
71	9	2	3	1	33.2	0.8	5.6	4.9	2.1	16.2	2.9	2.2	5.5	19.9	19.9	19.3	0.9	1.3
	9	2	3	2	3.3	95.7	35.4	27.9	16.3	98.8	31.8	11.2	37.0	101.2	108.5	78.6	39.5	18.9
	9	2	3	3	63.5	5.3	1.1	1.2	0.2	3.4	3.2	0.6	4.8	5.2	4.9	6.1	1.6	0.3
72	9	2	4	1	13.8	9.4	3.4	1.1	6.5	57.7	0.8	3.9	1.2	54.5	56.6	54.7	0.5	5.0
	9	2	4	2	0.8	213.3	56.5	52.0	33.9	22.7	11.9	15.9	16.0	8.9	8.0	4.0	30.5	35.4
	9	2	4	3	85.4	3.6	1.1	0.7	1.4	9.5	0.0	0.5	0.4	8.7	9.2	8.9	0.4	1.2
73	10	1	1	1	46.1	3.5	13.1	10.0	5.4	11.8	8.3	10.8	12.0	22.4	15.4	15.7	4.8	3.4
	10	1	1	2	15.3	10.2	4.4	2.0	1.1	18.6	7.0	17.9	4.6	22.9	16.5	21.7	5.5	11.9
	10	1	1	3	38.6	0.1	17.4	12.7	5.9	21.4	7.2	5.8	12.6	35.8	24.8	27.3	3.5	0.6
74	10	1	2	1	74.4	5.3	1.3	1.8	1.3	2.5	4.2	0.8	5.1	3.7	2.3	0.9	4.7	5.3
	10	1	2	2	12.3	6.0	8.0	0.1	8.0	18.7	2.4	12.3	2.1	30.0	13.4	25.0	8.6	0.9
	10	1	2	3	13.3	24.1	0.0	10.4	14.6	3.5	21.2	15.7	26.7	7.3	0.5	18.3	18.5	30.7
75	10	1	3	1	75.8	1.0	2.8	0.7	5.1	14.7	5.1	3.4	5.9	16.3	14.9	17.4	6.1	8.9
	10	1	3	2	11.8	22.4	6.5	3.5	6.9	10.9	11.0	0.7	10.8	22.1	10.8	26.5	9.0	9.1
	10	1	3	3	12.4	27.5	10.9	0.7	24.3	79.3	41.5	21.8	46.4	78.4	80.8	81.1	45.6	45.9
76	10	1	4	1	62.4	8.5	6.5	4.2	0.5	21.1	2.0	10.6	4.5	20.0	11.8	23.5	9.8	4.9
	10	1	4	2	8.7	26.7	10.9	3.3	7.8	16.1	18.3	11.2	16.9	0.5	14.1	4.4	21.8	7.5
	10	1	4	3	29.0	26.3	17.3	10.1	3.3	40.7	1.2	26.2	4.7	42.8	21.2	49.3	27.7	8.2
77	10	2	1	1	12.7	0.0	28.0	36.5	10.7	73.2	23.5	25.7	15.5	97.0	117.9	77.8	61.1	22.3
	10	2	1	2	7.7	26.4	1.4	4.0	8.9	29.8	11.7	1.6	7.4	11.6	23.5	17.3	6.7	3.8
	10	2	1	3	79.6	2.6	4.4	6.2	0.8	14.6	4.9	4.0	3.2	16.7	21.1	14.1	10.4	3.2
78	10	2	2	1	28.1	21.4	12.9	4.6	0.5	30.1	21.3	12.7	15.1	35.0	28.7	23.6	19.4	10.7
	10	2	2	2	6.2	18.2	14.9	0.2	9.1	63.2	33.3	24.7	35.3	38.1	49.0	33.7	10.1	13.7
	10	2	2	3	65.7	10.9	6.9	2.0	1.1	6.8	5.9	3.1	3.1	11.3	7.6	6.9	7.3	5.9
79	10	2	3	1	32.2	15.3	8.1	1.6	3.8	4.9	28.5	4.9	22.6	3.2	0.8	12.1	24.4	6.3
	10	2	3	2	3.1	71.6	18.7	13.7	12.7	130.2	55.0	17.5	51.7	74.1	111.1	84.6	40.6	22.2
	10	2	3	3	64.7	11.0	4.9	0.1	1.3	8.7	16.9	3.3	13.7	2.0	5.7	2.0	14.1	4.2
80	10	2	4	1	14.9	14.7	8.5	18.0	8.0	77.5	43.9	16.1	48.9	48.8	65.2	55.1	16.4	2.5
	10	2	4	2	1.0	130.1	14.8	30.5	4.3	72.3	56.6	22.3	59.4	19.8	48.7	17.6	24.6	15.2
	10	2	4	3	84.1	4.1	1.7	3.5	1.4	14.5	8.4	3.1	9.3	8.9	12.1	10.0	3.2	0.6

Table 7.1

Venezuela

1990 Census Proportions and Absolute Relative Residuals (ARB)
for Thirteen Different Models

Categories					Census Prop.	MODELS												
SG	STT	Sex	Age	LF		(a)	(b)	(c)	(d)	SPREE				Unsaturated SPREE				
										(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
81	11	1	1	1	46.8	1.0	8.3	5.7	3.6	3.2	15.5	8.5	15.4	20.3	8.0	15.3	11.3	2.9
	11	1	1	2	14.6	15.2	0.7	0.4	2.7	20.9	12.3	5.5	13.5	16.5	21.1	12.4	12.2	2.3
	11	1	1	3	38.6	7.0	9.8	7.0	3.3	11.8	14.1	8.1	13.6	30.9	17.7	23.3	9.0	2.6
82	11	1	2	1	76.3	4.9	1.2	1.9	0.3	7.5	1.0	1.0	1.1	5.0	8.8	3.5	1.8	2.5
	11	1	2	2	13.1	3.5	2.0	1.2	3.2	33.4	9.7	3.8	8.3	22.8	32.4	23.0	8.0	0.5
	11	1	2	3	10.6	39.1	11.5	11.8	1.8	12.8	4.6	11.8	2.7	7.9	23.2	3.6	2.8	17.5
83	11	1	3	1	80.7	4.6	1.4	0.7	0.1	9.7	1.3	2.0	1.4	9.5	9.5	10.0	0.7	2.1
	11	1	3	2	10.4	14.9	2.1	1.5	0.6	11.8	10.5	9.0	8.9	11.8	13.6	12.6	5.8	8.2
	11	1	3	3	8.9	58.6	10.2	8.4	0.2	74.3	24.0	28.4	23.5	72.2	69.8	75.2	12.9	28.3
84	11	1	4	1	61.2	1.9	0.0	2.2	3.6	20.6	0.6	2.5	0.3	23.8	18.8	26.2	1.3	7.6
	11	1	4	2	7.8	19.5	2.4	4.5	1.8	20.9	16.6	7.3	14.7	11.1	16.9	11.6	10.3	8.5
	11	1	4	3	31.0	8.7	0.6	5.4	6.8	46.0	5.5	6.8	4.3	49.9	41.5	54.8	0.0	17.2
85	11	2	1	1	16.5	13.7	9.9	15.3	2.1	57.3	10.8	0.6	11.1	60.0	88.6	51.5	38.6	4.5
	11	2	1	2	8.0	25.1	0.1	0.7	1.6	10.4	6.3	3.3	2.9	13.1	9.9	8.8	5.0	10.3
	11	2	1	3	75.5	5.6	2.2	3.4	0.3	13.6	1.7	0.5	2.1	14.5	20.4	12.2	7.9	0.1
86	11	2	2	1	32.4	14.1	6.5	4.1	0.8	11.0	0.2	6.5	0.6	22.4	8.6	16.3	1.5	4.3
	11	2	2	2	7.7	2.4	4.7	2.0	4.6	52.3	16.4	15.8	13.8	47.8	47.8	49.4	9.5	13.1
	11	2	2	3	59.9	7.3	2.9	2.0	0.2	0.8	2.0	1.5	1.5	5.9	1.5	2.5	2.0	0.6
87	11	2	3	1	37.5	7.2	1.0	1.5	4.5	15.0	2.3	3.4	2.6	12.6	18.2	17.2	0.3	1.2
	11	2	3	2	3.5	55.5	8.1	4.1	8.3	46.5	3.3	3.5	0.0	61.9	49.1	56.4	1.3	2.4
	11	2	3	3	59.0	7.8	1.1	1.2	2.4	6.8	1.3	1.9	1.7	4.3	8.7	7.6	0.3	0.6
88	11	2	4	1	16.1	18.3	12.2	7.5	2.9	57.9	3.3	4.1	4.1	49.1	58.5	53.2	4.5	4.9
	11	2	4	2	0.8	198.0	49.2	41.0	45.5	27.0	5.4	2.1	7.9	5.8	3.7	0.3	36.2	37.0
	11	2	4	3	83.1	5.4	2.8	1.8	1.0	11.5	0.6	0.8	0.7	9.5	11.4	10.3	0.5	1.3
89	12	1	1	1	50.4	0.4	10.4	6.1	5.5	21.5	1.6	2.9	4.2	23.4	19.7	24.4	0.9	5.5
	12	1	1	2	9.4	20.8	4.7	1.6	2.2	28.5	18.3	6.9	21.7	15.9	28.0	14.0	18.0	1.7
	12	1	1	3	40.2	4.3	12.0	7.3	6.4	33.6	6.3	5.3	0.2	33.1	31.2	33.9	5.3	7.3
90	12	1	2	1	76.0	0.0	3.9	0.4	5.8	0.7	5.3	4.8	7.4	1.2	0.9	0.7	5.0	6.3
	12	1	2	2	8.1	2.7	4.3	15.6	1.4	46.0	16.8	8.0	8.0	30.5	48.5	34.3	19.4	7.4
	12	1	2	3	15.9	1.6	21.0	10.0	27.2	20.1	33.9	26.9	39.5	21.5	20.5	21.0	33.7	33.8
91	12	1	3	1	84.3	5.6	1.8	0.6	0.3	8.6	0.8	2.0	1.0	8.5	10.0	8.2	2.7	0.7
	12	1	3	2	6.7	14.1	3.8	1.7	2.2	27.1	11.2	9.3	18.3	11.9	28.4	9.0	10.7	12.3
	12	1	3	3	9.0	63.3	14.4	6.8	4.8	59.9	15.6	11.4	4.1	70.9	72.2	70.4	17.1	15.3
92	12	1	4	1	65.7	6.7	4.3	4.3	0.8	17.0	0.8	3.7	3.0	18.9	16.0	18.3	2.5	0.8
	12	1	4	2	4.8	15.6	2.8	3.2	1.2	52.2	46.4	6.3	37.5	15.8	51.7	19.2	45.2	13.5
	12	1	4	3	29.5	17.4	9.2	10.1	1.9	46.3	5.9	9.2	12.7	44.6	44.1	44.0	1.8	0.4
93	12	2	1	1	17.2	15.5	8.4	17.1	0.8	71.0	22.3	21.2	11.1	57.3	64.0	60.3	16.2	11.9
	12	2	1	2	5.3	27.5	2.7	2.7	4.9	4.5	19.2	7.3	8.0	8.2	7.2	1.5	22.4	15.8
	12	2	1	3	77.6	5.3	1.7	3.6	0.2	15.4	3.6	4.2	1.9	13.2	13.7	13.4	2.1	1.6
94	12	2	2	1	31.5	20.2	12.1	1.0	4.8	17.6	7.9	1.1	0.5	26.3	21.8	27.7	10.6	17.1
	12	2	2	2	6.1	20.0	22.0	20.0	14.5	51.9	15.8	25.1	3.2	58.7	53.6	61.4	19.2	32.6
	12	2	2	3	62.4	8.3	4.0	2.4	1.0	3.8	2.5	1.9	0.6	7.5	5.8	8.0	3.5	5.4
95	12	2	3	1	39.3	4.0	2.7	1.2	8.9	5.2	12.7	10.8	5.4	15.9	13.5	14.7	2.8	1.4
	12	2	3	2	2.2	56.4	8.0	4.9	18.8	39.5	10.7	10.1	4.5	59.2	32.0	49.0	14.6	3.3
	12	2	3	3	58.5	4.8	1.5	1.0	5.2	2.0	8.1	7.6	3.8	8.4	7.9	8.0	1.4	0.8
96	12	2	4	1	15.5	23.6	16.6	17.5	5.9	58.1	2.2	9.6	12.6	47.9	50.7	47.3	10.6	18.8
	12	2	4	2	0.5	173.5	36.3	27.3	45.8	19.4	16.1	13.6	30.1	6.2	21.2	12.0	67.5	20.9
	12	2	4	3	84.0	5.5	3.3	3.4	1.4	10.8	0.3	1.7	2.1	8.9	9.2	8.8	2.4	3.6

