Integrated Methodology for Multiple Systems Estimation and Record Linkage Using a Missing Data Formulation

Stepthen E. Fienberg *and Daniel Manrique-Vallier†‡

September 19, 2008

Abstract

There are now three essentially separate literatures on the topics of multiple systems estimation, record linkage, and missing data. But in practice the three are intimately intertwined. For example, record linkage involving multiple data sources for human populations is often carried out with expressed goal of developing a merged database for multiple system estimation (MSE). Similarly, one way to view both the record linkage and MSE problems is as ones involving the the estimation of missing data. This presentation highlights the technical nature of these interrelationships and provides preliminary effort at their integration.

Keywords: Capture-recapture, Heterogeneity, Data fusion, EM algorithm, Fellegi-Sunter linkage, Missing data

1 Introduction

What do the following populations have in common?

- Women with breast cancer,
- Victims of killing in Kosovo,
- People in the U.S.,
- Size of WWW.

*e-mail: fienberg@stat.cmu.edu. Department of Statistics and Machine Learning Department. Carnegie Mellon University. Pittsburgh, PA 15213, USA

†e-mail: dmanriqu@stat.cmu.edu Department of Statistics. Carnegie Mellon University. Pittsburgh, PA 15213, USA

‡This research was supported in part by the National Institutes of Health under Grant No. R01 AG023141-01 to Carnegie Mellon University and in part through a project funded by the Asia Foundation in Sri Lanka. We thank Jana Asher for discussions that led to some of the formulation of the interconnections described in this paper.

At any given point in time they each represent closed populations whose total size is unknown. In this paper, we reconsider aspects of the estimation of population totals for these examples using multiple systems estimation (MSE), the version of multiple recapture estimation usually associated with human populations. In the process we raise issues such as the role of covariates, missing data, and the integration of multiple lists. For example, to implement MSE we need to be able to match lists—enter record linkage. Further, covariates for individual units on multiple lists often contain missing elements. We can use ideas from literature on missing data to integrate methodologies. The list of authors discussing one or even two of these topics is substantial, especially in the realm of MSE using log-linear models with and without heterogeneity. Herzog et al. (2007) introduce all three of these topics but without a discussion of their methodological links and the possibility of integration!

In the following three sections we review key elements of (i) MSE, (ii) missing data methodology, especially as it can be applied to MSE, and (iii) record linkage. Then we conclude with a discussion of the possibility of a "grand synthesis" of these ideas in the form of an integrated methodology and point to some difficult open problems. Our discussion is most relevant to population estimation from human populations but some elements are also applicable in the analysis of animal populations as well.

2 Multiple Systems Estimation (MSE)

Beginning with Sanathanan (1972b, 1973); Fienberg (1972); Bishop et al. (1975), the literature on multiple systems estimation and multiple recapture estimation began to focus in depth on departures from the core assumptions of the independence of lists and the homogeneity of population units as represented on lists. Integrated discussion of these appeared largely in the 1990s in papers by Darroch et al. (1993); Agresti (1994); International Working Group for Disease Monitoring and Forecasting (1995a,b); Norris and Pollock (1996); Fienberg et al. (1999). See also the discussion of various types of mixture models and other approaches in Chao et al. (2001); Pledger et al. (2003); Manrique-Vallier and Fienberg (2008). Estimation in most of the literature typically involved some version of maximum likelihood or Bayesian methods. Following the ideas of conditional estimation in Sanathanan (1972a), most authors adopted a 2-step approach, first fitting a model with dependence and/or heterogeneity to the incomplete 2^k table and then projecting that model once estimated, to the missing cell corresponding to not being included on any of the lists.

Consider a closed population of N individuals or units where N is unknown. For this population we have k lists, whose information we would like to merge to estimate N. We represent this integrated information in the form of a 2^k contingency table with entries $\{x_{i_1 i_2 \ldots i_k}\}$, for $i_j = 1, 2$ and $j = 1, 2, \ldots, k$, and we denote the expected counts correspondingly by $\{m_{i_1 i_2 \ldots i_k}\}$. Since we do not get to observe those who are on none of the lists, $x_{22\ldots2} = 0$ and our goal is to estimate $m_{22\ldots2}$ by fitting a model to the observed incomplete table.

