Discussion on "The central role of the identifying assumption in population size estimation" by Serge Aleshin-Guendel, Mauricio Sadinle, and Jon Wakefield

Daniel Manrique-Vallier.
Department of Statistics, Indiana University,
Bloomington, IN, 47405, USA.
Email: dmanriqu@iu.edu

Abstract

Since the now famous article by Link (2003), identifiability issues in Capture-Recapture (CR) have appeared sporadically in the literature, but neither a unified treatment nor general solution exist. Aleshin-Guendel, Sadinle, and Wakefield attempt to fill this gap by proposing a change of perspective in both the formulation and practice of CR. I contend that their proposal, while bold and thought-provoking, is not a step in the right direction: it fundamentally misunderstands the nature of real-life CR modeling; and attempts to solve problems by prescribing unnecessary and impractically broad restrictions to practice.

1 Introduction

I thank the authors for this thought-provoking piece and for the opportunity of engaging in this fascinating discussion. Since the famous paper by Link (2003), identifiability issues in Capture-Recapture (CR) have been sporadically studied but, to my knowledge, neither a unified treatment nor solution exist. The present work attempts to fill this void by proposing a change of perspective in both the formulation and practice of CR. The authors focus their attention on the constraints needed to ensure the statistical identifiability of the unknown population size ("Identifying Assumptions," IA). They contend that current practice does not pay due attention to IAs and, importantly, to the need of justifying them. This lack of attention, they judge, is due to the almost universal practice of specifying models for complete data (this is, together for observed and unobserved individuals), which implies both an observable-data model and IA, but does not make the latter explicit. Their proposal is prescriptive: they advocate for replacing the current practice of specifying complete-data sampling models, and replacing it with the direct specification and justification of IAs—and, separately and if needed, observable-data models. This perspective, they anticipate, will have the desirable effect of confronting practitioners with the need of selecting and justifying appropriate IAs. Justification for IAs, besides ensuring identifiability, should stem from their reasonable correspondence with the applied data context. Moreover, the authors judge that inability of justifying IAs in an applied situation should preclude the use of CR.

The proposal is intriguing and bold. I agree with the authors that identifiability in CR is an important and understudied topic, and that concepts from the (non-ignorable) missing data literature can be useful tools in the area (Manrique-Vallier et al., 2022). The strategy certainly "puts the [IA] front and center" of CR practice, as the authors claim. Furthermore, in some sense it also solves identifiability problems, as there cannot be identifiability problems where the restrictions needed to avoid them have been purposefully and correctly chosen to that effect.

Unfortunately, the proposal appears to me as more of a wild-goose chase than the workable and fertile framework envisioned by the authors. The framework is formally correct inasmuch as it is mathematically neutral: it is based on the equivalence between complete-data models and the classical conditional decomposition of the CR likelihood (Sanathanan, 1972). Thus, as the authors appropriately attempt to do, its value should be justified from its utility and feasibility, balanced against its potential disadvantages. I contend that not only the authors' evaluation of that balance is flawed, but also that the proposal itself is neither workable nor useful. Moreover, I find the proposal itself to be unnecessarily restrictive, and to be predicated upon unreasonable premises: an implausible idealization of CR modeling; the unwarranted reification of an otherwise useful analytical object; and an out-of-proportion concern regarding the dangers of non-identifiability.

2 Complete data processes and IAs

My first objection to the proposal is to its implied assumptions about CR modeling. These manifest prominently in the requirement of justifying IAs based on the real-life data context. The authors are adamant about this requirement, and in this they adopt the posture of considering both its desirability and feasibility as self-evident. To understand why I believe this is an unreasonable position, it might be useful to contrast it to what I would call a "traditional" understanding of CR.

The main insight behind multiple-list CR is that whenever we have two or more lists whose elements have been sampled from the same population (save unfortunate, but unlikely, sampling coincidences), some individuals will be present in some lists but not in others. Thus, we can learn about the sampling and non-sampling of individuals by studying how some of those individuals appear in some lists, while not appearing in others. For example an individual with capture vector $(0, 1, 0)$ provides information about not-appearing in lists 1 and 3, and information about appearing in list 2. This leads to the CR method: in essence, fit a joint model for the capture vector and use it to *extrapolate* the probability of not being captured by any list. Seen this way, the main limitation of CR results evident. Since all of the records in a multi-list dataset need, by definition, to be present in at least one of the lists, there cannot be any real data directly informing about not being in all lists simultaneously—which, lest we forget, is our inferential target. This means that any inference about non-observed individuals ultimately relies on whatever assumptions the joint model may impose on the relationship between individuals observed at least once and those fully unobserved. This relationship is the IA.