Table 7.1

Venezuela

1990 Census Proportions and Absolute Relative Residuals (ARB)
for Thirteen Different Models

Categories					Census	MODELS												
SG	STT	Sex	Age	LF	Prop.	(a)	(b)	(c)	(d)	SPREE				Unsaturated SPREE				
										(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
97	13	1	1	1	46.4	26.2	17.4	10.7	9.7	9.4	7.4	7.2	2.9	6.5	3.4	1.8	12.2	19.4
	13	1	1	2	12.9	12.8	1.1	3.1	3.9	21.6	15.6	10.3	17.0	18.1	20.8	20.2	14.2	13.6
	13	1	1	3	40.8	33.8	20.1	13.1	9.8	17.5	3.5	4.9	2.1	1.6	10.5	4.3	9.4	17.8
98	13	1	2	1	83.1	3.4	0.9	0.9	2.8	1.2	2.7	3.6	1.7	3.0	1.3	3.9	2.4	0.3
	13	1	2	2	9.5	1.9	2.3	3.2	2.5	16.6	9.0	16.8	7.3	22.1	22.3	23.0	2.7	1.6
	13	1	2	3	7.3	36.1	7.5	5.8	28.1	7.7	18.5	18.6	9.6	4.8	13.8	14.2	23.5	0.7
99	13	1	3	1	86.5	3.0	1.1	0.8	2.5	6.8	0.9	1.3	0.1	6.5	6.8	6.2	0.9	0.2
	13	1	3	2	7.9	14.7	1.1	4.7	1.6	24.9	5.9	1.5	4.2	20.8	23.1	19.6	1.9	1.9
	13	1	3	3	5.6	68.0	15.0	5.4	37.2	70.7	5.0	22.1	4.1	72.2	73.0	69.2	11.7	6.3
100	13	1	4	1	69.7	1.6	2.9	2.8	1.7	20.6	7.3	4.2	4.8	19.9	22.0	18.5	9.5	3.4
	13	1	4	2	5.8	11.2	7.0	7.4	4.2	16.6	17.0	6.1	17.5	7.3	19.0	7.8	23.0	7.9
	13	1	4	3	24.5	1.8	10.0	9.8	3.8	62.4	24.7	13.4	17.9	58.2	66.8	54.4	32.4	11.4
101	13	2	1	1	28.6	25.6	7.4	16.0	0.4	34.5	4.5	4.2	0.4	30.0	13.4	34.2	22.1	3.8
	13	2	1	2	8.3	25.1	2.5	4.4	2.3	4.4	20.1	14.4	19.7	0.4	6.7	0.3	23.6	13.6
	13	2	1	3	63.1	14.9	3.7	7.8	0.1	15.1	4.7	3.8	2.8	13.6	5.2	15.6	13.1	3.5
102	13	2	2	1	49.7	2.1	7.2	1.3	2.0	9.6	2.4	1.5	0.6	4.9	13.7	7.3	1.2	3.2
	13	2	2	2	6.6	5.1	4.1	4.2	6.9	34.1	11.4	3.9	10.0	47.5	43.8	47.4	3.1	8.5
	13	2	2	3	43.7	1.6	7.6	0.9	1.3	5.8	1.0	1.1	2.2	1.6	9.0	1.2	0.9	4.9
103	13	2	3	1	51.0	2.4	2.0	1.2	3.1	14.2	3.4	3.4	0.5	12.3	11.8	10.0	1.5	1.8
	13	2	3	2	4.0	21.9	13.9	8.3	11.7	18.6	28.4	17.0	29.0	24.7	22.0	25.4	25.0	21.8
	13	2	3	3	45.0	4.7	3.5	0.6	2.5	14.4	6.4	2.4	3.2	11.7	11.4	9.0	3.9	0.1
104	13	2	4	1	28.6	2.8	6.9	6.0	0.2	45.4	6.0	2.0	9.2	54.7	45.4	52.9	4.2	4.7
	13	2	4	2	1.6	51.1	23.6	23.6	20.0	22.7	7.8	24.0	8.4	45.7	27.3	44.5	13.6	30.7
	13	2	4	3	69.8	0.0	3.4	3.0	0.5	19.2	2.3	1.4	3.6	23.5	19.2	22.7	1.4	2.6
105	14	1	1	1	35.5	13.1	1.0	2.9	0.4	11.0	12.7	10.5	11.8	13.5	13.9	14.3	9.2	10.0
	14	1	1	2	16.3	13.0	2.8	5.7	2.1	21.0	8.3	11.4	4.2	18.7	17.3	21.3	4.2	8.4
	14	1	1	3	48.2	14.1	0.2	4.1	0.4	15.2	6.5	3.9	7.2	16.3	16.1	17.8	5.4	4.5
106	14	1	2	1	67.2	0.9	3.7	2.4	2.5	1.9	6.4	3.1	5.4	1.2	0.5	0.9	8.8	7.5
	14	1	2	2	17.1	12.5	10.4	5.0	5.4	9.6	9.8	0.8	5.9	13.7	4.8	11.3	14.7	9.2
	14	1	2	3	15.7	17.4	4.7	4.9	4.7	2.6	16.9	14.0	16.6	9.9	7.4	8.3	21.9	22.1
107	14	1	3	1	71.5	0.1	4.3	4.7	3.3	17.2	5.2	7.3	4.4	19.0	17.9	19.3	6.6	8.1
	14	1	3	2	13.8	23.1	6.9	4.7	1.6	6.0	13.0	5.6	17.8	15.5	8.3	17.2	9.7	1.1
	14	1	3	3	14.7	21.4	14.5	18.6	14.5	78.2	37.7	40.6	38.3	78.1	79.4	77.5	41.3	38.4
108	14	1	4	1	56.6	6.7	4.6	0.2	5.4	25.9	3.9	7.9	3.0	23.9	24.5	23.7	2.6	0.9
	14	1	4	2	8.6	17.1	0.9	4.0	6.7	23.0	15.6	11.3	20.8	23.3	25.8	20.1	16.9	10.2
	14	1	4	3	34.8	15.2	7.3	1.3	7.2	47.8	10.2	15.6	10.1	44.6	46.2	43.5	8.3	3.9
109	14	2	1	1	12.3	11.0	14.4	8.2	18.4	45.1	0.4	3.9	2.8	71.8	69.2	72.4	19.1	20.3
	14	2	1	2	6.6	11.1	19.5	13.9	4.7	50.2	22.6	16.7	9.2	42.7	44.7	51.6	19.0	24.7
	14	2	1	3	81.1	2.6	3.8	2.4	3.2	11.0	1.8	0.8	1.2	14.4	14.2	15.2	4.5	5.1
110	14	2	2	1	28.8	5.3	2.6	5.4	0.8	23.9	12.2	6.4	15.2	14.2	12.3	14.4	1.4	3.4
	14	2	2	2	7.3	9.1	5.5	11.3	7.9	48.8	11.1	0.7	21.8	39.8	46.1	35.8	7.7	9.6
	14	2	2	3	63.9	3.4	0.5	1.2	0.5	5.2	4.2	2.8	4.4	1.8	0.3	2.4	0.3	2.6
111	14	2	3	1	35.0	4.9	11.3	9.3	9.0	19.7	1.6	0.2	0.6	23.8	22.9	23.6	4.8	5.8
	14	2	3	2	3.5	64.0	12.9	17.6	2.2	83.7	21.9	14.1	7.2	71.7	91.1	82.9	27.2	22.1
	14	2	3	3	61.5	0.9	5.7	4.3	5.2	6.4	0.4	0.7	0.7	9.4	7.8	8.7	1.2	2.0
112	14	2	4	1	15.4	3.6	9.1	0.6	7.7	62.3	11.9	0.5	11.1	59.4	59.1	59.3	5.9	4.9
	14	2	4	2	0.8	188.1	43.4	42.3	23.4	0.8	46.1	47.6	27.1	2.3	6.6	9.4	49.8	54.8
	14	2	4	3	83.8	1.2	1.3	0.5	1.2	11.4	1.7	0.6	1.8	10.9	10.8	10.8	0.6	0.4

Table 7.1

Venezuela

1990 Census Proportions and Absolute Relative Residuals (ARB)
for Thirteen Different Models

Categories					Census	MODELS												
SG	STT	Sex	Age	LF	Prop.	(a)	(b)	(c)	(d)	SPREE				Unsaturated SPREE				
										(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
113	15	1	1	1	54.5	3.5	4.7	0.2	5.8	16.2	1.4	7.0	8.2	14.8	18.9	6.6	3.9	8.1
	15	1	1	2	11.6	12.0	0.6	1.0	2.2	4.6	3.5	8.4	1.1	21.0	8.6	16.0	0.0	9.4
	15	1	1	3	33.9	9.8	7.8	0.6	10.1	27.6	1.1	8.3	13.6	31.0	33.4	16.1	6.3	9.8
114	15	1	2	1	83.4	4.7	1.9	0.6	2.3	2.0	5.3	2.4	3.8	3.8	1.2	1.7	4.7	2.0
	15	1	2	2	8.8	1.2	2.0	1.9	1.4	5.3	25.4	5.2	24.4	21.4	5.5	20.2	25.7	3.6
	15	1	2	3	7.7	49.7	18.8	8.8	22.8	15.8	28.1	20.0	13.4	16.2	6.3	4.8	21.0	17.8
115	15	1	3	1	84.7	2.0	0.6	0.1	0.3	7.8	1.5	5.6	0.3	9.1	6.7	9.7	1.4	4.2
	15	1	3	2	7.5	18.1	2.3	1.1	2.8	4.1	21.8	3.1	23.3	23.2	3.9	25.1	18.9	3.7
	15	1	3	3	7.8	39.6	3.8	2.4	0.6	80.3	37.9	58.0	26.1	76.5	68.4	80.9	3.6	41.6
116	15	1	4	1	62.1	9.7	11.6	3.8	10.7	38.2	23.3	11.1	18.3	32.6	35.2	36.0	18.6	20.5
	15	1	4	2	5.1	10.0	8.9	0.3	7.6	11.3	9.2	8.9	10.6	10.9	12.7	10.1	12.6	12.9
	15	1	4	3	32.7	16.8	23.3	7.1	21.5	70.7	42.7	22.4	33.0	63.6	64.9	70.0	33.4	40.9
117	15	2	1	1	26.5	28.1	9.6	0.4	8.2	15.9	19.5	7.3	11.6	29.1	21.1	19.2	13.9	17.9
	15	2	1	2	7.3	29.8	7.5	1.6	5.2	8.2	10.0	15.3	5.2	4.3	9.9	7.7	7.8	22.9
	15	2	1	3	66.2	14.5	4.7	0.3	3.8	7.3	8.9	4.6	5.2	11.2	9.5	6.8	6.4	9.7
118	15	2	2	1	44.0	3.2	2.8	1.3	1.4	23.3	12.5	4.0	19.5	10.1	24.1	2.8	13.4	8.0
	15	2	2	2	5.7	4.6	3.2	3.3	5.2	60.0	29.6	13.1	28.7	48.2	56.6	49.2	23.8	11.6
	15	2	2	3	50.3	3.3	2.1	1.5	0.6	13.6	7.6	2.0	13.8	3.4	14.6	3.1	9.0	8.4
119	15	2	3	1	45.5	7.1	1.7	0.3	3.0	30.9	19.1	1.0	12.6	9.5	26.8	16.1	14.1	3.8
	15	2	3	2	3.1	36.0	4.6	2.9	2.8	37.4	12.4	15.8	9.4	36.7	42.3	32.9	9.0	15.2
	15	2	3	3	51.4	8.5	1.2	0.4	2.5	25.1	17.6	0.1	11.7	6.2	21.1	12.3	13.0	4.3
120	15	2	4	1	21.6	15.2	9.7	9.5	10.6	54.2	2.4	26.8	6.4	47.9	46.4	53.9	11.2	2.4
	15	2	4	2	0.8	152.1	26.8	1.6	29.1	56.3	42.0	44.6	39.3	12.3	29.1	18.3	9.1	7.6
	15	2	4	3	77.6	5.8	3.0	2.6	3.2	15.6	1.1	7.9	1.4	13.4	13.2	15.2	3.0	0.6
121	16	1	1	1	51.1	4.0	12.8	7.5	3.6	19.8	1.9	2.5	6.1	25.8	21.0	25.3	3.3	7.3
	16	1	1	2	13.8	9.3	4.6	4.8	1.9	15.8	4.7	9.7	2.7	21.0	16.3	22.9	5.5	12.9
	16	1	1	3	35.1	2.1	20.5	12.9	4.5	35.0	4.6	7.5	10.0	45.9	37.0	45.9	6.9	15.7
122	16	1	2	1	77.1	3.7	0.1	2.6	2.8	3.4	2.6	2.9	4.7	4.1	5.7	3.5	0.7	2.4
	16	1	2	2	12.2	7.9	6.8	3.1	7.1	7.5	12.8	5.0	11.1	18.8	11.1	15.7	10.0	7.1
	16	1	2	3	10.8	35.4	8.5	22.3	12.1	15.7	4.5	14.8	21.5	8.5	28.5	7.6	5.9	9.4
123	16	1	3	1	80.7	2.9	0.4	2.4	2.8	11.9	4.8	1.3	6.1	10.7	11.5	11.0	4.5	3.0
	16	1	3	2	9.5	18.0	1.7	0.3	2.5	17.2	4.8	0.9	5.6	14.0	14.3	16.5	7.2	2.7
	16	1	3	3	9.8	41.8	1.3	20.2	20.6	81.2	44.6	9.7	55.5	74.4	80.9	74.5	44.0	27.3
124	16	1	4	1	68.5	10.4	8.5	4.8	1.8	8.5	10.2	2.5	4.2	11.9	7.6	12.4	11.2	5.1
	16	1	4	2	6.3	11.4	7.2	0.3	11.3	11.1	5.6	16.2	12.5	23.0	13.5	19.4	6.7	14.3
	16	1	4	3	25.2	31.0	21.4	13.1	2.1	25.9	26.4	2.8	8.3	38.0	24.0	38.4	28.6	10.2
125	16	2	1	1	17.0	12.5	11.5	23.3	6.2	67.4	20.8	18.2	5.1	58.0	69.5	56.0	23.7	9.6
	16	2	1	2	6.7	20.2	6.1	10.3	3.1	35.0	16.8	9.4	3.2	19.2	28.3	26.4	10.3	5.4
	16	2	1	3	76.3	4.6	3.1	6.1	1.7	18.0	6.1	4.9	1.4	14.6	17.9	14.8	6.2	2.6
126	16	2	2	1	30.1	27.0	18.7	6.7	3.7	6.2	2.6	4.2	14.2	32.8	17.7	31.2	7.8	19.1
	16	2	2	2	6.0	12.2	9.3	6.2	3.4	55.2	20.7	5.9	29.5	40.6	49.7	36.6	11.8	9.4
	16	2	2	3	63.9	13.8	9.7	3.7	2.1	2.3	3.2	1.4	9.5	11.6	3.7	11.2	2.6	9.9
127	16	2	3	1	31.6	31.1	23.5	6.5	7.9	25.7	49.9	9.1	36.1	4.8	23.2	3.5	47.2	24.1
	16	2	3	2	3.0	61.8	12.2	0.9	6.3	113.2	38.7	0.5	27.7	68.0	96.4	78.8	27.6	16.8
	16	2	3	3	65.5	17.8	11.9	3.2	4.1	17.5	25.8	4.4	18.7	5.4	15.6	5.3	24.0	12.4
128	16	2	4	1	17.0	16.0	9.9	21.8	9.3	62.6	12.3	16.0	26.4	51.6	62.2	52.5	12.3	7.1
	16	2	4	2	1.0	109.0	4.3	1.8	6.4	33.2	2.1	25.5	16.6	26.8	44.1	22.2	20.2	6.7
	16	2	4	3	82.0	4.7	2.1	4.5	2.0	13.4	2.6	3.6	5.7	11.0	13.4	11.2	2.8	1.6