2.1 Log-linear Models and Maximum Likelihood Estimation

We illustrate a standard approach using log-linear models to estimate the $\{m_{i_1 i_2 \ldots i_k}\}$. Let $n = \sum x_{i_1 i_2 \ldots i_k}$ be the total number of observed individuals, and let $\hat{m}_{i_1 i_2 \ldots i_k}$ be the maximum likelihood estimate of $m_{i_1 i_2 \ldots i_k}$ under a log-linear model fit to the incomplete 2^k table of counts. Then we project the estimated model to the unobserved cell by maximizing

$$L(N|n, \{\hat{m}_{i_1 i_2 \ldots i_k}\}). \tag{1}$$

This yields a conditional maximum likelihood estimate of N (see Sanathanan (1972a)). Under the now familiar multinomial sampling model for the counts in the incomplete 2^k table, Fienberg (1972); Bishop et al. (1975) explain that this is equivalent to the following estimate for the unobservered cell:

$$\hat{m}_{22\ldots2} = \frac{M_{\text{odd}}}{M_{\text{even}}}, \tag{2}$$

$$\hat{N} = n + \hat{m}_{22\ldots2}. \tag{3}$$

where M_{odd} and M_{even} are the products of estimated values in the cells whose subscripts sum to odd and even values, respectively.

Example 1: Killings in Kosovo Ball and Asher (2002); Herzog et al. (2007) describe log-linear model analyses of data on deaths in Kosovo gathered from four different but dependent lists. We reproduce the table here as Table 1. Here $n = 4400$.

Using a parsimonious log-linear representation, Ball and Asher (2002) estimated a total number of killings of $\hat{N} = 10356$. Table 3 shows some estimates for N obtained under the log-linear models for independence, no third order interaction and a more parsimonious model selected by a stepwise search based on BIC, with

				ABA			
				Yes		No	
				EXH		List-3	
				Yes	No	Yes	No
HRW	Yes	OSCE	Yes	27	32	42	123
			No	18	31	106	306
	No	OSCE	Yes	181	217	228	936
			No	177	845	1131	-

Table 1: Number of Individual Victims of Killing by Documentation Status. Source: (Ball and Asher, 2002)

				ABA			
				Yes		No	
				EXH		EXH	
				Yes	No	Yes	No
HRW	Yes	OSCE	Yes	35	27	46	116
			No	15	43	97	306
	No	OSCE	Yes	173	222	224	943
			No	180	833	1140	–

Table 2: Estimated Counts of Numbers of Victims of Killing by Documentation Status Under BIC Selected Log-linear Model. (Rounded to the nearest integer.)

Model	df	Deviance	$\hat{N}$	95%-CI
Independence	10	245.9	7395	[7149, 7658]
Saturated	0	–	16942	[9320, 35986]
Loglinear-BIC ([12][23][134])	4	9.323	10357	[9012, 12138]
Bayesian Rasch	—	—	16065	[12266, 19686]
Bayesian GoM ($K = 2$)	—	—	9825	[8906, 11760]
Bayesian GoM ($K = 3$)	—	—	11239	[9954, 16404]

Table 3: Kosovo data

their respective confidence intervals computed by profile likelihood (Cormack, 1992). As an example we show the expected values under the BIC-selected log-linear model in Table 2. For this particular case, we get the estimate $\hat{m}_{22\ldots2} = 5997$, so that the conditional estimate is $\hat{N} = 4400 + 5997 = 10537$. For comparison, we also show estimates under some of the models to handle individual heterogeneity (GoM and Rasch) that we will discuss later.

2.2 Heterogeneity and Individual Level Mixtures: the Rasch and GoM models

For k lists and n observed individuals, let Y_{ij} be independent random variables taking the value 1 if the ith individual in the jth list, and 0 otherwise, for $i = 1, 2, \ldots .n$ and $j = 1, 2, \ldots, k$. Further let $p_{ij} = Pr(Y_{ij} = 1)$.