In this view, complete-data distributions stand as representations of joint sampling processes and, as such, should be justified in application-relevant ways. For example, independence (or main-effects log-linear) models can be appropriate if data were collected from a number of truly independent random samples. If instead we suspected that each individual had an intrinsic "capturability" to which all lists were sensitive in the same way, a Rasch model (Fienberg et al., 1999) might be appropriate. Full models such as these represent an understanding of a sampling process and *imply* IAs; e.g. the independence model, begin an instance of a non-saturated hierarchical log-linear model, implies the no-highest-order-interaction (NHOI) IA. These IAs make it in turn possible to estimate the unobserved population. Thus, the specification of a complete-data model can be understood as a method for translating knowledge and assumptions about listing processes into IAs that make sense for a problem—things are, admittedly, not that simple in practice, as some implied IAs might not constrain the CR problem enough; I will come back to this issue in the next section.

The authors reject this procedure and instead favor modeling the IAs directly, without resorting to complete-data models. Their stated position, however, requires us to accept two things. First,

that useful IAs can always be neatly interpreted as corresponding to some distinct and articulable real-life phenomenon or situation; and second, that only interpretable IAs should be considered acceptable. I agree with neither of those premises.

Not all useful IAs admit an intuitive interpretation. A good example is the NHOI restriction in hierarchical log-linear modeling. As the authors note, the NHOI can be notoriously difficult to intuitively interpret as a standalone model with more than $K = 3$ lists. In fact, the authors use it in their analysis of the Kosovo data as an alleged illustration the perils of unjustified IAs. This view, however, has the problem backwards. By definition, any unsaturated hierarchical log-linear model implies the NHOI condition. Therefore, NHOI IAs do not need to be justified on their own: the justification of any parsimonious hierarchical log-linear model as a plausible complete-data representation *is by itself* a justification for the NHOI condition.

The authors' own analysis of the Kosovo data provides a good illustration of why I believe they are wrong in this point. Conveniently, in this case we need not speculate about the plausibility of estimates, as the ground truth is known to be $N = 10,401$; see Manrique-Vallier (2016). The original CR analysis, by Ball et al. (2002), used parsimonious log-linear models (*and therefore, NHOI*) and produced the estimate $\hat{N}_{\text{BALL}} = 10,356$ (95% interval $[9,002,\ 12,122]$). The authors' estimate using only the NHOI IA is $\hat{N}_{\text{NHOI}} = 16,941$ (95% interval [5,304, 28,579]). They dismiss it as biased, and present it as an illustration of the dangers of using an IA that cannot be justified from the data. The problem is that both conclusions are *non-sequiturs*. First, the fact that Ball et al. (2002) assumed NHOI to obtain an essentially on-point estimate should be an indication that NHOI might, at the very least, be plausible. Second, given the former, the fact that both Ball et al. (2002) and the authors assumed NHOI, and that Ball et al. (2002)'s interval is completely contained within the authors', a more reasonable explanation is that the problem with $\hat{N}_{\text{NHOI}}$ is not of bias, but—unsurprisingly, as NHOI is a saturated model—of variance. Therefore, more reasonable conclusions would be that (1) there are IAs that are *useful* but difficult or impossible to justify except as the consequence of simpler-to-justify complete-data models, and (2) avoiding complete-data modeling and isolating IAs to use them on their own can lead to even bigger problems than what the procedure tried to avoid—thus should not be elevated to the category of principle.

The second premise, that only interpretable IAs should be considered acceptable is also problematic, as it unnecessarily rules out otherwise effective procedures, while offering limited or no alternatives. NHOI is an obvious example. Rasch models provide another obvious example: they offer a natural representation of sampling under symmetric individual heterogeneity, yet their IAs are not obviously interpretable (Fienberg et al., 1999). The same considerations apply to the regularized Latent Class models from Manrique-Vallier (2016), which offer a natural representation of the aggregation of an unknown number of homogeneous sub-populations. I will come back to this example to illustrate why I believe that the authors' apprehension about non-identifiability is not only exaggerated, but ultimately self-defeating.