Table 7.1

Venezuela

1990 Census Proportions and Absolute Relative Residuals (ARB)
for Thirteen Different Models

Categories					Census Prop.	MODELS													
SG	STT	Sex	Age	LF		(a)	(b)	(c)	(d)	SPREE				Unsaturated SPREE					
										(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
129	17	1	1	1	34.0	16.3	3.8	3.8	2.9	7.2	16.8	15.1	19.3	6.3	7.1	8.6	16.7	16.4
	17	1	1	2	18.3	0.4	13.4	1.7	14.5	25.9	15.1	8.5	20.1	32.0	29.4	29.1	18.7	18.0
	17	1	1	3	47.7	11.7	2.4	3.3	3.5	15.1	6.2	7.5	6.1	16.8	16.3	17.2	4.7	4.8
130	17	1	2	1	65.7	0.6	5.5	4.3	5.1	1.3	9.6	7.3	11.0	3.6	3.8	2.1	12.0	10.5
	17	1	2	2	15.5	3.1	0.3	2.1	0.8	19.6	2.5	5.9	8.8	16.3	12.6	22.4	8.3	0.2
	17	1	2	3	18.8	0.4	18.8	16.7	17.1	20.5	31.5	30.2	31.2	26.1	23.6	25.7	34.9	36.3
131	17	1	3	1	72.5	2.0	2.3	4.4	2.1	18.0	6.5	7.8	7.8	18.9	19.8	17.9	9.1	7.0
	17	1	3	2	11.6	8.5	11.4	1.2	10.9	3.3	16.1	7.0	8.2	8.2	12.7	2.7	5.5	16.6
	17	1	3	3	15.9	15.4	19.0	19.2	17.4	79.8	41.4	40.9	41.9	80.3	81.3	79.9	45.7	44.1
132	17	1	4	1	58.0	10.3	8.0	1.5	8.7	24.5	3.0	6.7	4.1	23.1	24.2	21.8	2.5	0.7
	17	1	4	2	5.8	22.9	50.3	1.2	48.9	76.9	65.6	17.8	54.5	68.7	70.0	78.0	58.9	63.8
	17	1	4	3	36.2	12.8	4.8	2.3	6.1	51.4	15.3	13.6	15.3	47.9	49.9	47.2	13.3	9.0
133	17	2	1	1	11.3	6.1	21.2	12.1	23.2	80.7	23.6	22.9	18.2	93.5	89.9	102.7	32.4	40.2
	17	2	1	2	8.5	32.3	8.3	3.8	6.1	9.6	27.0	20.5	12.0	3.8	7.5	9.4	12.1	26.0
	17	2	1	3	80.1	4.3	2.1	2.1	2.6	10.4	0.5	1.1	1.3	13.7	13.5	13.5	3.3	2.9
134	17	2	2	1	29.0	2.3	5.3	9.2	3.6	23.5	11.4	7.6	7.5	16.7	13.0	20.4	1.2	9.1
	17	2	2	2	8.0	1.5	4.3	3.9	2.2	56.8	25.8	18.9	10.4	48.8	51.8	55.8	17.9	24.1
	17	2	2	3	63.0	0.9	3.0	4.7	1.9	3.6	2.0	1.1	2.1	1.5	0.6	2.3	1.7	1.1
135	17	2	3	1	33.7	3.4	9.7	8.9	8.4	11.2	7.1	4.5	3.9	18.2	15.9	15.2	2.1	2.9
	17	2	3	2	3.3	71.5	18.6	4.0	21.0	44.4	6.4	16.3	14.4	69.5	65.2	47.1	8.6	3.2
	17	2	3	3	63.0	2.0	4.2	4.5	3.4	3.7	3.5	1.6	2.9	6.1	5.1	5.6	1.6	1.4
136	17	2	4	1	14.2	1.5	4.0	5.8	3.2	55.4	2.8	9.3	2.0	54.3	53.4	52.8	5.5	8.8
	17	2	4	2	0.8	170.7	35.6	7.4	37.8	35.2	8.2	35.9	10.4	8.1	4.7	20.2	32.3	11.5
	17	2	4	3	85.0	2.0	0.3	0.9	0.2	9.6	0.4	1.2	0.2	9.2	9.0	9.0	1.2	1.6
137	18	1	1	1	46.4	5.8	4.3	0.7	1.8	17.8	2.6	8.6	10.3	14.9	15.4	16.0	4.4	3.7
	18	1	1	2	11.6	24.6	8.4	5.0	8.7	30.0	21.4	14.1	17.9	17.4	25.6	17.4	16.1	6.7
	18	1	1	3	42.1	13.1	2.4	0.6	0.4	27.8	3.0	5.7	6.5	21.2	24.0	22.5	0.4	2.2
138	18	1	2	1	78.3	5.0	1.4	1.7	0.6	4.0	1.9	1.0	3.6	2.9	1.8	3.2	3.7	2.2
	18	1	2	2	10.8	1.6	0.3	0.5	1.5	23.6	2.3	11.5	3.3	14.9	21.7	15.8	3.2	7.1
	18	1	2	3	10.8	37.8	10.6	11.6	6.1	5.5	11.2	18.8	22.8	6.3	8.3	7.6	23.5	8.9
139	18	1	3	1	81.5	3.8	0.4	0.7	0.3	12.3	5.1	4.0	6.4	11.1	11.7	11.0	3.9	3.2
	18	1	3	2	9.2	19.4	3.3	2.8	4.5	32.3	16.6	2.5	18.1	24.5	30.4	23.7	14.2	5.6
	18	1	3	3	9.2	53.0	6.8	8.8	2.1	76.0	28.1	33.1	38.6	73.8	73.3	73.1	20.4	22.5
140	18	1	4	1	62.3	1.8	0.3	1.7	2.1	20.9	1.5	4.4	6.3	25.5	22.0	25.0	3.0	5.4
	18	1	4	2	7.1	26.0	10.3	4.2	10.6	15.0	7.9	3.2	9.9	7.2	14.6	6.5	9.5	10.4
	18	1	4	3	30.6	9.7	1.9	4.5	1.8	46.1	4.9	8.2	15.1	50.2	48.2	49.3	8.3	8.6
141	18	2	1	1	19.5	24.9	4.2	1.6	7.8	37.6	3.4	7.4	13.0	43.4	40.3	45.6	0.6	1.9
	18	2	1	2	7.2	29.7	6.4	7.4	5.8	10.9	27.0	15.9	30.1	2.1	12.7	3.3	27.6	19.5
	18	2	1	3	73.4	9.5	1.8	0.3	2.6	8.9	3.5	0.4	6.4	11.3	9.5	11.8	2.8	1.4
142	18	2	2	1	33.4	13.6	5.9	3.6	3.2	50.4	37.5	17.0	28.6	23.4	32.7	24.6	20.7	13.2
	18	2	2	2	6.2	1.8	0.9	0.8	0.6	31.5	17.9	4.4	16.5	50.6	44.1	51.2	3.3	15.4
	18	2	2	3	60.4	7.7	3.2	1.9	1.8	24.6	22.6	9.0	17.5	7.7	13.5	8.3	11.1	5.7
143	18	2	3	1	37.1	10.5	3.8	1.5	0.9	14.9	2.1	2.5	6.0	7.7	9.6	6.6	7.8	10.4
	18	2	3	2	2.8	60.1	10.8	9.0	12.5	27.9	18.5	2.6	20.3	53.2	38.4	51.3	10.9	2.2
	18	2	3	3	60.0	9.3	2.9	1.3	1.2	7.9	0.4	1.4	4.7	2.2	4.1	1.7	4.3	6.3
144	18	2	4	1	16.9	14.9	8.7	6.1	3.5	56.6	1.2	8.9	12.8	49.2	53.4	48.7	3.8	13.8
	18	2	4	2	0.7	164.8	31.9	38.3	31.6	29.2	2.0	10.0	8.0	13.5	2.7	14.3	40.0	16.2
	18	2	4	3	82.3	4.5	2.1	1.6	1.0	11.9	0.3	1.7	2.7	10.2	10.9	10.1	1.1	3.0