The Rasch model sets $\log p^{ij}/(1-p^{ij}) = \theta_i + \beta_j$. Individual heterogeneity represented by the $\{\theta_i\}$, and the independence of lists by the $\{\beta_j\}$. Thus if the $\theta_i = \theta$ for all i, then we have homogeneity and independent lists, which in effect reduces to the baseline model in the log-linear multinomial setup from the preceding subsection. The key to the Rasch model is that the latent structure allows for estimation of unobserved cell directly without the use of eqs. 2 and 3. Bayesian estimation makes approach work well and, as Fienberg et al. (1999) remark, they yield estimate for missing cell directly. We can let

$$\begin{aligned} \theta^i &\sim N(0, \sigma^2 I) \text{ with } \sigma^2 \sim \Gamma^{-1}(1,1), \\ \beta^j &\sim N(0, 10 \cdot I_k), \\ f_{\mathbf{N}}(N) &\propto \frac{1}{N} I_{\{n<N<N_{max}\}} \text{ and } N_{max} >> n. \end{aligned}$$

Then the information in observed data for estimating N is concentrated in the conditional posterior probability of unobserved cell: $P(\text{“unobserved cell”}|\{\beta_j\}, \sigma)$. The normality for the θ's and the β's is not necessarily important although there relative diffuseness is. The approach is also relatively robust with respect to the choice of the value of N_{max}, but what it important is to truncate the tail of the distribution for $f_{\mathbf{N}}(N)$. for further details, see Fienberg et al. (1999).

Using a similar setting, Manrique-Vallier and Fienberg (2008) adapted the Grade of Membership Model (GoM) (Woodbury et al., 1978; Erosheva, 2002; Erosheva and Fienberg, 2005) to the multiple recapture problem. Similar to the Rasch model, the GoM model specifies an individual level latent structure to accommodate heterogeneity, but it does it using the concept of *partial membership* or *soft clustering*. A small number, K, of *extreme profiles* or *pure types* is specified and homogeneity of capture probabilities is assumed within each class. Heterogeneity is modeled letting each individual to “belong” to every class up to a certain degree g_{ik} ($g_{ik} > 0$, $\sum_k g_{ik} = 1$). Calling λ_{jk} the probability of capture in list j for a complete member of pure type k, the distribution of capture probabilities for a generic individual is modeled by a convex combination of the distributions for the complete members of every class, weighted by the membership vector:

$$P(Y_{ij} = y_{ij}|g_i) = \sum_k g_{ik} \lambda_{jk}^{y_{ij}} (1-\lambda_{jk})^{1-y_{ij}}$$

Manrique-Vallier and Fienberg (2008) complete the specification of the model in a hierarchical Bayesian framework using the priors

$$\begin{aligned} g_i &\sim \text{Dirichlet}\left(\alpha_0 \cdot (\xi_1, \xi_2, ..., \xi_k)\right) \\ &\quad \text{with } \alpha_0 \sim \Gamma(\alpha, \tau) \text{ and } \xi \sim \text{Dirichlet}(1, 1, ..., 1)_K \\ \lambda_{ij} &\sim \text{Uniform(0,1)} \\ f_{\mathbf{N}}(N) &\propto I_{\{N<N_{max}\}} \end{aligned}$$

We give estimates from the posterior distribution for both the Rasch and GoM models for the Kosovo data in Table 3 above and in the following example.

Example 2: Estimating the size of the WWW Lawrence and Giles (1998) studied the coverage and recency of six (then) major Web search engines by submitting 575 queries in December 1997. Dobra and Fienberg (2003) used the multiple recapture dataset generated to estimate the size of the WWW, modeling the heterogeneity using the Rasch model (Fienberg et al., 1999; Dobra and Fienberg, 2003). In Table 4 we reproduce the multiple recapture contingency table for Query 140 as reported by Dobra and Fienberg (2003), where $n = 159$ unique pages were retrieved. Table 5 shows some estimates of the total number of documents matching Query 140, including log-linear models, the Bayesian Rasch model for heterogeneity (Fienberg et al., 1999) and the Grade of Membership model (Manrique-Vallier and Fienberg, 2008).

