



## Extending DSE methods to deal with domain misclassification when compiling Population Estimates

Sanela Smith<sup>1</sup>; John Dunne<sup>1</sup>

<sup>1</sup> Central Statistics Office, Ireland

### Abstract:

Since 1946, Ireland has typically conducted a traditional census where enumerators knock on every door at 5 yearly intervals. Such a census is costly and time consuming. Today, as with many other countries, CSO has access to significant amounts of administrative data for statistical purposes that may make alternative census models, with lower cost and higher frequency, more feasible.

One such model showing promise at CSO is based on first compiling a Statistical Population Dataset (SPD) based on signs of life and then adjusting counts from the SPD using a second administrative list and Dual System Estimation (DSE) methods to obtain population estimates.

However, DSE methods can break down when stratifying by domain due to domain misclassification. This paper describes an extension to DSE methods that can correct for domain misclassification in one list (list A) by adjusting the list A count for misclassification before compiling the population estimate. The methods were then applied in a real-world scenario.

### Keywords:

Administrative data; Census; Dual system estimation

### 1. Introduction:

For the countries that do not have a Central Population Register (CPR) on which demographics statistics can be compiled, the production of reliable demographic statistics on population counts and migration flows can prove challenging. This is particularly true for those countries, like Ireland, that have relatively high migration flows that are difficult to estimate. In the absence of the CPR, the simple idea is to compile a Statistical Population Dataset (SPD) using available data sources. The ideal SPD will have a record for each statistical unit (person) in the target population – each unit identified with a unique identification number.

The approach taken to date in the Irish PECADO (Population Estimates Compiled from Administrative Data Only) project has been to use a signs of life approach to build an SPD and then adjust SPD counts using DSE methods with a second unused administrative list satisfying the necessary assumptions. Use of high quality identification numbers and signs of life eliminates errors associated with overcoverage and linkage error. High quality information for age and sex ensures no domain misclassification for age and sex, enabling post stratification by these variables in applying DSE methods to adjust for undercoverage.

There is however a challenge in disaggregating by geography as geography is not always categorised correctly or consistently across data sources. Therefore, post-stratification by geography to obtain a geographical breakdown will lead to inflated estimates for affected geographical domains.

This paper presents an extension<sup>1</sup> of previous DSE methods by Zhang and Dunne (2018) used in the PECADO project that adjusts for domain misclassification in one list (list A). The idea behind the method is to first estimate the number of units in list A that are misclassified and then reduce the size of list A by this amount before using the standard DSE formula.

As a suitable list B is yet to be identified or built with the correct geography within the PECADO project, an alternative use case was identified, namely investigating undercoverage in the 2016 Census, to see if the method extension would provide plausible results. Undercoverage in the 2016 Census is explored in a counter intuitive way by considering the Census itself as a coverage survey on an administrative list suffering from undercoverage (list A), whereby the Census list is assumed to satisfy the assumptions associated with list B in a DSE setup. The Census list is also considered to have the correct geography for each unit.

This paper presents the extension to the DSE methods that incorporate adjustments for misclassification in list A before presenting the Use case where an administrative list is used to explore undercoverage in a traditional Census.

## 2. Methodology:

### 2.1. DSE structure

There are several modifications of the basic DSE which uses two lists (A and B), both of which are subsets of the population, to estimate the true size of the whole population. Here we will consider the modification presented by Zhang and Dunne (2018).

Let A be the first list of size  $x$ . Suppose list A is subject to undercoverage so that  $x < N$  and  $A \subset U$ , where  $U$  is the total population.

Let B be the second list of size  $n$  and also subject to undercoverage so that  $n < N$  and  $B \subset U$ .

Suppose the records in list A and list B can be linked in an error free manner and doing so will provide the matched list AB with  $m$  records common to both list A and list B.

Let  $\delta_{iB} = 1$  if  $i \in B$ , noting  $B \subset U$ , and 0 otherwise. We assume that the probability  $P(\delta_{iB} = 1) = \pi$  is constant across  $i \in U$ . This is the basis for the development of the estimator. This gives the three assumptions underpinning the DSE method

- i) no erroneous records in either list A or list B
- ii) error free matching between list A and list B and
- iii) homogeneous capture with respect to list B or each unit in the population has equal probability of being included in list B.

Heterogeneous capture can be accommodated through post stratification to ensure that the homogeneous capture assumption holds within each stratum.

Given the assumption of homogeneous capture, we have

$$E[n] = N\pi \quad (1)$$

Assuming that homogeneous capture and error free matching assumptions are valid, we have:

$$E[m|\delta_A] = x\pi \quad (2)$$

which is the expected number of records in list  $A \cap B$  after applying the constant capture probability  $\pi$  to the  $x$  records in the list A. replacing  $E[n]$  by  $n$  and  $E[m\delta_A]$  by  $m$ , we get the following:

$$\hat{N} = \frac{nx}{m} \quad (3)$$

This estimator works perfectly well on the State level or when there is no misclassification. When trying to get estimates for the population subgroups, for example smaller geographical areas, the misclassification can cause some problems. Let's look at the data structure when there is misclassification in the data.