Table 7.1

Venezuela

1990 Census Proportions and Absolute Relative Residuals (ARB)
for Thirteen Different Models

Categories					Census Prop.	MODELS													
SG	STT	Sex	Age	LF		(a)	(b)	(c)	(d)	SPREE				Unsaturated SPREE					
										(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
145	19	1	1	1	45.8	1.4	11.6	9.0	5.3	16.2	3.8	1.2	9.0	23.6	17.6	22.7	2.3	3.1
	19	1	1	2	13.3	11.8	3.3	1.9	4.9	16.7	5.1	1.4	11.3	20.3	21.4	14.9	10.1	2.7
	19	1	1	3	41.0	2.3	14.0	9.4	7.5	23.5	2.6	1.8	6.3	33.0	26.6	30.1	0.7	4.3
146	19	1	2	1	73.6	2.8	1.3	0.0	3.9	5.2	1.9	3.2	4.3	2.0	5.3	2.2	1.9	4.4
	19	1	2	2	11.6	1.3	0.8	7.4	4.7	42.4	16.4	14.1	6.2	27.0	37.8	33.5	12.4	7.9
	19	1	2	3	14.8	15.1	7.3	5.7	15.9	7.5	22.3	27.2	26.2	11.3	3.1	15.1	19.2	28.1
147	19	1	3	1	79.5	4.8	0.8	0.3	1.3	13.2	4.8	1.4	6.5	11.5	14.1	11.1	6.2	2.2
	19	1	3	2	8.7	7.4	12.2	0.9	5.6	14.6	6.2	19.6	4.1	1.9	21.6	3.1	2.8	25.9
	19	1	3	3	11.8	37.5	3.6	2.4	13.1	78.3	36.6	23.9	41.0	75.8	79.2	76.7	39.4	34.1
148	19	1	4	1	61.7	7.4	5.1	5.7	0.4	19.2	1.0	4.1	2.3	21.0	16.7	21.2	4.3	1.0
	19	1	4	2	5.6	1.2	20.5	3.2	16.3	56.3	48.5	15.8	36.7	41.2	49.7	49.0	39.7	40.8
	19	1	4	3	32.8	14.1	6.1	11.4	2.1	45.7	6.4	10.4	10.6	46.5	39.9	48.3	1.4	8.8
149	19	2	1	1	13.2	4.3	23.1	30.4	9.0	82.0	27.6	23.6	15.3	84.0	99.2	84.5	43.0	27.7
	19	2	1	2	7.3	33.9	10.9	3.4	4.5	10.3	25.0	20.2	7.6	3.5	3.2	11.0	12.7	26.8
	19	2	1	3	79.5	3.8	2.8	4.7	1.1	12.7	2.3	2.1	1.8	14.3	16.8	13.0	6.0	2.1
150	19	2	2	1	30.4	12.0	4.1	0.1	5.5	11.5	1.7	6.8	6.5	20.8	10.1	20.2	0.0	9.4
	19	2	2	2	7.3	12.9	15.3	11.0	7.2	58.5	27.2	24.5	9.3	53.1	51.8	59.9	16.3	30.4
	19	2	2	3	62.3	4.3	0.2	1.3	3.5	1.3	2.3	0.4	4.3	3.9	1.2	2.8	1.9	1.0
151	19	2	3	1	34.5	7.2	0.2	0.6	8.8	1.8	21.9	4.1	13.9	12.6	2.7	13.0	17.3	4.8
	19	2	3	2	2.9	59.9	10.5	2.6	21.5	39.5	9.2	9.1	15.3	66.6	45.4	42.9	4.7	6.2
	19	2	3	3	62.6	6.7	0.6	0.4	3.9	2.8	11.7	1.8	8.4	3.9	0.6	5.2	9.3	2.4
152	19	2	4	1	12.3	37.7	30.0	26.1	14.2	55.4	5.6	19.9	5.2	41.0	49.9	42.1	16.0	32.4
	19	2	4	2	0.6	229.9	64.6	27.4	75.3	28.2	5.8	14.8	29.3	16.7	10.8	1.0	58.1	38.0
	19	2	4	3	87.2	6.8	4.6	3.9	2.5	8.0	0.8	2.7	0.5	5.7	6.9	5.9	2.6	4.8
153	20	1	1	1	39.1	9.1	1.8	4.1	1.5	4.3	17.3	14.0	20.2	13.7	8.4	13.2	13.4	8.8
	20	1	1	2	17.3	5.7	8.4	3.9	4.6	15.4	4.6	7.6	5.1	21.8	16.3	20.9	5.0	9.8
	20	1	1	3	43.6	10.4	5.0	5.2	0.5	10.0	13.6	9.5	16.1	21.0	14.0	20.1	10.1	4.0
154	20	1	2	1	70.9	3.0	1.1	0.5	1.9	3.3	4.2	3.0	5.1	3.0	3.6	2.8	4.2	4.7
	20	1	2	2	14.6	1.5	0.5	0.3	2.4	20.9	1.0	8.8	2.8	27.8	18.9	28.6	2.9	4.6
	20	1	2	3	14.5	16.4	6.0	2.1	11.8	4.9	19.4	23.4	22.0	13.4	1.4	15.0	17.6	27.6
155	20	1	3	1	75.2	2.3	1.4	1.2	2.1	13.8	4.0	3.5	4.8	14.6	14.6	14.5	5.1	4.9
	20	1	3	2	12.7	19.4	2.8	5.2	1.1	5.4	16.1	9.8	13.5	13.0	10.9	12.6	8.5	5.8
	20	1	3	3	12.1	34.5	5.9	1.8	12.0	79.9	41.6	31.8	44.3	77.0	79.2	77.2	40.4	36.2
156	20	1	4	1	61.3	9.5	7.5	3.0	5.2	17.0	3.6	2.3	2.0	18.5	16.5	18.8	4.0	1.1
	20	1	4	2	8.0	12.1	6.7	3.5	10.4	23.4	16.1	12.5	14.8	28.0	18.5	28.7	10.7	21.1
	20	1	4	3	30.7	22.0	13.3	6.8	7.6	40.0	3.0	7.8	0.2	44.1	37.7	44.9	5.3	3.2
157	20	2	1	1	13.6	11.1	13.9	11.6	8.3	56.9	9.3	3.2	4.5	73.2	85.2	72.7	33.9	20.6
	20	2	1	2	7.9	23.0	3.1	8.5	4.3	26.7	5.3	0.7	8.4	24.2	34.8	21.8	14.3	0.7
	20	2	1	3	78.4	4.3	2.7	2.9	1.0	12.6	2.2	0.5	1.6	15.2	18.3	14.8	7.3	3.6
158	20	2	2	1	29.4	11.4	3.5	1.1	0.3	17.1	5.9	7.3	2.3	23.2	12.1	22.3	1.8	10.4
	20	2	2	2	7.5	7.3	4.2	0.6	2.4	49.2	11.8	1.4	8.8	39.9	48.5	40.9	11.0	1.1
	20	2	2	3	63.1	6.2	2.1	0.5	0.4	2.1	1.4	3.6	0.0	6.0	0.1	5.5	0.5	5.0
159	20	2	3	1	33.1	8.1	1.3	2.5	2.7	1.2	19.5	5.9	15.8	10.6	6.2	11.2	14.9	7.9
	20	2	3	2	3.1	85.0	27.9	20.2	19.5	100.4	31.5	11.5	37.0	98.7	114.8	95.1	42.4	28.8
	20	2	3	3	63.7	8.4	2.1	0.3	0.5	4.3	11.7	3.6	10.0	0.7	2.5	1.1	9.8	5.5
160	20	2	4	1	15.3	6.8	1.0	12.0	5.6	57.9	2.4	17.5	7.8	53.1	59.2	53.8	5.3	5.9
	20	2	4	2	0.9	181.7	40.6	32.6	28.4	33.0	3.5	1.5	1.8	3.9	13.3	1.5	23.3	41.3
	20	2	4	3	83.8	3.1	0.6	2.5	0.7	10.9	0.5	3.2	1.5	9.7	10.9	9.8	0.7	1.5

Table 7.1 **11/12**

Venezuela

1990 Census Proportions and Absolute Relative Residuals (ARB)
for Thirteen Different Models

Categories					Census Prop.	MODELS												
SG	STT	Sex	Age	LF		(a)	(b)	(c)	(d)	SPREE				Unsaturated SPREE				
										(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
161	21	1	1	1	43.1	5.3	5.0	3.8	0.5	13.7	6.5	4.7	11.9	15.2	12.0	15.7	7.7	4.4
	21	1	1	2	14.6	13.4	1.4	1.7	0.0	24.1	13.1	12.2	12.5	21.5	24.1	21.1	14.1	10.3
	21	1	1	3	42.2	10.1	5.5	4.5	0.5	22.3	2.1	0.6	7.8	23.0	20.6	23.4	3.0	1.0
162	21	1	2	1	74.8	4.7	0.8	1.7	0.6	5.0	1.4	2.5	3.0	3.1	4.9	3.2	1.7	3.1
	21	1	2	2	12.8	0.9	1.0	3.2	0.5	29.0	5.7	2.8	3.1	20.2	27.9	21.2	4.1	1.8
	21	1	2	3	12.3	29.5	4.1	6.7	3.4	0.3	14.6	18.0	21.3	2.5	0.8	2.5	14.3	16.7
163	21	1	3	1	80.0	5.3	1.8	1.4	0.5	11.9	3.3	2.3	4.5	10.7	11.7	10.6	3.3	1.9
	21	1	3	2	10.3	12.8	5.0	0.1	3.3	22.2	2.7	0.6	5.7	13.8	20.1	13.0	1.2	6.9
	21	1	3	3	9.8	56.7	9.3	11.6	0.9	74.4	23.9	19.3	31.1	73.4	75.0	73.1	25.6	23.0
164	21	1	4	1	61.2	5.1	3.1	2.2	0.2	21.5	1.0	2.6	4.4	23.4	22.5	23.2	2.5	2.9
	21	1	4	2	7.8	20.2	3.0	0.5	2.8	11.2	6.6	5.0	5.7	6.3	10.2	7.1	5.3	1.8
	21	1	4	3	31.0	15.2	6.9	4.4	0.4	45.2	3.7	6.4	10.0	47.8	46.9	47.6	6.2	6.2
165	21	2	1	1	15.7	15.8	7.7	10.2	0.5	80.0	27.2	24.1	18.5	64.4	69.8	65.9	19.8	15.2
	21	2	1	2	7.4	23.4	2.6	3.3	2.2	14.2	3.0	2.5	0.7	13.8	10.2	11.9	8.0	7.3
	21	2	1	3	77.0	5.5	1.8	2.4	0.3	17.6	5.2	4.7	3.7	14.4	15.2	14.6	3.3	2.4
166	21	2	2	1	30.2	15.8	7.8	3.9	2.3	20.4	8.5	11.8	2.4	28.1	22.3	28.6	9.4	15.8
	21	2	2	2	7.5	3.2	5.7	5.2	4.8	55.5	22.3	21.3	20.0	50.1	48.1	50.9	10.8	15.5
	21	2	2	3	62.3	7.3	3.1	1.3	0.5	3.2	1.4	3.2	1.3	7.6	5.0	7.7	3.2	5.8
167	21	2	3	1	34.0	12.1	5.2	3.2	0.3	3.0	16.0	8.1	9.9	6.1	2.8	5.6	15.8	12.5
	21	2	3	2	3.4	54.3	7.0	0.4	8.1	55.4	1.6	0.3	5.6	54.6	47.6	52.0	4.0	0.8
	21	2	3	3	62.6	9.5	3.2	1.7	0.3	1.4	8.8	4.4	5.7	0.3	1.1	0.2	8.4	6.8
168	21	2	4	1	16.2	9.0	3.2	7.4	4.9	53.0	6.0	14.9	2.8	51.5	53.7	51.4	3.1	8.7
	21	2	4	2	1.1	97.4	1.3	3.0	2.7	35.6	8.9	8.5	8.3	32.2	31.1	33.3	5.3	8.6
	21	2	4	3	82.7	3.1	0.6	1.5	1.0	10.9	1.1	2.8	0.7	10.5	10.9	10.5	0.5	1.6
169	22	1	1	1	33.2	39.4	24.5	15.4	1.4	10.8	12.3	38.4	5.1	2.7	14.3	4.4	8.7	23.7
	22	1	1	2	12.0	11.5	4.5	5.9	7.7	21.7	9.4	9.9	20.4	25.6	22.4	34.0	10.1	22.7
	22	1	1	3	54.8	26.4	13.9	8.1	2.6	11.3	5.4	21.2	7.6	4.0	13.6	10.1	3.1	9.4
170	22	1	2	1	68.4	6.3	10.8	12.0	1.3	9.6	15.5	12.5	10.6	6.5	11.2	5.2	17.3	12.4
	22	1	2	2	13.0	21.2	19.9	12.1	10.4	6.2	24.9	18.0	25.0	0.8	9.8	6.7	28.5	25.1
	22	1	2	3	18.6	8.5	25.8	35.8	2.6	31.1	39.9	33.4	21.5	23.5	34.5	14.6	43.7	28.1
171	22	1	3	1	71.2	7.8	12.3	13.6	4.3	22.4	9.2	11.2	4.7	25.7	23.2	25.6	12.4	14.7
	22	1	3	2	7.5	4.8	14.8	1.9	30.5	20.1	43.8	7.6	43.2	1.1	25.4	5.8	51.4	13.2
	22	1	3	3	21.3	24.2	46.5	46.2	25.0	82.0	46.3	40.1	30.9	85.5	86.6	83.5	59.8	53.9
172	22	1	4	1	52.8	9.8	12.3	15.3	5.5	53.3	30.1	12.2	18.6	41.1	54.8	38.0	33.2	11.5
	22	1	4	2	5.1	3.7	16.7	4.6	20.1	6.2	9.1	11.0	13.7	33.3	3.6	24.1	6.3	12.8
	22	1	4	3	42.1	11.8	17.5	19.8	4.4	66.0	36.7	16.6	21.6	55.5	68.3	50.6	40.9	16.0
173	22	2	1	1	19.5	33.6	15.0	23.8	16.9	25.7	52.5	33.3	38.2	20.2	3.7	28.2	35.8	10.3
	22	2	1	2	4.9	11.0	18.8	13.1	9.9	8.8	17.2	5.8	1.7	29.2	30.3	55.5	1.5	28.2
	22	2	1	3	75.6	9.4	2.6	5.3	5.0	6.0	14.6	8.9	9.7	7.1	1.0	10.9	9.1	0.8
174	22	2	2	1	42.1	17.6	23.8	17.9	1.7	1.1	9.9	17.2	6.6	16.1	6.3	11.1	16.4	19.0
	22	2	2	2	5.1	12.8	8.8	28.6	6.0	41.2	0.6	5.9	11.5	43.6	44.9	32.3	6.7	16.6
	22	2	2	3	52.8	12.9	18.1	11.6	1.9	3.1	7.9	13.1	6.4	17.0	9.4	12.0	13.8	13.6
175	22	2	3	1	45.6	17.4	23.1	23.4	2.3	28.0	15.4	21.4	0.1	36.3	35.5	32.3	23.4	18.2
	22	2	3	2	2.6	56.9	7.4	6.1	8.5	148.0	57.9	8.7	75.7	52.2	100.4	83.4	28.4	19.8
	22	2	3	3	51.8	12.4	20.0	20.9	2.5	17.1	10.6	18.4	3.9	29.3	26.1	24.2	19.2	15.0
176	22	2	4	1	26.1	34.2	38.2	34.4	15.7	57.9	11.9	34.4	9.1	74.0	62.3	71.6	20.0	35.0
	22	2	4	2	1.0	65.7	18.2	25.3	24.9	22.3	1.1	1.7	17.9	46.8	48.2	33.9	32.9	8.5
	22	2	4	3	72.9	11.3	14.0	12.7	6.0	21.1	4.3	12.3	3.5	27.2	23.0	26.1	7.6	12.7