3 Missing Data and MSE

Missing data are common occurrences in most applications. Sometimes we are missing entire records or cases. Other times we are missing information for some variables for some cases. Basic MSE with multiple lists involves the former, and a number of authors over the years have represented MSE as a missing data problem. But we can also have missing values on covariates as in Baker (1990), with complex stratification, as in Zwane et al. (2004); Sutherland et al. (2007), and combinations of these, e.g., as in (Zwane and van der Heijden, 2007). The basic structure we and others exploit here involves the assumption of "missing at random."

						NL (6)							
						Yes				No			
						LY(5)				LY(5)			
						Yes		No		Yes		No	
						HB(4)		HB(4)		HB(4)		HB(4)	
						Yes	No	Yes	No	Yes	No	Yes	no
AV(3)	Yes	IS(2)	Yes	EX(1)	Yes	1	0	2	0	0	0	1	0
					No	2	0	3	2	0	0	0	2
			No	EX(1)	Yes	1	0	2	1	0	0	3	4
					No	1	3	0	8	2	0	3	19
	No	IS(2)	Yes	EX(1)	Yes	0	0	0	1	0	0	0	0
					No	0	0	1	1	0	0	5	4
			No	EX(1)	Yes	0	0	0	1	0	0	4	22
					No	0	0	7	17	2	3	31	–

Table 4: Web Query 140 data. EX = Excite, IS = Infoseek, AV= Altavista, HB = HotBot, LY = Lycos, NL = Northern Light.

Model	df	Deviance	$\hat{N}$	95%-CI
Independence	56	107.8	202	[185, 226]
Saturated	0	–	1339	–
Loglinear-BIC ([13][23][24][26][35][36][45])	49	43.62	325	[253, 451]
Bayesian Rasch	—	—	483	[335, 1787]
Bayesian GoM ($K = 2$)	—	—	247	[213, 338]
Bayesian GoM ($K = 3$)	—	—	253	[214, 336]
Bayesian GoM ($K = 4$)	—	—	246	[211, 335]
Bayesian GoM ($K = 5$)	—	—	238	[208, 323]

Table 5: WWW query 140

Following Rubin (1976); Little and Rubin (2002), we note that missing data are missing at random (MAR) if response probabilities depend on observed random variables but not the unobserved random variables. A common approach to dealing with data that are MAR is the Expectation-Maximization (EM) algorithm, see Dempster et al. (1977).

- **E-step:** Estimate missing data $m_{22\ldots2}$ via expectation conditional on current estimated value of $m_{i_1 i_2 \ldots i_k}$ for the cells in the incomplete 2^k table. This produces "complete" data table consisting of the observed counts plus the estimate for the missing cell.
- **M-step:** Maximize likelihood for the "complete" data under a log-linear model to get new estimate of $m_{22\ldots2}$.

We then alternate between the E and M steps until the process converges. The EM algorithm works especially well for exponential family problems, e.g., those involving normal and multinomial distributions, such as those described briefly in the preceding section. The Idea works not only for log-linear models but also for Rasch model and other mixture models. We emphasize EM not so much because of its computational efficiency, but more because it captures the essence of the treatment of missing data without an excess of notation and formal theory, c.f., Zwane et al. (2004).

Example 3: MSE with Covariates Baker (1990) illustrated both kinds of missingness and the use of the EM algorithm. His example included two lists for the detection of breast cancer, based on M(ammogram), and P(hysical exam), and two covariates, Age and Screen Number.

More generally we could consider a setup with k lists and C strata for covariates yielding a $2^k \times C$ table with missing counts in C cells, $(2, 2, , 2, j)$ for $j = 1, 2, \ldots, C$. As before we fill in the missing cells with estimates and the then fit log-linear model to the complete $2^k \times C$ table, alternating using EM.

For Baker's example with $k = 2$ lists this is especially easy to do and he illustrates the application of the EM approach to show the impact of Age as a covariate separately on each of the lists, with the lists being reasonably modeled by the conditional independence of lists given age.