---

<sup>1</sup> Original DSE extension is presented in unpublished paper by L.-C. Zhang, University of Southampton

The structure of the data for each domain in this case can be represented with the following form

	n		
x	m	(m')	(α)
	u	(u')	
	a	(a')	

Cells in the first row form a partition of the list A. x is the total number of enumerated records, m is the number of records matched to list B, m' is additional number of records in case list B covers the whole population and α is the number of misclassified records that actually belong to other domains. Only x and m are observed. Similarly, the cells in the first column form a partition of the list B. n is the total number of enumerated records, u is the number of records that don't belong anywhere, u' is the remaining number of records that don't belong anywhere in the case that list B covers the whole population, a is the number of records that are matched to the list A elsewhere and a' is the additional number of such records in case list B covers the whole population. In case there is no misclassification, structure of the data and formula for estimating the population becomes the ideal DSE.

## 2.2 Calibration

Let U be the total population. Let  $\delta_l = 1$  if a unit  $l \in U$  is found in the list B, and 0 otherwise. Suppose that catch probability within list B varies across domains but is constant within each domain

$$\pi_i = P(\delta_l | l \in U_i) \quad (4)$$

for domain population  $U_i$ , where  $i = 1, \dots, A$ . Let x denote all the domain specific list A's. from the data structure shown above

$$x_i = m_i + m'_i + \alpha_i \quad (5)$$

Following the equal catchability assumption for each domain, total domain specific population can be estimated with

$$N_i = m_i + m'_i + u_i + u'_i + a_i + a'_i \quad (6)$$

We know that misclassification happened between the domains and it doesn't have an impact on overall population, therefore, total inflow population must be equal to total outflow population.

$$\sum_{i=1}^A \alpha_i = \sum_{i=1}^A (a_i + a'_i) = \sum_{i=1}^A \xi_i x_i, \xi_i = \frac{E(a_i|x)}{E(m_i|x)} \quad (7)$$

where A is the number of domains within the population.

In the best-case scenario when there is no misclassification, the catch rate,  $\pi_i$  can be estimated by  $\frac{m_i}{x_i}$ . However, if the misclassification does exist, it must be considered in calculating the catch rate. For that purpose, the misclassification must be estimated first. Let  $a_{ji}$  be the number of records in list B in  $j^{th}$  domain that are matched to  $i^{th}$  domain in the list A. The estimator for  $\alpha_i^*$  is given as

$$\alpha_i^* = \sum_{j \neq i} d_j \frac{a_{ji}}{\pi_j} \quad (8)$$

where  $d_j$  is design weight for the  $j^{th}$  domain in list B. By solving the fixed-point equation system

$$\pi_i = \frac{m_i}{x_i - \sum_{j \neq i} d_j \frac{a_{ji}}{\pi_j}} \quad (9)$$

With calculated catch rate in this way, DSE is given by

$$N_i = \frac{n_i}{\pi_i} \quad (10)$$

This estimator could be biased due to non-linear nature of the equation system, so it needs to be adjusted to satisfy the natural formula for total misclassification.

Finally, we get the following solution:

$$\begin{aligned} \hat{\alpha}_i &= \alpha_i^* g_i \\ g_i &= 1 + \lambda(1 + \xi_i) \\ \lambda &= \frac{\sum_{i=1}^A \xi_i x_i - \sum_{i=1}^A (1 + \xi_i) \alpha_i^*}{\sum_{i=1}^A (1 + \xi_i)^2 \alpha_i^*} \end{aligned} \quad (11)$$

And the corresponding calibrated DSE (CDSE)

$$\hat{N}_i = \frac{n_i(x_i - \hat{\alpha}_i)}{m_i} \quad (12)$$

### 3. Results

The methods are used to explore the possibility of undercoverage of the Irish Census of population. Counterintuitively the Census is considered a coverage survey of an administrative list. In summary list B is compiled from the Census removing all records where stratification variables are incomplete or invalid and all records that are missing the identification key to match to administrative records. List B is therefore a trimmed census list and that trimming is assumed to not violate the homogenous capture assumption for list B. List A is compiled from valid health administrative records and social welfare records. The size of the lists/datasets is shown in Table 1.

Datasets	Full size	List size
2016 Census (list B)	4.74 million	4.28 million
Admin data (list A)	2.6 million	2.3 million

*Table 1 Datasets used for CDSE*

Both lists were stratified by sex, age (81 age category, 80+ were grouped together) and 26 counties. That gave 4212 domains.

It is assumed that geographical domain variables on census dataset or list B are correct as the methodology only corrects for misclassification in list A.

After the data was prepared, the above methodology was applied to each of the domains to estimate the true population within that domain.