Table 7.1

Venezuela

1990 Census Proportions and Absolute Relative Residuals (ARB)
for Thirteen Different Models

Categories					Census Prop.	MODELS												
SG	STT	Sex	Age	LF		(a)	(b)	(c)	(d)	SPREE				Unsaturated SPREE				
										(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
177	23	1	1	1	34.9	30.4	16.5	4.4	2.0	31.7	9.5	7.1	21.2	14.3	31.5	23.6	9.8	3.1
	23	1	1	2	14.5	3.6	17.1	2.9	21.5	26.2	11.4	5.8	23.7	35.9	23.9	41.3	9.6	28.7
	23	1	1	3	50.7	19.9	6.5	2.2	7.5	29.3	9.8	3.2	21.4	20.1	28.5	28.0	9.5	6.1
178	23	1	2	1	66.4	8.3	13.0	12.3	5.9	3.5	11.5	1.7	6.7	3.9	11.4	0.5	18.6	9.4
	23	1	2	2	11.2	3.8	2.0	0.3	3.9	11.3	10.2	17.2	14.8	29.4	3.1	28.4	23.4	4.3
	23	1	2	3	22.4	22.8	37.6	36.4	19.6	16.2	29.0	13.9	12.6	26.5	32.3	15.6	43.4	30.1
179	23	1	3	1	69.3	9.7	14.4	17.4	8.5	21.4	6.0	11.6	2.3	26.5	24.3	25.3	11.8	12.6
	23	1	3	2	6.5	16.1	40.5	3.6	50.5	66.3	92.2	10.8	83.2	33.1	57.4	35.5	85.1	59.3
	23	1	3	3	24.2	32.2	52.2	50.8	37.9	79.1	42.0	36.1	29.0	84.7	84.8	81.9	56.7	52.0
180	23	1	4	1	58.1	1.5	0.9	10.6	11.3	30.6	9.1	8.3	0.5	21.6	32.9	16.5	12.3	7.9
	23	1	4	2	4.5	13.4	38.0	3.6	36.9	47.4	38.6	8.3	26.5	68.5	43.7	65.6	35.3	46.3
	23	1	4	3	37.4	0.8	6.0	16.1	13.0	53.4	18.8	13.9	4.1	41.9	56.5	33.6	23.4	6.6
181	23	2	1	1	15.8	20.1	2.6	10.1	35.2	12.5	41.4	24.6	27.3	27.1	1.9	45.4	33.2	3.5
	23	2	1	2	6.5	30.6	6.8	6.6	5.2	2.4	19.5	6.6	9.1	1.2	14.1	7.4	10.3	10.5
	23	2	1	3	77.7	6.6	0.0	1.5	6.7	2.3	10.0	5.6	4.8	5.4	0.8	9.8	7.6	0.2
182	23	2	2	1	39.2	13.2	19.5	22.1	0.5	9.3	1.2	16.1	16.4	21.0	13.2	11.3	20.9	17.4
	23	2	2	2	5.8	3.5	0.3	0.5	3.6	66.3	40.9	28.4	26.3	48.4	60.2	44.7	32.1	3.3
	23	2	2	3	55.1	9.0	13.8	15.6	0.0	0.3	3.5	14.4	8.9	20.0	15.7	12.7	18.3	12.7
183	23	2	3	1	46.9	21.3	26.6	26.9	9.4	43.0	30.1	33.5	18.1	46.5	48.3	39.4	36.3	24.7
	23	2	3	2	2.4	82.2	25.3	10.5	19.3	90.8	25.0	34.3	59.6	73.8	103.4	88.0	33.5	25.5
	23	2	3	3	50.7	15.9	23.5	25.4	7.8	35.5	26.7	32.6	14.0	39.6	39.9	32.4	32.0	21.7
184	23	2	4	1	26.4	36.4	40.2	28.0	21.5	68.9	29.6	30.2	14.5	78.7	70.4	75.0	33.6	40.3
	23	2	4	2	0.5	278.2	87.7	42.0	90.6	5.3	31.4	2.0	74.2	15.0	26.9	29.6	71.8	85.5
	23	2	4	3	73.1	11.4	14.0	9.8	7.2	25.0	10.5	10.9	4.8	28.4	25.3	26.9	11.7	14.0

Table 7.2

Venezuela - 1990 Census Proportions and Subgroup Absolute Relative Residuals Averages (ARB) for Thirteen Different Models

Categories				MODELS												
SG	STT	Sex	Age	(a)	(b)	(c)	(d)	SPREE				Unsaturated SPREE				
								(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
1	1	1	1	27.4	15.1	11.8	2.9	7.2	7.5	9.7	4.2	1.2	4.4	0.8	9.0	13.8
2	1	1	2	6.8	5.5	10.1	8.4	20.3	12.3	10.7	5.1	10.8	22.0	12.2	14.4	4.5
3	1	1	3	24.0	3.8	9.1	10.5	31.6	2.1	11.6	13.2	32.0	30.9	31.3	4.2	2.9
4	1	1	4	12.6	15.2	11.8	2.9	40.1	24.0	8.4	17.0	31.1	38.9	30.7	23.0	14.1
5	1	2	1	28.0	8.7	13.0	1.0	12.8	13.4	10.9	6.6	15.0	8.0	15.8	16.3	8.1
6	1	2	2	6.2	10.7	2.9	3.4	13.2	10.5	4.7	9.9	17.9	15.7	17.8	7.3	10.2
7	1	2	3	9.6	11.1	4.0	3.0	27.2	22.5	4.5	16.1	24.5	26.4	23.6	19.5	10.5
8	1	2	4	23.0	10.0	14.9	5.7	25.0	4.0	8.0	10.7	37.0	33.6	36.7	12.9	9.3
9	2	1	1	13.8	1.5	2.2	1.0	18.9	1.0	3.4	2.1	19.4	18.0	21.2	0.4	2.8
10	2	1	2	13.8	2.7	2.9	0.7	12.0	5.8	2.0	3.6	12.6	10.3	12.6	6.0	4.1
11	2	1	3	30.8	2.1	2.6	0.6	30.3	2.4	2.7	5.2	32.0	29.6	32.2	3.1	3.6
12	2	1	4	12.4	3.8	1.8	1.9	24.0	6.2	1.2	4.1	22.0	24.6	22.3	6.2	5.3
13	2	2	1	16.8	5.2	3.8	2.2	33.8	9.4	6.6	3.2	31.4	32.5	34.4	6.9	7.7
14	2	2	2	10.3	5.9	6.1	0.4	19.1	1.6	2.8	6.8	14.6	15.6	13.7	3.1	8.2
15	2	2	3	29.7	7.1	6.9	0.9	44.0	13.4	7.9	7.4	35.5	42.8	38.6	12.3	9.0
16	2	2	4	61.7	13.6	15.4	6.9	31.0	7.3	10.7	4.4	24.7	27.9	25.0	11.1	16.2
17	3	1	1	3.2	12.1	5.4	8.9	23.7	7.7	5.3	6.1	31.8	31.2	31.2	11.9	11.4
18	3	1	2	3.0	7.8	4.2	9.3	11.2	9.3	6.7	12.4	10.8	7.2	11.5	9.8	8.3
19	3	1	3	4.8	18.7	6.4	19.6	31.1	18.6	2.0	16.2	30.0	30.6	29.9	19.6	19.2
20	3	1	4	9.0	12.1	2.3	8.3	27.5	7.1	3.3	3.4	27.2	26.0	27.8	6.6	11.9
21	3	2	1	18.2	11.0	11.3	5.7	11.5	20.8	9.8	16.8	29.8	29.0	28.7	7.1	11.2
22	3	2	2	8.8	4.3	4.2	1.2	28.1	12.3	8.1	3.9	21.9	24.0	21.8	7.5	4.6
23	3	2	3	47.1	17.6	3.8	16.3	48.6	27.6	1.7	39.3	41.2	43.9	40.1	19.1	16.3
24	3	2	4	45.4	2.4	6.3	7.0	21.3	15.8	9.2	24.0	30.5	30.5	31.4	2.2	1.8
25	4	1	1	17.0	3.9	1.9	3.5	13.7	3.7	1.5	4.3	15.5	14.1	14.4	3.6	3.6
26	4	1	2	22.7	9.7	8.3	6.7	24.8	8.0	4.0	7.6	26.8	30.1	23.9	11.9	7.4
27	4	1	3	46.4	10.8	10.8	9.8	31.2	9.0	6.1	8.5	28.0	28.3	29.3	12.3	9.9
28	4	1	4	11.5	3.4	1.0	5.6	19.7	3.9	2.9	3.5	23.3	20.2	25.0	2.8	4.9
29	4	2	1	21.8	1.6	2.9	2.7	25.7	1.0	2.0	1.0	23.5	25.2	23.5	5.3	1.7
30	4	2	2	7.1	3.3	3.7	2.5	16.3	2.8	1.6	3.2	21.9	19.4	19.9	1.3	2.0
31	4	2	3	15.5	8.1	5.7	4.9	17.1	6.6	3.0	6.1	14.9	14.7	17.6	8.5	5.8
32	4	2	4	31.4	5.8	2.5	3.8	32.0	0.9	3.4	1.3	33.6	35.3	33.6	6.0	3.5
33	5	1	1	9.1	10.0	6.9	4.5	20.6	4.0	4.0	6.8	28.4	23.6	28.2	7.4	8.3
34	5	1	2	17.5	4.5	8.2	5.0	13.2	6.2	1.6	6.8	19.8	22.5	16.6	10.4	4.7
35	5	1	3	30.2	3.4	5.4	10.5	35.3	8.4	6.6	11.9	33.0	33.9	34.6	7.2	5.8
36	5	1	4	23.7	14.5	8.7	4.9	11.0	20.8	9.0	13.4	14.2	9.8	16.6	23.1	14.4
37	5	2	1	13.8	8.6	11.9	2.2	36.7	10.6	6.8	4.3	35.8	42.3	37.8	15.9	10.1
38	5	2	2	16.9	10.7	4.4	1.2	18.7	6.2	2.3	13.6	24.1	21.7	20.4	4.7	12.6
39	5	2	3	34.9	10.9	3.8	1.4	45.4	31.1	6.0	22.2	23.8	36.3	30.0	24.4	13.0
40	5	2	4	53.9	11.1	20.0	3.4	31.9	4.4	21.2	7.0	25.8	34.4	24.5	4.9	12.7
41	6	1	1	15.7	1.2	1.0	3.3	19.4	0.8	5.3	5.5	17.2	15.2	22.4	3.0	3.7
42	6	1	2	15.7	3.1	4.7	1.2	17.9	13.4	9.2	8.2	17.2	15.5	18.2	10.3	7.5
43	6	1	3	28.4	2.3	2.0	2.0	33.7	8.4	5.3	13.0	32.2	31.3	32.6	9.4	4.7
44	6	1	4	18.5	9.9	1.3	7.8	19.9	16.7	3.6	11.1	20.5	21.5	20.5	12.6	14.6
45	6	2	1	17.9	3.5	1.8	0.7	41.0	14.6	8.6	5.4	29.0	31.0	36.7	6.0	10.2
46	6	2	2	12.7	6.9	3.8	3.6	19.3	1.6	7.3	6.0	19.1	19.5	17.4	2.8	13.5
47	6	2	3	26.7	3.4	3.8	0.2	42.3	17.9	7.6	8.4	30.8	39.7	37.5	18.9	10.0
48	6	2	4	38.8	4.9	7.3	8.9	38.2	9.2	9.2	16.9	34.4	37.9	29.8	10.1	8.4
49	7	1	1	14.3	1.3	1.8	1.4	16.8	1.1	2.0	1.9	18.3	17.3	17.3	0.6	0.9
50	7	1	2	14.3	1.7	2.9	1.3	13.2	2.5	2.8	1.2	17.3	15.8	15.9	1.0	1.0
51	7	1	3	34.6	3.7	3.7	2.7	30.6	4.6	2.3	3.7	30.2	29.6	30.6	5.4	2.9
52	7	1	4	11.0	4.0	0.5	4.1	21.4	2.8	1.3	2.1	24.0	21.1	24.7	2.1	4.9
53	7	2	1	22.6	5.2	2.6	4.4	22.4	4.0	2.8	2.8	21.9	22.5	21.3	3.9	5.8
54	7	2	2	9.5	3.9	1.6	2.8	17.2	5.0	3.9	3.4	19.9	20.2	19.1	4.0	3.4
55	7	2	3	21.7	3.1	1.6	3.9	24.4	1.8	0.6	4.1	22.3	23.1	23.1	2.3	2.3
56	7	2	4	38.2	2.1	1.9	2.7	30.6	3.4	4.8	1.2	31.6	32.5	32.1	2.1	1.2