4 MSE and Record Linkage

Record linkage is a missing data problem. Suppose we have two list, potentially overlapping but we do not know in advance which items on different lists go together. We may have considerable information about the units on each list and there can also be auxiliary information used for matching or record linkage.

Following Fellegi and Sunter (1969) we can setup the problem as follows. Given the two lists to match, say A and B, we consider the set of pairs $A \times B = \{(a, b) : a \in A \text{ and } b \in B\}$ belonging to any of two mutually exclusive classes: the unmatched pairs (U) and the matched pairs (M). To explicitly cast this problem as a missing data problem in the sense of Rubin (1976) we can consider "augmented" pairs of the form $A \times B \times \{M, U\} = \{(a, b, s) : a \in A, \text{ and } b \in B, \text{ and } s \in \{M, U\}\}$ where the last component represent the status of the pair (matched or unmatched). Now we can make the missing data connection by recognizing that the last component of each of these augmented pairs is missing in all cases.

4.1 Methods for Linking Lists

Of course the simplest way to match items on lists is using unique identifiers. These might be names or social security numbers in the U.S., but as anyone who has tried to do this knows, U.S. social security numbers are not in fact unique and names come in various forms of abbreviations and misspellings. Even then, we typically need some form of probabilistic approach to match. In Example 1, Killings in Kosovo, the four lists involved no unique identifiers and thus considerable effort was required to match the individuals on the lists—sometimes even names were absent. In Example 2, Estimating the size of the WWW, the investigators actually downloaded the html code for each webpage and thus they had the equivalent of unique identifiers for matching.

The most widely used method is due to Fellegi and Sunter (1969) and even today most modern methods are variants on the ideas they set forth, c.f. the discussion in Bilenko et al. (2003) and in Herzog et al. (2007). Fellegi-Sunter methods have implicit assumptions of overlap of lists used in the matching process; they rely on the accuracy of variables for blocking to enable comparison of records within lists. If two lists contain n_1 and n_2 records, then we make $n_1 \times n_2$ comparisons for possible matches. Clearly we need to do this smartly as the sizes of our lists grow!

4.2 Fellegi-Sunter Key Ideas

The idea behind most applications of Fellegi-Suntner methodology is to represent every pair of records using vector of features (variables) that describe similarity between individual record fields, e.g., Boolean (e.g., last-name matches), discrete (e.g., first-d-characters-of-name-agree), or continuous (e.g., string-edit-distance-between-first-names)). We use the matching procedure to place feature vectors for record pairs into three classes: matches (M), non-matches (U), and possible matches (PM), by performing record-pair classification by calculating the ratio $P(\gamma|M)/P(\gamma|U)$ for each candidate record pair, where γ is a feature vector for pair (functions of covariates used for matching), and $P(\gamma|M)$ and $P(\gamma|U)$ are probabilities of observing that feature vector for matched and non-matched pair, respectively. We choose two thresholds based on desired error levels—T_μ and T_λ—to optimally separate ratio values record pairs into three groups. For further details see Bilenko et al. (2003) and in Herzog et al. (2007).

When we have no training data available in form of duplicate and non-duplicate record pairs, we can do "unsupervised" matching by estimating conditional probabilities for feature values using observed frequencies. Because most record pairs are clearly non-matches, we need not consider all of them for matching. The way to manage this is to "block" databases, e.g., using geography (e.g., province) or some other variable in both databases) so that we only compare records in comparable blocks.

Example 4: U.S. Census Coverage Evaluation The U.S. Census Bureau uses a version of the Fellegi-Sunter approach to actually do matching between census records and follow-up sample records for same area. Here list 1 is corresponds to the census results for the entire nation and list 2 corresponds to the results for a sample of census blocks (city blocks) with total sample size of approximately 300,000. The sample includes all persons in every household in the sample of census blocks and thus they literally do matching within blocks. In 1990, About 75% of the sample data individuals were matched to there census records by a version of the Fellegi-Sunter approach. The remaining 25% were matched by hand! For details of the 1990 U.S. census, a discussion of the matching process, and the use of dual systems estimation with the resulting data, see Anderson and Fienberg (2001). There are related papers by (Ding and Fienberg, 1994,?) on approaches to dual systems and MSE in the presence of matching error.