The common thread to these flawed premises is the unwarranted reification of IAs. Indeed, IAs are mathematical objects that correspond to the analytical operation of extrapolation. Then, different from complete-data processes, which by definition correspond to the very concrete and mundane "operation" of generating a population, they need not correspond to any obvious real-life phenomenon. Thus, even though there might be instances of happy coincidences (like those from Section 4.3, which are nonetheless extremely limited), expecting IAs to be both interpretable and interpreted is an unreasonable foundation to predicate practice on.

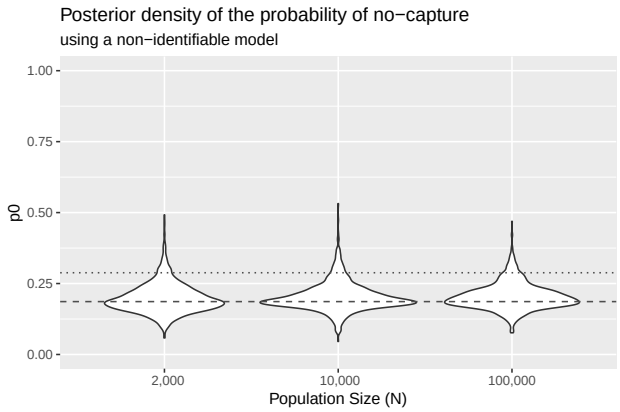

Figure 1: Posterior density of probability of no-capture, p_0, obtained from four CR datasets simulated from the same non-identifiable Latent Class model, for increasing populations sizes. Both the true and the spurious but undecidable values are indicated with dashed lines.

3 Identifiability in CR

My second main objection is to the wisdom of rejecting non-identifiability (NI) as a matter of principle. I contend that this is an impractically maximalist position. Moreover—and here I adopt an unapologetic Bayesian perspective—I believe that the proposal is essentially an exercise at fixing things that are not broken, at an unacceptably high cost.

Identifiability is a desirable property. It is often a necessary condition for establishing fundamental properties, like global consistency and central limit theorems. Statistical procedures that rely on models that exhibit some form of NI often cannot offer the strong guarantees that their identifiable counterparts can. In the case of CR, (conditional) NI occurs when there are points in the parameter space that imply the same observable-data distributions but different population sizes. Such procedures are structurally incapable of discriminating the plausibility of those population sizes based on data. This is an important *theoretical* deficiency. Its *practical* relevance, though, depends on our need for that discrimination in order to obtain *reliable* and *useful* knowledge.

The authors consider that CR procedures under NI are intrinsically unreliable, and therefore it is imperative to avoid them. As evidence they offer a series of examples of the—alleged—failure of CR under NI, under plausible applied scenarios. I have re-evaluated their investigation into this matter, and found it to be thoroughly unconvincing; I provide my detailed account in a Web Supplement accompanying this article. However I did find the exercise illuminating in a different way. For me, far from showcasing NI as the intolerable dead-end that the authors dread, the exercise shows it as something far less nefarious: as a potential limit (sometimes large, sometimes not) to the extent of the knowledge that we can expect to obtain from data.

A simulated example, taken from the authors' work, can help clarifying this idea. In this exercise I consider one of the examples of non-identifiable models described by the authors in their Web Supplement: a Latent Class Model (LCM) with $K = 3$ lists, $J = 2$ latent classes, and parameters calculated according to the authors' Theorem A.2; see Online Supplement for details. As proven by the authors, this is a complete-data model whose induced observed-data distribution is the same as that of another similar LCM, which nonetheless induces a different probability of no-capture, p_0. I generate four CR datasets, taken from four populations of increasing sizes $N = 2000, 10^4$, and 10^5, and obtain their corresponding posterior distributions of p_0 using the R package LCMCR, which implements the regularized LCMs from Manrique-Vallier (2016).

Figure 1 shows the posterior distribution of p_0, estimated from the three simulated datasets. We note that all posterior densities seem to concentrate within a limited region, which contains both the true value $p_0^T = 0.186$ and the spurious $p_0^S = 0.288$. This is significant. It shows that this procedure, despite being based on a non-identifiable model, still provides potentially *useful* data-based knowledge in the form of posterior probability mass concentration around a relatively small set which contains the true value. A second salient feature is that posterior densities do not further concentrate as the population (and thus the sample) size increases. This suggests the procedure's inability of extracting information from data past certain limit. This can be seen as a flaw—e.g. it correctly suggests lack of statistical consistency. However, we can also note that in all cases the regions that concentrate almost all probability mass are as small as they can be while still containing the two (lest we forget) undecidable-based-on-data p_0^T and p_0^S. Moreover, as the sample sizes increases there seems to be a modest but noticeable bimodal concentration around p_0^T and p_0^S, as we would expect. Thus, these results can also be read as evidence of the *reliability* of the procedure: it will extract knowledge from data, but only up to the point allowed by both the data and the method's own intrinsic capacity of using it.