Table 7.2

Venezuela - 1990 Census Proportions and Subgroup Absolute Relative Residuals Averages (ARB) for Thirteen Different Models

Categories				MODELS												
SG	STT	Sex	Age	(a)	(b)	(c)	(d)	SPREE				Unsaturated SPREE				
								(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
57	8	1	1	6.0	9.0	6.8	3.4	20.2	7.3	3.5	6.3	28.6	22.8	26.9	7.3	7.4
58	8	1	2	15.4	2.8	4.3	1.8	13.8	9.8	2.6	8.7	18.6	15.2	16.5	11.9	1.6
59	8	1	3	22.6	4.0	4.2	7.7	30.8	8.9	4.8	10.3	32.0	30.5	32.8	7.1	5.4
60	8	1	4	11.1	5.0	3.4	3.5	22.6	4.4	1.6	4.4	21.8	19.6	22.6	7.4	3.5
61	8	2	1	13.8	8.1	11.4	4.9	35.1	8.8	5.9	6.6	35.3	42.6	33.7	15.7	6.6
62	8	2	2	11.8	5.7	2.7	0.2	22.7	9.2	2.1	9.5	20.0	21.2	18.5	6.4	4.6
63	8	2	3	33.8	7.8	6.2	7.7	43.5	16.0	6.5	13.9	36.4	41.5	37.7	13.2	8.4
64	8	2	4	68.3	14.4	12.9	8.5	32.0	9.0	10.9	9.1	22.0	29.0	23.0	4.5	13.1
65	9	1	1	10.1	4.7	3.5	2.7	14.5	6.9	2.3	5.9	24.0	18.6	22.0	3.5	2.6
66	9	1	2	14.4	3.3	4.7	1.6	12.3	7.4	2.8	9.0	12.2	11.3	12.6	10.2	2.4
67	9	1	3	24.5	4.6	5.5	4.4	30.3	6.4	3.9	5.2	33.0	32.4	31.9	3.8	4.4
68	9	1	4	7.9	1.7	2.5	3.4	24.2	5.5	0.8	4.4	25.2	24.2	27.1	5.5	3.3
69	9	2	1	17.5	4.3	6.3	2.8	22.7	3.2	4.2	3.6	30.7	38.1	26.2	11.9	1.5
70	9	2	2	6.1	7.0	7.3	4.5	22.2	13.0	2.5	10.7	13.7	18.7	16.5	9.2	4.3
71	9	2	3	42.2	15.2	13.2	7.9	43.8	14.7	7.4	17.9	46.5	49.0	38.6	17.0	9.5
72	9	2	4	84.3	22.9	16.9	15.1	28.4	6.3	6.5	8.6	24.1	22.9	20.8	13.3	16.8
73	10	1	1	10.1	13.8	9.5	7.8	22.6	6.4	5.4	4.3	32.8	24.1	27.1	9.3	8.9
74	10	1	2	8.6	3.9	2.6	9.9	12.4	3.7	4.4	6.1	15.0	12.2	10.1	4.9	6.8
75	10	1	3	20.9	9.2	1.5	15.5	32.0	9.9	2.9	12.3	35.7	32.7	38.7	11.4	14.5
76	10	1	4	22.2	13.1	6.7	4.5	21.7	16.5	6.3	12.9	17.3	10.4	22.3	31.8	8.5
77	10	2	1	11.4	11.3	17.2	3.3	39.1	14.0	9.4	9.4	40.1	52.4	35.6	25.4	7.3
78	10	2	2	15.7	9.7	2.0	1.7	29.2	16.0	9.5	13.8	23.7	24.2	17.2	8.1	7.1
79	10	2	3	30.9	8.0	1.5	1.2	39.8	27.1	3.4	23.1	22.6	33.7	30.1	20.4	5.5
80	10	2	4	46.2	5.5	14.8	4.7	55.9	38.1	16.9	40.9	28.7	43.9	30.4	17.7	1.9
81	11	1	1	7.3	7.5	4.9	3.0	15.7	8.1	1.5	8.4	26.4	19.4	20.7	5.0	3.4
82	11	1	2	17.2	3.9	5.1	0.3	26.6	9.6	4.1	9.9	20.3	30.9	15.3	12.2	0.3
83	11	1	3	36.0	4.3	4.0	0.5	30.2	2.8	0.2	2.6	29.1	28.6	30.9	8.5	0.2
84	11	1	4	9.1	0.3	3.6	4.3	23.3	5.5	1.5	5.5	22.7	19.8	25.4	6.2	4.6
85	11	2	1	17.8	3.4	7.2	1.0	27.7	2.9	2.6	4.3	29.8	39.8	24.8	14.3	1.9
86	11	2	2	8.2	3.9	2.4	1.4	18.9	4.7	4.0	3.4	22.3	16.7	19.7	2.7	2.0
87	11	2	3	26.2	3.3	2.7	4.6	25.5	0.9	0.5	1.1	29.1	28.1	29.9	2.2	3.9
88	11	2	4	73.9	18.9	13.2	13.5	32.6	2.0	3.3	3.2	20.5	25.2	21.9	12.4	13.5
89	12	1	1	10.0	8.6	5.7	3.4	29.3	9.8	6.5	6.9	25.2	27.8	25.1	9.1	9.9
90	12	1	2	3.5	13.1	11.0	13.2	22.5	16.8	10.8	16.4	16.9	23.6	18.2	17.5	13.7
91	12	1	3	36.1	4.8	2.3	3.1	28.3	25.3	18.2	21.8	28.6	34.7	27.3	9.8	6.4
92	12	1	4	13.7	8.8	5.5	2.8	32.5	18.8	3.1	12.2	20.6	31.2	21.3	20.6	10.2
93	12	2	1	17.8	6.7	9.3	3.4	30.8	14.8	10.4	6.4	29.8	28.1	28.5	13.3	9.3
94	12	2	2	15.0	12.2	9.5	4.9	21.1	4.6	8.7	3.6	27.6	23.7	29.1	7.0	14.5
95	12	2	3	16.2	4.3	2.3	6.2	11.3	12.7	5.8	4.4	22.7	13.8	19.2	8.8	5.7
96	12	2	4	53.5	9.2	8.2	9.1	33.8	1.8	11.7	9.3	26.0	21.6	27.5	17.8	7.7
97	13	1	1	26.2	13.7	9.6	8.4	14.1	3.3	3.1	3.2	3.3	9.5	5.5	6.3	11.2
98	13	1	2	11.6	2.4	5.4	9.4	9.4	5.4	8.5	1.3	15.9	13.6	19.9	7.6	4.4
99	13	1	3	34.1	4.6	2.2	14.4	30.2	8.8	1.5	12.3	29.3	30.6	27.5	4.8	8.7
100	13	1	4	4.6	7.0	7.5	4.4	30.6	12.7	4.0	9.6	25.8	33.4	24.1	18.2	3.6
101	13	2	1	24.8	5.0	10.4	1.4	19.4	6.0	3.4	3.8	18.9	8.3	21.2	15.9	2.9
102	13	2	2	2.9	5.6	1.7	2.5	13.3	9.7	6.7	8.2	15.7	19.3	15.9	2.7	3.9
103	13	2	3	11.3	7.3	3.1	4.7	21.9	11.8	2.8	10.0	22.6	21.3	21.2	9.1	5.5
104	13	2	4	20.5	13.1	11.9	7.7	27.0	2.5	7.2	3.8	39.9	28.7	38.6	3.2	11.0
105	14	1	1	15.5	1.6	4.0	2.3	18.9	3.4	3.5	1.9	19.2	18.8	20.9	0.5	2.1
106	14	1	2	12.6	5.4	3.1	4.0	11.9	6.1	3.5	4.6	10.0	7.6	9.8	7.5	5.4
107	14	1	3	24.6	7.3	9.0	5.9	31.5	3.6	6.2	5.3	34.9	33.0	35.2	5.1	7.4
108	14	1	4	15.0	5.5	1.7	6.9	21.5	3.3	1.3	5.2	19.7	21.3	18.4	4.8	7.2
109	14	2	1	10.3	13.6	7.9	9.1	35.3	9.9	8.7	4.3	42.5	42.3	46.0	14.0	16.5
110	14	2	2	8.2	2.7	5.2	2.9	23.5	6.4	2.4	11.1	16.1	17.8	15.0	2.6	5.2
111	14	2	3	28.8	11.7	12.2	5.9	41.2	11.7	8.5	5.9	39.4	45.3	43.0	15.2	14.0
112	14	2	4	51.4	9.4	8.7	3.8	31.3	12.4	8.0	7.0	29.8	28.4	27.7	11.1	12.0

Table 7.2

Venezuela - 1990 Census Proportions and Subgroup Absolute Relative Residuals Averages (ARB) for Thirteen Different Models