Belin and Rubin (1995) suggest use of variations on logistic regression models for matching in the census context of Example 4. We are missing information on whether or not two records in different lists corre-

spond to the same individual , i.e., if they match and we estimate the probability of a match for each pair of records using a logistic function. In this context Belin and Rubin explicitly use an EM approach but estimate probabilities directly instead of indirectly as in Fellegi-Sunter. This approach still has problems and so they develop what refer to as "Mixture Model Calibration Method," using very fancy versions of EM. Many other authors adopt versions of EM for other variations on record linkage and these methods are actually in widespread use to create merged data files in many U.S. statistical agencies, not just in the context of census estimation.

4.3 Practical Issues for Record Linkage

What makes the Fellegi-Sunter approach work well? The techniques yield high quality record linkage locking (to reduce number of comparisons) and the knowledge that there should be a 1-1 match between elements in the two files or lists to be matched. But MSE using multiple lists is something we do when the lists deviate substantially from the 1-1 matching situation.

When does the method have problems? When we have list or files with little overlap, wthere are undetected duplications within files, and when we need to perform linkage with $k \geq 3$ lists. In the case of $k \geq 3$ lists we essentially need to match all lists in pairs, and then resolve discrepancies! Unfortunately, there is no unique way to do this! In human population problems and especially in census-style applications practitioners spend enormous amounts of time and effort organizing lists, de-duplicating (another application of record linkage techniques), and then examining pair of potential matches to determine match status. For example, see the discussion in Zaslavsky and Wolfgang (1993) on determining both match status and whether items (indivudual households) on lists are "within scope."

Example 5: Administrative Records Census For at least the last six decennial censuses, the U.S. Census Bureau has used a variety of methods for assessing the accuracy and coverage of the decennial census counts such as the approach described above in Example 4. In a separate effort, the Bureau has investigated the use of administrative records either as a substitute for or as a supplement to the traditional enumeration and other surveys. Specifically, the Bureau staff produced, on an annual basis, an administrative records "superlist" called the Statistical Administrative Records System (StARS), which was the end result of the careful merging of six or more administrative records sources:

- Internal Revenue Service Tax Year 1998 Individual Master File.
- Internal Revenue Service Tax Year 1998 Information Returns File.
- 1999 Medicare Enrollment Database.
- Indian Health Service Registration System.
- Selective Service System Registration File.
- Department of Housing and Urban Development 1999 Tenant Rental Assistance Certification.

As one can, see the overlap between some lists is substantial while at least some overlap hardly at all with the others. Thus Fellegi-Sunter variations were not sufficient to handle the matching. Asher and Fienberg (2001) contains further details.

Then as part of the evaluation program for Census 2000, the Bureau developed an administrative records experiment (AREX 2000) in which the feasibility of using StARS for an administrative records census was explored in five counties of Colorado and Maryland. Then an effort was undertaken to develop multiple systems estimation methods that would combine decennial census 2000 data, post-enumeration survey data, and the AREX 2000 administrative records file in order to create block-level population counts. Several issues arose during this research including:

- Matching across the three files creates 3-way cross-classifications for ACE sample blocks and 2-way cross-classifications for non-sample blocks. How does this fit within the framework outlined here?
- Inconsistencies between the reference times for the list. The AREX data were collected well prior to 2000, the Census 2000 data were collected around April 2000, and the post-enumeration data were collected later in 2000.

The first step in the multiple systems estimation project was to create a cross-classification table of population counts for each block using the three data sources described above. Because each data source references a different time period, it was possible for an individual to have different addresses in different lists, and therefore be in different blocks in different lists. To address this issue, (Asher and Fienberg, 2001) assumed multiple addresses are always the result of different reference times for lists (not multiple residency). Since the desired reference point was that of the decennial census (April 1, 2000), they placed all triple matches in their Census 2000 addresses. Double matches mirrored the same strategy; Census 2000 with post-enumeration and Census 2000 with AREX 2000 matches were put in their Census 2000 addresses,

but post-enumeration with AREX 2000 matches were put in their ACE addresses.