Firstly, all misclassified records were identified.

Domain	County name	Should be elsewhere ( $\alpha_i$ )	Should be here ( $a_i$ )	Census 2016	DSE estimate	CDSE estimate
1	Cork	2,077	1,778	542,868	569,100	564,870
2	Kerry	908	658	147,707	153,580	150,780
3	Limerick	1,803	1,315	194,899	205,490	201,590
4	Clare	976	1,183	118,817	126,920	124,770
5	Mayo	1,149	1,363	130,507	139,100	136,830
6	Galway	1,997	1,474	258,058	271,470	267,760
7	Leitrim	543	506	32,044	35,350	34,510
8	Sligo	733	536	65,535	69,570	68,560
9	Donegal	561	398	159,192	167,460	166,050
10	Waterford	2,388	1,970	116,176	129,110	125,840
11	Wexford	2,045	1,007	149,722	162,140	159,340
12	Wicklow	2,835	1,447	142,425	155,890	141,450
13	Kildare	2,797	2,522	222,504	242,000	233,720
14	Kilkenny	986	2,872	99,232	105,020	103,270
15	Tipperary	2,660	1,651	159,553	173,520	170,320
16	Offaly	1,373	2,314	77,961	84,970	82,780
17	Longford	573	468	40,873	45,070	44,030
18	Monaghan	476	540	61,386	65,770	64,960
19	Louth	1,989	755	128,884	140,270	136,750
20	Roscommon	913	1,720	64,544	70,410	68,600
21	Meath	2,783	3,414	195,044	213,150	202,100
22	Westmeath	2,073	1,164	88,770	97,960	95,090
23	Carlow	2,947	1,755	56,932	66,820	63,590
24	Cavan	722	1,043	76,176	81,250	79,610
25	Laois	1,215	3,716	84,697	91,760	89,310
26	Dublin	4,256	6,173	1,347,359	1,412,610	1,405,940
Total				4,761,865	5,075,760	4,982,420

Table 2 Misclassification and final estimate by county

After the misclassification was identified, the CDSE methodology was applied that created and adjusted the estimates. Estimates are shown in Table 2, compared with Census data. These population estimates show slight Census undercoverage of around 4.6%. There is slightly higher undercoverage of male population aged 20 to 60, but that was expected.

If ideal DSE was applied, without the adjustment for misclassification, the total population would be overestimated at 5.08 million.

#### 4. Discussion and Conclusion:

This paper presents an extension to DSE methods that corrects for domain misclassification in list A. The purpose of this evaluation at CSO is to disaggregate population estimates by geography where there is the possibility of incorrect geographic attributes on individual records on list A, geographic attributes for list B are assumed to be correct. The consequence of not identifying and adjusting for such domain misclassification is inflated population estimates. While the focus of the application is population estimates compiled from administrative data, in the absence of a suitable list B an alternative use case is used to determine the plausibility of the methods. To investigate the plausibility of the method a novel use case was identified where an administrative list (list A) could be used to estimate undercoverage in a Census (list B). Typically, list B is identified as the coverage survey but, in our application, we consider the Census as a coverage survey on an administrative list and obtain an estimate of undercoverage on the Census. This is a reasonable use case to consider as only one list is required to satisfy the *homogeneous capture assumption*.

The application considers misclassification in list A at a county level (26 counties) and adjusts and calibrates the DSE estimator accordingly. The results look plausible with the unadjusted population estimate being 2% higher than the adjusted population estimate. Furthermore, this application points to undercoverage in the Census of Population which requires further investigation. Ireland is committed to conducting a coverage survey in the 2022 Census for the first time. If an administrative list could reasonably be used to evaluate coverage in Census operations, similar as to what is done here, then potentially this would simplify Census operations with respect to coverage surveys while at the same time saving significant time and money.

Finally, the application of this method, as described here, provides for significant promise in the application of these methods in the PECADO setting where the SPD is compiled using signs of life from administrative data sources and a list B is compiled from another SPD excluded list that satisfies the necessary assumptions and has accurate geographical information.

Further work planned includes evaluating to what extent stratification and clustering can be used and combined for providing population estimates at a higher geographic resolution (Electoral District - Ireland has approximately 3,500 EDs).

**References:**

1. Dunne, J. and Graham, P. (2019). New Population Estimation Methods: New Zealand and Ireland. In ISI World Statistics Congress 2019, number August, Kuala Lumpur
2. Wolter, K. M. (1986). Some Coverage Error Models for Census Data. Journal of the American Statistical Association
3. Zhang, L.-C. (2019). A Note on Dual System Population Size Estimator. Journal of Official Statistics
4. Zhang, L.-C. and Dunne, J. (2018). Trimmed Dual System Estimation. In Bohning, D., van der Heijden, P. G., and Bunge, J., editors, Capture-recapture methods for the Social and Medical Sciences