Categories				MODELS												
SG	STT	Sex	Age	(a)	(b)	(c)	(d)	SPREE				Unsaturated SPREE				
								(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
113	15	1	1	8.6	5.4	0.5	7.0	19.0	8.1	2.7	12.9	22.5	20.4	12.9	9.4	3.5
114	15	1	2	23.8	9.2	5.4	11.6	2.7	11.9	0.9	10.8	25.7	7.1	15.8	8.8	1.3
115	15	1	3	31.2	2.3	0.7	4.5	25.3	12.8	14.1	13.4	30.3	19.2	33.3	23.3	5.4
116	15	1	4	14.0	15.2	4.2	14.2	39.8	24.5	10.6	19.9	32.7	37.2	35.7	20.7	21.4
117	15	2	1	25.2	5.7	0.8	4.9	17.7	10.0	3.2	8.0	18.5	20.8	12.4	8.2	11.4
118	15	2	2	4.9	5.8	3.7	3.9	27.3	9.7	1.5	13.7	14.9	26.6	15.2	8.2	9.4
119	15	2	3	19.1	1.8	0.5	1.3	40.5	15.4	4.3	12.4	27.0	39.7	29.7	13.5	7.0
120	15	2	4	71.5	18.1	7.1	18.0	37.8	10.5	21.0	13.3	21.3	23.3	22.1	10.2	11.1
121	16	1	1	4.3	13.4	8.6	1.8	25.6	5.4	8.3	4.8	33.1	26.8	33.8	6.9	13.8
122	16	1	2	17.3	4.1	9.5	5.0	19.6	8.6	1.6	4.5	20.9	26.9	19.3	12.5	4.7
123	16	1	3	26.8	1.9	7.1	9.2	34.4	10.2	8.3	15.6	29.9	33.2	30.9	10.6	1.2
124	16	1	4	22.2	12.6	8.7	2.1	9.7	23.4	9.0	12.7	13.1	8.0	13.4	24.3	13.2
125	16	2	1	15.0	6.7	13.6	4.3	43.3	17.1	13.3	5.5	33.5	41.6	35.3	15.9	8.2
126	16	2	2	18.5	12.5	4.4	2.6	19.1	8.0	1.1	17.0	24.8	20.6	22.8	3.3	14.3
127	16	2	3	44.4	18.2	3.8	7.9	59.2	41.4	6.4	30.5	31.6	51.3	36.0	35.8	20.4
128	16	2	4	57.2	11.7	18.2	1.3	31.6	5.3	24.6	9.7	24.4	35.6	23.1	5.0	13.7
129	17	1	1	9.8	4.7	3.6	4.5	21.0	2.4	3.7	3.7	23.7	22.8	23.4	4.3	4.2
130	17	1	2	7.1	7.5	5.2	8.4	11.7	3.3	3.7	5.7	7.3	7.1	10.0	7.6	5.1
131	17	1	3	16.7	9.2	8.0	9.5	28.8	9.0	6.0	7.0	31.0	33.2	28.6	8.3	10.5
132	17	1	4	12.2	16.3	2.3	15.2	37.7	16.6	1.9	13.0	33.5	35.0	35.5	15.3	19.4
133	17	2	1	15.3	11.4	7.8	9.9	26.3	11.7	9.2	4.8	35.6	35.9	34.0	9.9	17.2
134	17	2	2	3.3	2.8	3.1	2.8	23.7	8.3	4.9	1.5	18.4	19.0	21.9	6.8	7.2
135	17	2	3	25.2	8.0	4.5	9.6	23.7	1.8	6.5	6.6	35.2	32.7	26.5	5.6	2.6
136	17	2	4	73.9	16.4	5.9	17.7	30.9	3.0	10.1	11.9	24.3	25.3	24.0	18.2	11.3
137	18	1	1	15.0	7.2	3.0	4.2	27.5	9.0	2.3	1.8	19.3	23.6	20.1	5.1	7.9
138	18	1	2	19.5	5.7	6.4	3.1	25.8	7.4	20.7	1.9	22.6	18.5	23.7	2.2	6.6
139	18	1	3	37.9	6.3	6.7	4.2	35.0	7.9	1.2	10.0	30.9	32.8	30.3	10.5	6.7
140	18	1	4	13.0	4.0	4.1	3.8	23.6	9.5	3.5	5.9	20.7	24.6	20.0	8.3	4.6
141	18	2	1	22.5	2.4	4.9	4.4	16.3	7.5	3.4	12.8	22.4	16.7	22.8	6.6	2.6
142	18	2	2	7.3	4.3	2.9	0.9	29.6	27.9	8.4	22.8	22.4	24.9	23.2	10.8	5.4
143	18	2	3	29.9	6.5	4.6	5.1	22.8	4.1	2.8	9.1	27.8	23.6	26.6	3.0	6.6
144	18	2	4	56.1	10.3	10.1	7.4	32.8	1.4	7.9	8.1	24.6	21.4	24.7	14.5	10.7
145	19	1	1	6.7	8.5	6.3	3.5	21.1	2.2	5.6	3.0	28.0	24.3	24.8	4.9	7.6
146	19	1	2	12.0	0.4	2.8	6.3	24.3	6.9	5.4	2.0	16.5	24.7	17.0	7.0	3.4
147	19	1	3	25.3	3.0	0.6	6.9	35.4	7.4	3.5	13.0	30.0	38.2	28.8	11.7	6.1
148	19	1	4	10.2	11.4	6.9	5.4	28.7	16.3	3.6	10.2	25.1	23.7	28.2	17.9	12.5
149	19	2	1	15.8	11.5	12.0	3.9	30.2	14.6	11.6	4.4	33.6	39.2	31.4	16.4	15.3
150	19	2	2	9.9	7.2	4.4	4.6	22.1	9.1	7.9	6.1	23.5	19.2	25.3	5.7	10.8
151	19	2	3	26.8	3.9	0.6	8.0	12.1	14.4	5.2	10.2	26.5	15.0	19.5	10.6	4.6
152	19	2	4	82.4	25.0	16.0	21.1	34.3	2.2	12.4	9.7	18.8	20.6	21.4	17.4	17.2
153	20	1	1	10.0	4.9	4.7	0.4	15.1	3.8	0.3	5.4	24.2	18.0	23.4	1.4	4.0
154	20	1	2	12.1	1.8	0.7	5.2	15.7	5.3	1.7	5.0	14.0	16.8	13.5	6.7	2.1
155	20	1	3	26.7	2.5	0.8	6.0	30.6	7.5	1.0	8.2	31.8	32.2	31.7	4.9	2.8
156	20	1	4	14.3	10.2	4.6	6.9	17.2	11.6	4.5	9.6	20.6	14.7	21.2	11.0	9.9
157	20	2	1	14.3	7.7	8.2	3.1	30.0	4.0	0.5	3.4	35.2	43.6	34.1	16.7	6.6
158	20	2	2	7.7	3.1	0.5	1.7	21.2	6.0	1.8	6.4	20.8	20.2	20.7	7.4	3.3
159	20	2	3	27.8	4.6	1.9	3.1	23.7	10.1	3.2	9.9	27.5	29.4	26.8	11.0	3.6
160	20	2	4	56.5	8.7	10.2	6.5	38.1	8.2	11.1	9.9	25.8	33.0	26.8	3.8	7.6
161	21	1	1	10.5	4.2	3.1	0.6	23.2	4.3	5.0	1.8	22.9	22.1	23.1	3.9	4.3
162	21	1	2	15.0	3.2	4.7	0.7	24.4	7.8	5.0	3.5	19.4	24.1	19.8	7.3	3.7
163	21	1	3	34.7	5.7	5.9	1.9	31.4	4.9	6.4	2.7	27.8	31.0	27.3	3.6	6.2
164	21	1	4	12.5	3.3	1.5	2.1	20.2	4.5	2.5	0.8	20.2	20.8	20.3	2.7	3.3
165	21	2	1	18.4	2.9	4.8	1.2	35.4	8.8	7.8	6.4	29.4	30.0	29.2	4.1	2.6
166	21	2	2	7.8	5.3	3.6	2.5	21.4	5.5	6.1	6.6	23.3	19.9	23.8	1.6	6.1
167	21	2	3	27.1	4.8	2.5	2.6	22.9	7.0	2.4	5.4	24.2	20.3	23.1	5.8	4.4
168	21	2	4	39.8	0.6	2.9	3.1	33.1	1.2	3.5	4.3	31.3	31.7	31.7	2.1	1.2

Table 7.2

Venezuela - 1990 Census Proportions and Subgroup Absolute Relative Residuals Averages (ARB) for Thirteen Different Models

Categories				MODELS												
SG	STT	Sex	Age	(a)	(b)	(c)	(d)	SPREE				Unsaturated SPREE				
								(a)	(b)	(c)	(d)	(a)-(b)	(a)-(c)	(a)-(d)	(b)-(c)	(b)-(d)
169	22	1	1	22.3	6.6	6.4	9.5	17.6	10.0	12.7	14.8	12.8	19.6	20.6	11.5	1.0
170	22	1	2	9.7	16.7	19.9	7.7	20.0	20.1	16.7	12.8	18.9	20.1	13.1	23.4	15.5
171	22	1	3	16.2	24.2	21.7	17.7	35.1	24.5	16.6	17.1	40.7	38.6	41.2	32.7	21.2
172	22	1	4	11.4	14.9	12.2	0.9	44.2	27.3	8.6	19.7	37.4	44.7	32.0	29.0	8.0
173	22	2	1	29.0	10.6	10.5	7.1	13.4	32.7	20.9	21.0	12.7	2.5	23.8	20.0	2.5
174	22	2	2	14.1	19.0	14.4	4.2	18.0	9.5	12.2	5.2	27.7	22.2	21.0	15.9	11.6
175	22	2	3	32.4	17.0	16.1	10.1	75.5	33.9	19.2	39.6	44.9	62.4	54.0	27.9	21.4
176	22	2	4	34.1	22.4	24.4	8.1	29.5	2.9	14.1	17.3	46.0	40.8	40.2	14.8	13.5
177	23	1	1	15.6	10.7	3.0	15.4	25.5	6.3	5.4	18.5	20.4	24.3	28.0	5.7	9.1
178	23	1	2	15.4	23.5	23.1	15.2	15.9	21.4	16.6	16.5	25.1	19.9	20.4	32.5	19.7
179	23	1	3	25.0	44.6	26.8	39.2	55.5	47.1	21.4	39.1	48.4	55.4	47.8	51.3	42.0
180	23	1	4	9.6	16.7	8.9	23.1	41.2	19.6	8.1	8.0	41.0	41.8	35.6	21.2	18.3
181	23	2	1	26.6	8.4	5.5	15.2	1.5	21.6	9.8	10.0	16.7	7.4	25.5	14.8	9.2
182	23	2	2	14.1	19.6	22.6	3.5	28.5	17.7	19.5	21.8	29.1	28.6	22.2	23.5	11.7
183	23	2	3	48.6	28.5	26.4	21.2	63.0	29.4	23.2	34.9	58.8	71.4	59.7	36.7	26.1
184	23	2	4	71.7	26.9	14.4	25.7	38.8	10.6	19.3	12.2	38.5	31.5	33.1	20.9	27.5

REFERENCES

Agresti Alan (1990), "Categorical Data Analysis", New York: John Wiley and Sons.

Aitchiso J. and Aitken C.G.G. (1976), "Multivariate Binary Discrimination by the Kernel Method", Biometrics, Vol.63, 413-420.

Battese G.E. and Harter M. and Fuller W.A. (1988), "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data", Journal of the American Statistical Association, Vol.83, 28-36.

Berkson J. (1972), "Minimum Discrimination Information, the "No Interaction" Problem, and the Logistic Function", Biometrics, Vol.28, 443-468.

Bhapkar V.P. (1966), "A Note on the Equivalence of Two Test Criteria for Hypotheses in Categorical Data", Journal of the American Statistical Association, Vol.61, 228-235.

Binder D.A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys", International Statistical Review, Vol.51, 279-292.

Birch M.W. (1963), "A New Proof of The Pearson-Fisher Theorem", Annals of Mathematical Statistics, Vol.35, 817-824.

Bogue D.J. (1950), "A Technique for Making Extensive Population Estimates", Journal of the American Statistical Association, Vol.45, 149-163.

Booth J.G. and Hobert J.P. (1998), "Standard Errors of Predictors in Generalized Linear Mixed Models", Journal of the American Statistical Association, Vol.93, 362-372.

Brackstone G.J. (1987), "Small Area Data: Policy Issues and Technical Challenges", In: Small Area Statistics (Platek R., Rao J.N.K., Sarndal C.E. and Singh M.P. Eds.), New York: John Wiley, 1-22.

Cassel C.M. and Kristiansson K.E. and Råbäck G. and Wahlström S (1987), "Using Model-Based Estimation to Improve the Estimate of Unemployment on a Regional Level in the Swedish Labour Force Survey", In: Small Area Statistics (Platek R., Rao J.N.K., Sarndal C.E. and Singh M.P. Eds.), New York: John Wiley, 141-159.

Centro Interamericano para la Enseñanza de Estadísticos (CIENES) (1995), "Encuestas sobre Fuerza de Trabajo: Metodología Comparada", Santiago: CIENES Press.

Chambers, R.L. (1997), "Small-Area Estimation: A Survey Samplers' Perspective", In: Population Counts in Small Areas: Implications for Studies of Environment and Health (Arnold, Elliot, Wakefield and Quinn Eds), Studies on Medical and Population Subjects No.62, Governmental Statistical Service, UK, London: The Stationary Office.

Citro F.C. and Cohen M.L. and Kalton G. and West K.K. (Eds.) (1997), "Small-Area Estimates of School-Age Children in Poverty, Interim Report I: Evaluation of 1993 County Estimates for Title I Allocations", National Research Council, National Academy Press.

Citro F.C. and Cohen M.L. and Kalton G. and West K.K. (Eds.) (1998), "Small-Area Estimates of School-Age Children in Poverty, Interim Report II: Evaluation of 1993 County Estimates for Title I Allocations", National Research Council, National Academy Press.

Citro F.C. and Kalton G. (2000), "Small-Area Income and Poverty estimates: Priorities for 2000 and Beyond", ASA Proceedings of the Section on Survey Research Methods, 69-74

Cohen M.L. (2000), "Evaluation of the Census Bureau's Small-Area Poverty Estimates", ASA Proceedings of the Section on Survey Research Methods, 62-68.

Cohen M.P. (1999), "Small Area Estimation for the Distribution of the Parameters with Covariates", ASA Proceed. of the Sec. on Survey Research Methods, 655-659.

Cressie N. (1992), "REML Estimation in Empirical Bayes Smoothing of Census Undercount", Survey Methodology, Vol.18, 75-94.

Cronkite F.R. (1987), "Use of Regression Techniques for Developing State and Area Employment and Unemployment Estimates", In: Small Area Statistics (Platek R., Rao J.N.K., Sarndal C.E. and Singh M.P. Eds.), New York: John Wiley, 160-174.

Datta G.S. and Day B. and Basawa I. (1999), "Empirical Best Linear Unbiased and Empirical Bayes Prediction in Multivariate Small Area Estimation", Journal of Statistical Planning and Inference, Vol.75, 269-279.

Datta G.S. and Ghosh M. and Nangia N. and Natarajan K. (1996), "Estimation of Median Income of Four-Person Families: A Bayesian Approach.", In: Bayesian Analysis in Statistics and Econometrics (Berry, D.A., Chaloner, K.M. and Geweke, J.K. Eds.) , New York: John Wiley and Sons, 129-140..

Deming W.E. and Stephan F.F. (1940), "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known", Annals of Mathematical Statistics, Vol.11, 427-444.

Dobson A.J. (1990), "An Introduction to Generalized Linear Models", Chapman & Hall/CRC Press, 2nd. Ed.

Draper N. and Smith H. (1998), "Applied Regression Analysis", New York: John Wiley and Sons, 3rd Ed. edition.

Drew D. and Singh M.P. and Choudhry G.H. (1982), "Evaluation of Small Area Techniques for the Canadian Labour Force Survey", Survey Methodology, Vol.8, 17-47.

Efron B. and Morris C. (1972), "Limiting the Risk of Bayes and Empirical Bayes Estimates - Part II: The Empirical Bayes Case", Journal of the American Statistical Association, Vol.67, 130-139.

Ericksen E.P. (December 1974), "A Regression Method for Estimating Population Changes of Local Areas", Journal of the American Statistical Association, Vol.69, 867-874.

Falorsi P.P and Falorsi S. and Russo A. (1994), "Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey", Survey Methodology, Vol.20, 171-176.

Falorsi P.D. and Russo A. (1999), "A Conditional Analysis of Some Small Area Estimators in Two Stage Sampling", Journal of Official Statistics, Vol.15, 537-550.

Farrel P.J. and McGbbon B. and Tomberlin T.J. (1997b), "Empirical Bayes Estimates of Small Area Proportions in Multistage Designs", Vol.7, 1065-1083.

Fay R.E. (1985), "A Jackknife Chi-squared Test for Complex Samples", Journal of the American Statistical Association, Vol.80, 148-157.