Asher and Fienberg (2002) developed several multiple systems estimation frameworks for analyzing the resulting three-waycross-classification tables, incorporating assumptions about missingness of post-enumeration sample blocks by including either stratification or covariates based on the post-enumeration sampling frame. Modeling within each stratum allowed them to ignore the missingness. Including covariates based on the characteristics by which blocks were stratified for the ACE sample allows the missingness to be accounted for within the model. The models included a common set of parameters (within stratum for the stratified model) across all blocks for the ACE effect and interaction effects between ACE and the other lists, and individual parameters for each block for the decennial census and AREX 2000 effects and interaction.

5 Interrelationships and Grand Synthesis

In the separate sections above we have explored what the literature treats as three distinct topics:

- **Multiple Systems Estimation**: To estimate those missed by several merged lists. We used log-linear models and related models with latent variables to provide structure for MSE.
- **Missing Data Methods and EM**: Based on MAR assumptions. We used these methods to describe an alternative to the traditional MSE approach. Extensions allow for covariates (with possible missing values) and missing lists for some covariate combinations.
- **Record Linkage and EM** : This is essentially missing data problem and thus EM plays a prominent role in methods for accomplishing linkage. Record linkage is a prelude to many applications of MSE and logistic regression is a common model structure lying at its center.

Herzog et al. (2007) consider all three topics under the broad rubric of data quality, but their focus does not emphasize what for us are obvious links, some of which we have described here.

Missing data and EM ideas show up in various forms throughout our discussion and thus it seems rather natural to ask whether there is a way to combine record linkage, covariates, and MSE methodologies using missing data framework and assumptions such as MAR. While we have not attempted such a grand unification in this paper, we think that considering the problems in an integrated form will lead to new and improved statistical methodology. One of the main benefits we foresee in this unification is the acknowledgment and

incorporation of the inherent uncertainty that probabilistic record linkage methods for merging multiple lists in a form directly suitable for MSE introduce in MSE estimates, which is ignored in virtually all applications.

References

Agresti, A. (1994), "Simple capture-recapture models permitting unequal catchability and variable sampling effort," *Biometrics*, 50, 494–500.

Anderson, M. and Fienberg, S. (2001), *Who Counts? The Politics of Census-Taking in Contemporary America*, New York: Russell Sage Foundation, revised paperback ed.

Asher, J. and Fienberg, S. (2001), "Statistical variations on an administrative census," in *Proceedings of the Statistical Research Methods Section*, Alexandria, VA: American Statistical Association, available at http://www.amstat.org/sections/SRMS/Proceedings/y2001/Proceed/00554.pdf.

— (2002), "The administrative records experiment in 2000: An application to population count estimation via triple systems estimation," in *Proceedings of the Government Section*, Alexandria, VA: American Statistical Association, available at https://www.amstat.org/sections/SRMS/Proceedings/y2002/Files/JSM2002-000827.pdf.

Baker, S. (1990), "A simple EM algorithm for capture-recapture data with categorical covariates," *Biometrics*, 46, 1193–2000.

Ball, P. and Asher, J. (2002), "Statistics and Slobodan," *Chance*, 15, 17–25.

Belin, T. and Rubin, D. (1995), "A method of calibrating false-match rates in record linkage," *Journal of the American Statistical Assocation*, 90, 694–707.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., and Fienberg, S. (2003), "Adaptive name matching in information integration," *IEEE Intelligent Systems*, 5, 16–23.

Bishop, Y., Fienberg, S., and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press, reprinted in 2007 by Springer-Verlag, New York.

Chao, A., Tsay, P., Lin, S., Shau, W., and Chao, D. (2001), "Tutorial in biostatistics: The applications of capture-recapture models to epidemiological data," *Statistics in Medicine*, 20, 3123–3157.

Cormack, R. M. (1992), "Interval Estimation for Mark-Recapture Studies of Closed Populations," *Biometrics*, 48, 567–576.

Darroch, J., Fienberg, S., Glonek, G., and Junker, B. (1993), "A three-sample multiple-recapture approach

to census population estimation with heterogeneous catchability," *Journal of the American Statistical Assocation*, 88, 1137–1148.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society. Series B*, 39, 1–38.