The main lesson of this exercise should be evident: NI is real and it does have an effect in CR procedures; however its mere presence does not automatically render a method useless. Moreover, the limited knowledge obtainable from a reasonably well justified non-identifiable procedure is still more useful than the no-knowledge afforded by the authors' nihilistic advise (*"[if] none of the [IAs] discussed in this article [is appropriate, then] no estimation of the population size should be produced based on the data at hand."*). Similarly, if data were actually generated from a process unable to produce datasets with enough information for discriminating the true population size from an incorrect alternative—as the one in our example—I would consider the humble, but trustworthy, inferences based on procedures that abide by that limitation to be preferable to the false sense of security afforded by more comforting, but potentially misleading, identifiable alternatives.

There are still several important additional factors to consider in this discussion. It will not always be the case that NI allows relatively concentrated posterior distributions—this happened in our example because both the actual and spurious values of p_0 were relatively close to one another. If that were not the case, we would end up with either clearly multimodal or fairly wide posterior distributions. These might not be the easiest to interpret, but would nonetheless be informative. A related issue is that practical Bayesian procedures seldom distribute prior probability mass uniformly over the parameter space, thus procedures under true NI conditions will actually favor some of the undecidable alternatives over others; not because of the data, but because of the prior. This can be observed in our example, in Figure 1. There we see that the posterior density of the spurious value p_0^S is smaller than that of the true value p_0^T, even though they are in principle undecidable from data. In this instance, this occurs because the current implementation of the procedures, in LCMCR, in addition to *a priori* favoring sparse mixtures (due to the regularizing effect of a stick-breaking prior), also *a priori* favors mixture components with small probabilities of no-capture (due to hard-coded hyper-parameter choices). This weakly informative prior can be adequate for applications in which conservative estimates are preferred, but may not be appropriate for others. I am currently working in a revision to the LCMCR package that will give the user a greater level of control over this specification.

None of these issues, however, are solved by proscribing procedures, nor by attempting to fix them with the addition of not-yet-discovered but supposedly justifiable additional IAs. If anything, they call for good old-fashioned research into the properties of specific procedures and, importantly, into their potential effects in statistical practice. It also showcases the need for more research into the specification of weakly informative prior distributions—e.g. I am currently working on methods for translating the almost always available knowledge about reasonable bounds on the extent of

sampling coverage into workable prior specifications for RLCMCR models.

4 Conclusion

In this note I have laid out my case against the authors' proposal. Summarizing, I find it to be neither a practical solution to a problem, nor a good starting point for better practice. The posture is of unyielding prescriptivism. This leads to a logically consistent and elegant framework, that is nonetheless unworkable and hindering. Unworkable, because it unreasonably requires the general ability of interpreting IAs. Hindering, because it rules out potentially useful tools without offering good alternatives: methods that do not neatly conform to the prescription, e.g. those that imply too-complex-for-interpretation IAs; and methods that can be sub-optimal, but that nonetheless can still provide useful knowledge in practice, e.g. methods that may exhibit NI.

References

Ball, P., Betts, W., Scheuren, F., Dudukovic, J., and Asher, J. (2002), "Killings and Refugee Flow in Kosovo, March–June, 1999," Report to ICTY.

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.

Link, W. A. (2003), "Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities," *Biometrics*, 59, 1123–1130.

Manrique-Vallier, D. (2016), "Bayesian Population Size Estimation Using Dirichlet Process Mixtures," *Biometrics*, 72, 1246–1254.

Manrique-Vallier, D., Ball, P., and Sadinle, M. (2022), *Capture-Recapture for Casualty Estimation and Beyond: Recent Advances and Research Directions*, Cham: Springer International Publishing, pp. 15–31.

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

Supporting Information

Web Appendices referenced in Section 3 are available with this paper at the Biometrics website on Wiley Online Library. Computer code to reproduce the results presented in this article and the Supplement is also available there.