Fay R.E. and Herriot R. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data", Journal of the American Statistical Association, Vol.74, 269-277.

Feeney G.A. (1987), "The Estimation of the Number of Unemployed at the Small Area Level", In: Small Area Statistics (Platek R., Rao J.N.K., Sarndal C.E. and Singh M.P. Eds.), New York: John Wiley, 198-218.

Freedman D.A. and Navidi W.C. (1992), "Should We Have Adjusted the U.S. Census of 1980? (with discussion)", Survey Methodology, Vol.18, 3-74.

Ghosh M. and Natarajan K. and Stroud T.W.F. and Carlin B.P. (1998), "Generalized Linear Models for Small Area Estimation", Journal of the American Statistical Association, Vol.93, 273-282.

Ghosh M. and Rao J.N.K. (1994), "Small Area Estimation: An Appraisal", Statistical Science, Vol.9, 55-93.

Gokhale D.V. and Kullback S. (1978), "The Information in Contingency Tables", New York: Marcel Dekker.

Gonzalez M.E. (1973), "Use an Evaluation of Synthetic Estimates", Journal of the American Statistical Association, 33-36.

Gonzalez, M. E. and Hoza, C. (March 1978), "Small-Area Estimation with Application to Unemployment and Housing Estimates", Journal of the American Statistical Association, Vol.73, 7-15.

Goodman L. A. (1965), "On Simultaneous Confidence Intervals for Multinomial Proportions", Technometrics, Vol.7, 247-254.

Gozalez J.F. and Placek P.J. and Scott C. (1996), "Synthetic Estimation in Followback Surveys at the National Center for Health Statistics.", In: Indirect Estimators in U.S. Federal Programs (Schaible, W.L. Ed.), New York: Springer-Verlag, 16-27.

Griffiths R. (1996), "Current Population Survey Small Area Estimation for Congressional Districts", ASA Proceedings of the Section on Survey Research Methods, 314-319.

Grizzle J.E. and Starmer C.F. and Koch G.G. (1969), "Analysis of Categorical Data by Linear Models", Biometrics, Vol.25, 489-504.

Haberman S.J. (1973), "Log-linear Models for Frequency Data: Sufficient Statistics and Likelihood Equations", The Annals of Statistics, Vol.1, 617-632.

Haberman S.J. (1974), "The Analysis of Frequency Data", Chicago: University of Chicago Press.

Hansen M. and Hurwitz W.N. and Madow W.G. (1953), "Sampling Survey Methods and Theory", New York: John Wiley and Sons, Vol.1.

Harte R. (2000), "County Estimates of Employment using CES and ES202 Data", ASA Proceedings of the Section on Government Statistics and Social Statistics, 166-171.

Harville D.A. (1997), "Matrix Algebra From a Statistician's Perspective", New York: Springer-Verlag.

Haskey J. (1991), "The Ethnic Minority Population Resident in Private Households - Estimates by Country and Metropolitan Districts of England and Wales", Population Trends, Vol.63, 22-35.

Heady Patrick and Clarke P. and Brown G. and D'Amore A. and Mitchell B. (1999), "Small Area Estimates Derived from Surveys", Paper created for IASS Conference on Small Area Estimation.

Hidiroglou M.A. and Sarndal C.E. (1985), "An Empirical Study of some Regression Estimators for Small Domains", Survey Methodology, Vol.11, 65-77.

Hoaglin D.C. and Welch R.E. (1978), "The Hat Matrix in Regression and ANOVA", The American Statistician, Vol.32, 17-22.

Holt D. and Holmes D.J. (1994), "Small Domain Estimation for Unequal Probability Survey Designs", Survey Methodology, Vol.20, 23-31.

Holt D. and Smith T.M.F. and Tomberlin T.J. (1979), "A Model Based Approach to Estimating for Small Subgroups of a Population", Journal of the American Statistical Association, Vol.74, 405-410.

James W. and Stein C. (1961), "Estimation with Quadratic Loss", In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley, 361-379.

Judkins D.R. and Liu J. (2000), "Correcting the Bias in the Range of a Statistics Across Small Areas", Journal of Official Statistics, Vol.16, 1-13.

Kalton G. (1994), "Issues and Strategies for Small Area Data: Comment", Survey Methodology, Vol.20, 18-20.

Kalton G. and Kordos J. and Platek R. (Eds) (1993), "Small Area Statistics and Survey Designs", Vol. I: Invited Papers; Vol. II: Contributed Papers and Panel Discussion. Central Statistical Office, Warsaw.

Kish L. (1965), "Survey Sampling", New York: John Wiley and Sons, 1st Ed.

Knaub J.R. Jr. (1999), "Using Prediction-Oriented Software for Model-Based and Small Area estimation", ASA Proceedings of the Section on Survey Research Methods, 660-665.

Kullback S. (1959), "Information Theory and Statistics", New York: John Wiley and Sons.

Larsen M.D. (2000), "Estimation of the Small Area Proportions Using Survey Weights", ASA Proceedings of the Section on Government Statistics and Social Statistics, 148-153.

Lundström S. (1987), "An Evaluation of Small Area Estimation Methods: The case of Estimating the Number of Nonmarried Cohabiting Persons in Swedish Municipalities", In: Small Area Statistics (Platek R., Rao J.N.K., Sarndal C.E. and Singh M.P. Eds.), New York: John Wiley, 239-256.

Madow L.H. (1956), "U.S. Television Households by Region State and County - March 1956", Advertisind Research Foundation, New York.

Malec D. and Sedransk J. and Moriarity C.L. and Leclere F. (1997), "Small Area Inference for Binary Variables in the National Health Interview Survey", Journal of the American Statistical Association, Vol.92, 815-826.

Marker D.A. (1999), "Organization of Small Area Estimators using a Generalized Linear Regression Framework", Journal of Official Statistics, Vol.15, 1-24.

McCullagh P. and Nelder J.A. (1983), "Generalized Linear Models", London: Chapman and Hall, 1st. Ed.

Molina E.A. and Skinner C.J. (1992), "Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes", Computational Statistics and Data Analysis, Vol.13, 395-405.

Moura F.A.S. and Holt D. (1999), "Small Area Estimation Using Multilevel Models", Survey Methodology, Vol.25, 73-83.

Nelder J.A. (1974), "Log-Linear Models for Contingency Tables: A Generalization of Classical Least Squares", Applied Statistics, Vol.23, 323-329.

Neyman J. (1949), "Contributions to the Theory of the Chi-squared Test", In: Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability, Edited by Neyman, J., University of California Press, Berkeley, 361-379.

Nichol S. (1977), "A Regression Approach to Small Area Estimation", Unpublished Manuscript, Australian Bureau of Statistics, Canberra, Australia.

Oficina Central de Estadística e Informática (OCEI) (1987), "Encuesta de Hogares por Muestreo: Documento Técnico", Caracas: Imprenta OCEI.

Oficina Central de Estadística e Informática (OCEI) (1997), "30 Anos de la Encuesta de Hogares por Muestreo", Caracas: Imprenta OCEI.

Olsen C.H. and Schirm A.L. and Zaslavsky A.M. (2000), "An Evaluation of State Estimates Produced by Model-Based Reweighting of a National Database", ASA Proceedings of the Section on Government Statistics and Social Statistics, 154-159.

Pfeffermann D. (1999), "Small Area Estimation - Big Developments", Paper Presented at the Conference on Analysis of Survey Data, Southampton, England.

Pfeffermann D. and Burk L. (1990), "Robust Small Area Estimation Combining Time Series and Cross-Sectional Data", Survey Methodology, Vol.16, 217-237.

Pfeffermann D. and Feder M. and Signorelli D. (1998), "Estimation of Autocorrelations of Survey Errors with Application to Trend Estimation in Small Areas", Journal of Business and Economic Statistics, Vol.16, 339-348.

Platek R. and Rao J.N.K. and Sarndal C.E. and Singh M.P. (ed.) (1987), "Small Area Statistics", International Symposium on Small Area Statistics - Mayo 1985 - Ottawa.

Platek R. and Singh M.P. (1986), "Small Area Statistics: Contributed Paper", Laboratory For Research in Statistics and Probability, Carleton University.

Pregibon D. (1981), "Logistic Regression Diagnostics", The Annals of Statistics, Vol.9, 705-724.

Purcell N.J. and Kish L. (1979), "Estimation for Small Domains", Biometrics, Vol.35, 365-384.

Purcell N.J. and Kish L. (1980), "Postcensal Estimates for Local Areas (or Domains)", International Statistical Review, Vol.48, 3-18.

Rao J.N.K. (1999), "Some Recent Advances in Model-Based Small Area Estimation", Survey Methodology, Vol.25, 175-186.

Rao J.N.K (2000), "Introduction to Small Area Estimation", Monograph Published by EUSTAT, Spain, March 2000.

Rao J.N.K. and Kumar S. and Roberts G. (1989), "Analysis of Sample Survey Data Involving Categorical Response Variables: Methods and Software", Survey Methodology, Vol.15, 161-186.

Rao J.N.K. and Scott A.J. (1981), "The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Test of Goodness of Fit and Independence in Two-Way Tables", Journal of the American Statistical Association, Vol.76, 221-230.

Roberts G. and Rao J.N.K. and Kumar S. (1987), "Logistic Regression Analysis of Sample Survey Data", Biometrika, Vol.74, 1-12.

Royall R.M. (1986), "Models Robust Confidence Intervals Using Maximum Likelihood Estimators", International Statistical Review, Vol.54, 221-226.

Sarndal C. and Swensson B. and Wretman J.H. (1989), "The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total", Biometrika, Vol.76, 527-537.

Sarndal C. and Swensson B. and Wretman J.H. (1992), "Model Assisted Survey Sampling", New York: Springer-Verlag.

Schaible W.L. (1992), "Use of Small Area Statistics in U.S. Federal Programs", Vol.1, 95-114.

Siegel J.S. and Shryock H.S. and Greenberg B.Jr. (1954), "Accuracy of Postcensal Estimates of Population for States and Cities", American Sociological Review, Vol.19, 440-446.

Simon G. (1973), "Additivity of Information in Exponential Family Probability Laws", Journal of the American Statistical Association, Vol.68, 478-482.

Simonoff J. (1983), "A Penalty Function Approach to Smoothing Large Sparse Contingency Tables", The Annals of Statistics, Vol.11, 208-218.

Singh A.C. and Mantel H.J. and Thomas B.W. (1994), "Time Series EBLUP's for Small Areas Using Survey Data", Survey Methodology, Vol.20, 33-43.

Singh M.P. and Gambino J. and Mantel H.J. (1994), "Issues and Strategies for Small Area Data", Survey Methodology, Vol.20, 3-22.

Skinner C. (1991), "The Use of Synthetic Estimation Techniques to produce Small Area Estimates", NM18 - New Methodology Series, Office of Population Census & Surveys.

Skinner C.J. and Holt D. and Smith T.M.F (Eds.) (1989), "Analysis of Complex Surveys", Chincester: John Wiley and Sons.

Thomas R. D. and Rao J.N.K. (1987), "Small-Sample Comparisons of Level and Power for Simple Goodness-of-fit Statistics Under Cluster Sampling", Journal of the American Statistical Association, Vol.82, 630-636.

Thompson J.R. (1968), "Some Shrinkage Techniques for Estimating the Mean", Journal of the American Statistical Association, Vol.68, 113-122.

Tiller R.B. (1992), "Times Series Modelling of Sample Survey Data from the U.S. CPS", Journal of Official Statistics, Vol.8, 149-166.

Titterington D.M. and Bowman A.W. (1985), "A Comparative Study of Smoothing Procedure for Ordered Categorical Data", Journal of Statistics and Computing Simulation, Vol.21, 291-312.

U.S. Government Printing Office, Washington, D.C. (1968), "Synthetic State Estimates of Disability.", P.H.S. Publication 1759..

United Nations Statistical Office (1950), "The Preparation of Sampling Survey Reports", New York: U.N. Series C, 1 edition.

Wang J. and Fuller W.A. and Opsomer J. (1999), "Small Area Estimation in the National Resources Inventory", ASA Proceed. of the Sec. on Survey Research Methods, 650-654.

Wang S. and Chambers R.L.C. and Douglas A. and Caplen D. (1999), "Small Area Estimation of Unemployment in Great Britain", Paper created for IASS Conference on Small Area Estimation.

Wolter K.M. (1985), "Introduction to Variance Estimation", New York: Springer-Verlag.

Woodruff R.S. (1966), "Use of a Regression Technique to Produce Area Breakdowns of the Monthly National Estimates of Retail Trade", Journal of the American Statistical Association, Vol.61, 496-504.

You Y. and Rao J.N.K. (1999), "Hierarchical Bayes Estimation of Small Area Means Using Multilevel Models", Proceedings of IASS Satellite Conference on Small Area Estimation, 171-185, Riga, Latvia.

You Y. and Rao J.N.K. and Gambino J. (2000), "Hierarchical Bayes Estimation of Unemployment Rates for Sub-Provinctional Regions Using Cross-Sectional and Time Series Data", ASA Proceedings of the Section on Government Statistics and Social Statistics, 160-165.