Ding, Y. and Fienberg, S. (1994), "Dual system estimation of census undercount in the presence of matching error," *Survey Methodology*, 20, 149–158.

Dobra, A. and Fienberg, S. (2003), "How large is the World Wide Web?" in *Web Dynamics*, eds. Levene, M. and Poulovassilis, A., New York: Springer-Verlag, pp. 23–45.

Erosheva, E. (2002), "Grade of membership and latent structures with application to disability survey data," Ph.D. thesis, Department of Statistics Carnegie Mellon University.

Erosheva, E. and Fienberg, S. (2005), "Bayesian mixed membership models for soft clustering and classification," *Classification-the Ubiquitous Challenge: Proceedings of the 28th Annual Conference of the Gesellschaft Für Klassifikation EV, University of Dortmund, March 9-11, 2004*.

Fellegi, I. and Sunter, A. (1969), "A theory of record linkage," *Journal of the American Statistical Assocation*, 64, 1183–2010.

Fienberg, S. (1972), "The Multiple recapture census for closed populations and incomplete 2^k contingency tables," *Biometrika*, 59, 591–603.

Fienberg, S., Johnson, M., and Junker, B. (1999), "Classical multilevel and Bayesian approaches to population size estimation using multiple lists," *Journal of the Royal Statistical Society. Series A*, 162, 383–406.

Herzog, T., Scheuren, F., and Winkler, W. (2007), *Data Quality and Record Linkage Techniques*, New York: Springer-Verlag.

International Working Group for Disease Monitoring and Forecasting (1995a), "Capture-recapture and multiple-record systems estimation I: History and theoretical development," *American Journal of Epidemiology*, 142, 1047–1058.

— (1995b), "Capture-recapture and multiple-record systems estimation II: Applications in human diseases," *American Journal of Epidemiology*, 142, 1059–1068.

Lawrence, S. and Giles, C. (1998), "Searching the World Wide Web," *Science*, 280, 98.

Little, R. and Rubin, D. (2002), *Statistical Analysis With Missing Data*, New York: Wiley, 2nd ed.

Manrique-Vallier, D. and Fienberg, S. (2008), "Population Size Estimation Using Individual Level Mixture

Models," *Biometrical Journal*, in press.

Norris, J. and Pollock, K. (1996), "Nonparametric MLE under two closed capture-recapture models with heterogeneity," *Biometrics*, 52, 639–649.

Pledger, S., Pollock, K., and Norris, J. (2003), "Open capture-recapture models with heterogeneity: I. Cormack-Jolly-Seber Model," *Biometrics*, 59, 786–794.

Rubin, D. (1976), "Inference and missing data," *Biometrika*, 63, 581–590.

Sanathanan, L. (1972a), "Estimating the size of a multinomial population," *Annals of Mathematical Statistics*, 43, 142–152.

— (1972b), "Models and estimation methods in visual scanning experiments," *Technometrics*, 14, 813–829.

— (1973), "A comparison of some models in visual scanning experiments," *Technometrics*, 15, 67–78.

Sutherland, J., Schwarz, C., and Rivest, L.-P. (2007), "Multilist population estimation with incomplete and partial stratification," *Biometrics*, 63, 910–916.

Woodbury, M., Clive, J., and Garson Jr., A. (1978), "Mathematical typology: A grade of membership technique for obtaining disease definition." *Computers in Biomedical Research*, 11, 277–98.

Zaslavsky, A. and Wolfgang, G. (1993), "Triple-system modeling of census, post-enumeration survey, and administrative-list data," *Journal of Business & Economic Statistics*, 11, 279–288.

Zwane, E. and van der Heijden, P. (2007), "Analysing capture–recapture data when some variables of heterogeneous catchability are not collected or asked in all registrations," *Statistics in Medicine*, 26, 1069–89.

Zwane, E., van der Pal-de Bruin, K., and van der Heijden, P. (2004), "The multiple-record systems estimator when registrations refer to different but overlapping populations," *Statistics in Medicine*, 23, 2267–2281.