CSAC, September 18-19, 2014 Discussion: 2015 Census Address Validation Test (AVT)

Noel Cressie
National Institute for Applied Statistics Research Australia (NIASRA)
University of Wollongong
Australia
(ncressie@uow.edu.au)

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG

2015 ADDRESS VALIDATION TEST

The 2015 AVT consists of two parts:

1. Master Address File Model Validation Test (MMVT)
2. Partial Block Canvassing (PBC)

- The MMVT has sampling design at its core but not the PBC.
- The MMVT seems to have more methodological maturity than the PBC.
- The two modeling approaches of the MMVT *do* have a lot in common! (Each could be used to measure MAF quality and identify blocks to canvas.)
- The PBC should be subjected to an initial validation, where not only "positives" are evaluated, but "negatives" too.

“POSITIVES” AND “NEGATIVES”

In the PBC component of the 2015 AVT, suppose a “positive” is a block identified for PBC and a “negative” is a block identified for regular canvassing.

- There do not appear to be any negatives considered in the PBC component? All the assessments concentrate on the positive blocks.
- In the presence of uncertainty, some of the negatives are “true” and some are “false”; and some of the positives are “true” and some are “false.”
- A part of the validation is to determine the False Discovery Rate (FDR) as well as the False Nondiscovery Rate (FNR).

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG

A 2x2 TABLE

Let A_N denote the number of negatives, and let A_P denote the number of positives. Then, in obvious notation, we have

$$A_{TN} + A_{FN} = A_N \qquad (1)$$
$$A_{FP} + A_{TP} = A_P \ ,$$

where $A_N + A_P$ is the number of items being validated. Equation (1) implies a 2x2 table, whose row totals are given by A_N and A_P .

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG

Specificity (alpha) and Sensitivity (pi)

Medical testing uses:

Specificity: $Sp = A_{TN} / (A_{TN} + A_{FP})$

Sensitivity: $Se = A_{TP} / (A_{FN} + A_{TP})$

The "best" value is (Sp, Se) = (0,1).
This approach looks at the columns of the 2 x2 table.

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG

FDR and FNR (look at the rows)

Modern statistical genetic testing uses:

False Discovery Rate:

$FDR = A_{FP} / (A_{FP} + A_{Tp})$

False Non Discovery Rate:

$FNR = A_{FN} / (A_{TN} + A_{FN})$

The "best" value is (FDR, 1 - FNR) = (0,1).
When there are a lot of individual validations, this approach is highly preferable.

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG

ANSWERS TO QUESTIONS

1. Run periodic validation tests, being sure to have a sampling design at the core of each.
2. They are either covariates in the logistic regression or they determine the parameters of the statistical model fitted to the number of “adds.”
3. Yes, what you do should be determined by a cost-benefit analysis.

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG