Downloaded from www.sacemaquarterly.com | The quarterly update on epidemiology from the South African Centre for Epidemiological Modelling and Analysis (SACEMA).

Published: September 2012

The use of clusters to estimate recent transmission of TB

Pieter Uys - Part-time researcher at SACEMA

Primary versus reactivation TB

A new case of TB is the outcome of a recent infection event (primary TB) or is the result of the reactivation of a latent infection acquired some years previously. In a community where TB is endemic it is important to know the extent to which primary cases contribute to the overall burden as this can inform strategies to deal with the epidemic. For instance in many developed countries in Europe it is known that most cases are reactivation cases. These result from transmission events many years previously when the incidence of TB was considerably higher. Eventually this reservoir of latent TB will be exhausted and then, in principle, mostly primary cases can be expected. These would arise mainly because of immigration or infection acquired outside of these countries. Controlling TB in such circumstances is relatively easy. By contrast, in high incidence communities on-going transmission is a problem that requires the identification of more stringent performance targets for TB control. For example, TB control programs should vigorously pursue improvements in case detection, reductions in diagnostic delay and time to effective treatment. Again, the relative contribution made by primary TB needs to be monitored in order to assess the effectiveness of current control measures.

Clustered individuals as recent transmission indicator

Cluster analysis provides a way to estimate the proportion of recent transmission. Sputum specimens from cases reporting to clinics are cultured and the TB strains are identified, commonly using molecular techniques of DNA ‘fingerprinting’ such as restriction fragment length polymorphism (RFLP). By comparing these fingerprints from various patients it becomes possible to classify them as unique or clustered. Unique cases each form what is termed a ‘singleton cluster’. Such cases are apparently not the result of recent infection events within the community but must be due to reactivation. There is also the possibility that they involve immigration either of the individual or of infection acquired outside the community. Two cases yielding the same type, and hence in the same cluster, are usually considered likely to be directly ‘linked’ in the following sense: either one case is the result of infection by the other, or they have both been infected by the same person. Again, the immigration effect may play a role. The proportion of clustered individuals can then be used as an indicator of the proportion of on-going or recent transmission.

'n' and 'n-1' method

There are two common rules of thumb for estimating the proportion of cases due to recent transmission: the ‘n method’ and the ‘n-1 method’. The former uses the proportion of cases in clusters, p_n, as a proxy for the proportion of cases due to recent transmission. In the latter, one case from each cluster is assumed to be an index case, and the proportion of non-index cases, p_{n-1}, is used as a measure of recent transmission.

By adopting the following notation: A is the number of cases, M is the total number of clusters, and A_1 is the number of unclustered cases (i.e. the number of singleton clusters), the formulae for these two methods are respectively given by $p_n = (A - A_1)/A$ and $p_{n-1} = (A - M)/A$. (The latter formula gives the proportion of cases which are not the first case in a cluster and hence gives the proportion of non-index cases). Since the total number of clusters, M, includes the singleton clusters, A_1, it will be seen that the ‘n-1 method’ always leads to a lower estimate of the proportion of recent transmission. This is not an issue as long as one method is used consistently in an investigation.

Underestimation of recent transmission

This may seem straightforward enough, at least in theory, but the practical implementation is beset with all manner of difficulties that tend to lead to under-estimation of the extent of recent transmission. For a start it is unlikely that one will be able to identify every active case in a community. This difficulty is compounded by the necessity, in principle, of finding all cases with the same strain as any one particular case, and to do this for all cases. It is extremely unlikely that this could be achieved. As a consequence clusters that actually contain two members might be ‘observed’ to be singletons and this will contribute towards an under-estimate of transmission. Moreover, in practice, data concerning cases will have been collected only over some limited period of years and cases outside of this period that are possibly related to cases within the period will be missing. This too will lead to under-estimation of the proportion of cases due to recent transmission. Even among the sputum confirmed TB subjects encountered (typically self-reporting to clinics), not all sputum specimens will be successfully typed. This may be partly due to difficulties in culturing certain of the strains.

The process of typing a specimen requires categorizing the specimen according a hierarchy of genotype families and strains within the families. The problem is that the strains in circulation mutate. Possibly then, two related cases will be categorized as different when in fact one is just a mutated version of the other. This, again, will lead to under-estimation of recent transmission.

However, the proportion successfully typed (sampling rate) is, of course, known. To use statistical nomenclature, the cases that are successfully typed thus constitute our 'sample' from the 'population' of the sputum confirmed TB subjects. Throughout this discussion the words 'sample' and 'population' will have these specific meanings.

Frequency distribution bias

Finally, even with the population and sample at hand there is a further source of bias. Bias in the number of unique cases, for example, exists due to contributions resulting from sampling the larger clusters, e.g. 10 clusters of size 4 when sampled at a rate of 0.5 may present as 3 singleton clusters (uniques) together with 3 doublet clusters, 1 triplet cluster and 2 clusters of size 4. For the same reason, bias arises in the number of clusters of each size and in the total number of clusters. This source of bias will be called frequency distribution bias and is purely combinatorial.

To clarify this point we denote the actual number, in the population, of clusters containing *n* related cases, by A_n, $n = 1, 2 \dots N$. We denote the observed number, in the sample as determined by the successfully typed sputum specimens, by S_k, $k = 1, 2 \dots$ The S_k are then sums of the numbers of clusters of corresponding size obtained by sampling from the population and depend on the values of the A_n for $n \geq k$. This is illustrated in table 1 and figure 1 which show the result of a sampling at the rate of 0.3 from a hypothetical population. The frequency distribution bias can be seen by calculating $p_n = (A - A_1)/A$ using values from the population and comparing with values from the sample $p_n = (S - S_1)/S$. For the population, the number of cases, A, is given by the sum of the number of clusters of each size multiplied by the size of the cluster. This is 1383. So $p_n = \frac{A-A_1}{A} = \frac{1383-682}{1383} = 0.51$.

For the sample, the same calculation gives $p_n = \frac{S-S_1}{S} = \frac{427-290}{427} = 0.32$. We see that the value of p_n determined from the sample is considerably smaller than the value found from the population. This illustrates that estimates, p_n, of the proportion of recent transmission made using sample data is biased and underestimates the actual extent of recent transmission. We should therefore use inferred population data. However the fact that different samples can yield different histograms such as the one shown in figure 1 makes it obvious that it is impossible to infer the population histogram from a sample histogram.

Table 1. Contributions to S(k) from A(n) at a sampling rate of 0.3 from a hypothetical population.

Cluster size (n)	1	2	3	4	5	6	7
A(n)	682	95	44	0	37	30	2
			Contributions to S(k) from A(n)				
	212						
	34	11					
	20	14	2				
	0	0	0	0			
	12	6	6	1	0		
	12	7	6	2	1	0	
	0	1	0	0	0	0	0
S(n)	290	39	14	3	1	0	0

Notes: The column headed 1 shows the contributions to the total observed number of singleton clusters, S(n) = 290, arising from sampling the clusters of various sizes in the population. For example the 95 doublet clusters in the population yield 34 singleton clusters (and 11 doublet clusters). Similarly for the total observed number of clusters of other sizes. A total of 3 clusters of size 4 is observed even though there are no clusters of that size in the population. It is important to note that this table shows the outcome of just one particular sampling. One could expect other samplings to be different.

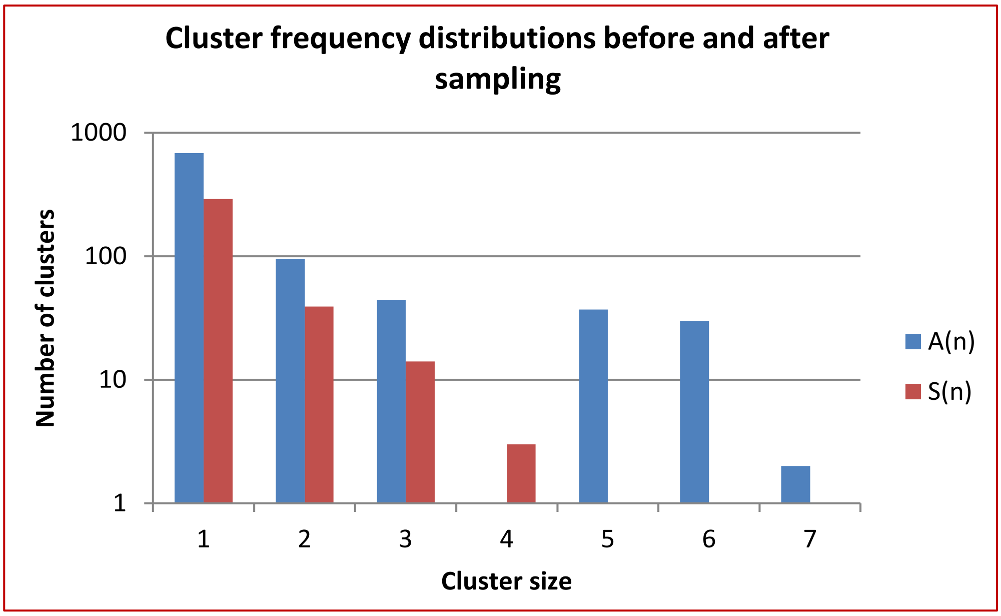

Figure 1. Cluster frequency distribution in a hypothetical population and the distribution in a sample taken at the rate 0.3.
Note that the clusters of sizes 5, 6, and 7 in the original population are all 'lost'. They are represented in the sample by smaller clusters.

Fortunately, however, it is not necessary to estimate the population histogram. All that is actually needed to estimate p_n or p_{n-1} are estimates of the following population parameters: A, the total number of cases, A_1, the number of singleton clusters, and M, the total number of clusters. Of course, A, the reported number of positive TB diagnoses in the study, is known. Appendix 1 shows how to derive unbiased estimators for M and A_1. Estimators for p_n and p_{n-1} may be directly derived by substituting these values into their definitions. The uncertainty (standard error) inherent in the estimators are also analyzed in order to obtain confidence intervals.

The inversion method presented in the Appendix provides a simple way of estimating cluster parameters for a population and hence obtaining practically unbiased estimates for p_n and p_{n-1}. The inversion method was tested by applying it to hypothetical populations at various sampling rates. It was found that with a sampling rate of 0.7, the biases in the estimates of p_n and p_{n-1} are in the range 0% - 0.5%. In contrast to this, using the naïve method the biases in the values of these quantities lie in the range 5% - 10% (figure 3). With the naïve methods, the bias becomes very large at sampling rates of less than 0.5. However the inversion method, applied to estimating p_{n-1}, shows little bias (0.5%) even at low sampling rates.

In summary, the bias using the inversion method is negligible for sampling rates in excess of 0.5 and even lower sampling rates are acceptable when estimating p_{n-1}. The naïve method, on the other hand, provides excessively biased estimates. There is therefore no justification for persisting with the naive approach when a simple and effective method is available.

A moot question is to what extent has the contribution to TB incidence made by transmission been underestimated in the past? This could have serious implications for epidemiological models of TB and control strategies informed by those models especially in high incidence regions.

Pieter Uys, Part-time researcher at SACEMA. Area of interest is the epidemiology of TB focussing on transmission, reinfection, acquisition of resistance and the evolution of resistant strains. The effect of delays in diagnosis is also a theme. *pieter@edserve.co.za*

This article is based on: van Schalkwyk C, Cule M, Welte A, van Helden P, van der Spuy G, Uys P. (2012) Towards Eliminating Bias in Cluster Analysis of TB Genotyped Data. PLoSONE 7(3): e34109.

Appendix

Firstly, some notation is needed:

The population vector of cluster size frequencies is given by $\boldsymbol{A} = (A_1, \ldots, A_N)$. The sample vector of cluster frequencies (a histogram of observed cluster sizes) is given by $\boldsymbol{S} = (S_1, \ldots, S_N)$. Of course, it is not possible to know N, the true size of the largest cluster. Thus, a truncated vector $\tilde{\boldsymbol{S}} = (S_1, \ldots, S_{\tilde{N}})$, is observed, where $\tilde{N}$ ($\tilde{N} \leq N$) is the largest observed cluster size.

Let **P** denote the matrix **P** = **P**(*N*) = $\begin{bmatrix} p(1,1) & p(1,2) & p(1,3) & \cdots & p(1,N) \\ 0 & p(2,2) & p(2,3) & & p(2,N) \\ 0 & 0 & p(3,3) & & p(3,N) \\ \vdots & & & \ddots & \vdots \\ 0 & & \cdots & & p(N,N) \end{bmatrix}$

where

$$p(k,n) = \frac{\binom{n}{k}\binom{A-n}{S-k}}{\binom{A}{S}}$$

is the hypergeometric probability mass function, which represents the probability that a population cluster of size n presents as a cluster of size k in the sample.

The expected value of S_k is then given by

$$\mathrm{E}(S_k) = \sum_{n \geq k} p(k,n) A_n$$

or, in matrix notation, $\mathrm{E}(\mathbf{S}) = \mathbf{P}\boldsymbol{A}$. **(1)**

From equation 1, a crude estimate $\widehat{\boldsymbol{A}}$ of $\boldsymbol{A}$ may be obtained by simply matching the first moment, that is,

$$\widehat{\boldsymbol{A}} = \mathbf{P}^{-1}\boldsymbol{S} \quad \textbf{(2)}$$

The crucial quantities of interest are the total number of clusters $M = \sum_{i=1}^{N} A_i$ and the number of singletons, A_1. An estimator of M can be derived as:

$$\widehat{M} = \sum_{i=1}^{N} \widehat{A}_{\imath} \quad \textbf{(3)}$$

The first element of the vector $\widehat{\boldsymbol{A}}$ is an estimate for A_1. These are unbiased as a consequence of equation 1. Similarly the estimate of $\widehat{\boldsymbol{A}} = \mathbf{P}^{-1}\boldsymbol{S}$ can be used to obtain estimates $\hat{\sigma}^2_{\hat{M}}$ and $\hat{\sigma}^2_{\hat{A}_1}$ of the variances σ^2_M and $\sigma^2_{A_1}$.

Since $\hat{M}$ is a linear combination of random variables (albeit with some dependence), a normal approximation to the distribution of $\hat{M}$ seems reasonable. Assuming this approximation is valid, the estimate $\hat{\sigma}^2_{\hat{M}}$ of σ^2_M can be used to provide the approximate (1 - α) - level confidence interval of

$$\left[\hat{M} - z_{\alpha/2}\hat{\sigma}_{\hat{M}}, \hat{M} + z_{\alpha/2}\hat{\sigma}_{\hat{M}}\right]$$

Similarly, an approximate (1 - α) - level confidence interval for A_1 is given by:

$$\left[\hat{A}_1 - z_{\alpha/2}\hat{\sigma}_{\hat{A}_1}, \hat{A}_1 + z_{\alpha/2}\hat{\sigma}_{\hat{A}_1}\right]$$

We have still not arrived at a solution, however, since the estimates in equation 3 involve an unknown quantity, namely N, the maximum population cluster size. However, from the upper triangular structure of **P** and the fact that $S_n = 0$ for $n > \tilde{N}$, the estimates are unchanged if $\mathbf{P}(N)$ is replaced by $\mathbf{P}(\tilde{N})$ and $\boldsymbol{S}$ by the observed vector $\tilde{\boldsymbol{S}}$.

There is still one final issue: The matrix **P** is close to singular when $\tilde{N}$ is large. For example, if A = 500, S = 400 (i.e. the sampling rate is 0.8), n = 20 and s = 5, then the hypergeometric probability mass function $p(k,n)$ has the value 2.589E-05. With $\tilde{N}$ large, many such small entries will occur in **P**.

To overcome this computational difficulty, a truncation approach to the use of the **P** matrix is now introduced. The observed vector $\tilde{\boldsymbol{S}}$ is divided into two parts $\boldsymbol{S}_0$ (the first C components, which lead to a numerically stable inversion of **P**) and $\boldsymbol{S}_1$ (the remaining $\tilde{N} - C$ components). (Figure 2)

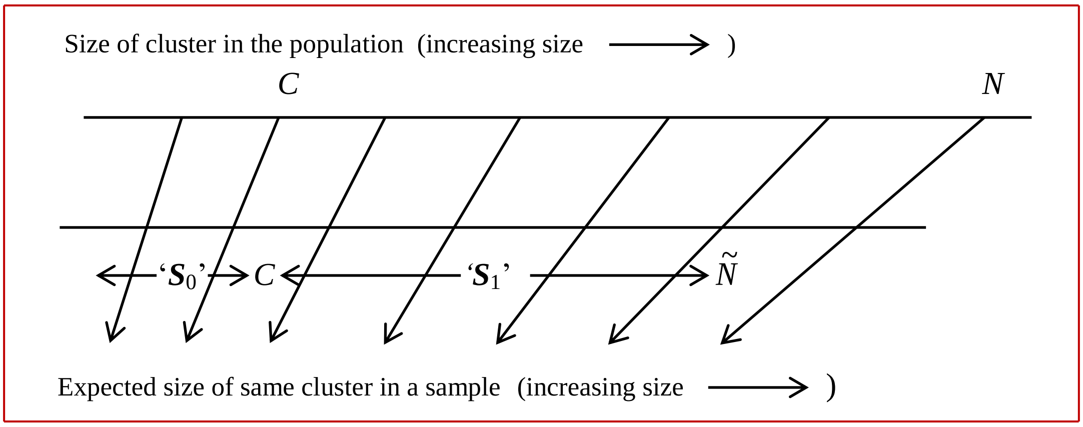
Size of cluster in the population (increasing size ⟶)
C
N
'S0'
C
'S1'
Ñ
Expected size of same cluster in a sample (increasing size ⟶)

Figure 2. The clusters of size greater than C in a sample arise from larger clusters in the population.

Since the clusters of size greater than C in a sample arise from larger clusters in the population it follows that an estimate of the total number of clusters in the population is the number of clusters of size greater than C appearing in the sample together with the estimate of the number of clusters in the population obtained using $\widehat{\boldsymbol{A}} = \mathbf{P}^{-1}\,(C)\,\boldsymbol{S}_0$.

The key question that arises is whether the truncation leads to unbiased estimates of the number of clusters, M, and the number of singletons, A_1. Firstly, the maximum cluster size for which the **P** matrix is still numerically non-singular will be the maximum cluster size at which the vector $\boldsymbol{S}_0$ can be truncated. Within this range, it is now possible to explore bias and variance of estimates of M and A_1 and hence for p_n and p_{n-1}. This was done for hypothetical populations at various sampling rates and truncations. It was found that with a sampling rate of 0.7 and truncating with C = 6, the biases in the estimates of p_n and p_{n-1} are in the range 0% - 0.5%. In contrast to this, the naïve values of these quantities lie in the range 5% - 10% (figure 3). With the naïve methods, the bias becomes very large at sampling rates of less than 0.5. However the inversion method, applied to estimating p_{n-1}, shows little bias (0.5%) even at low sampling rates.

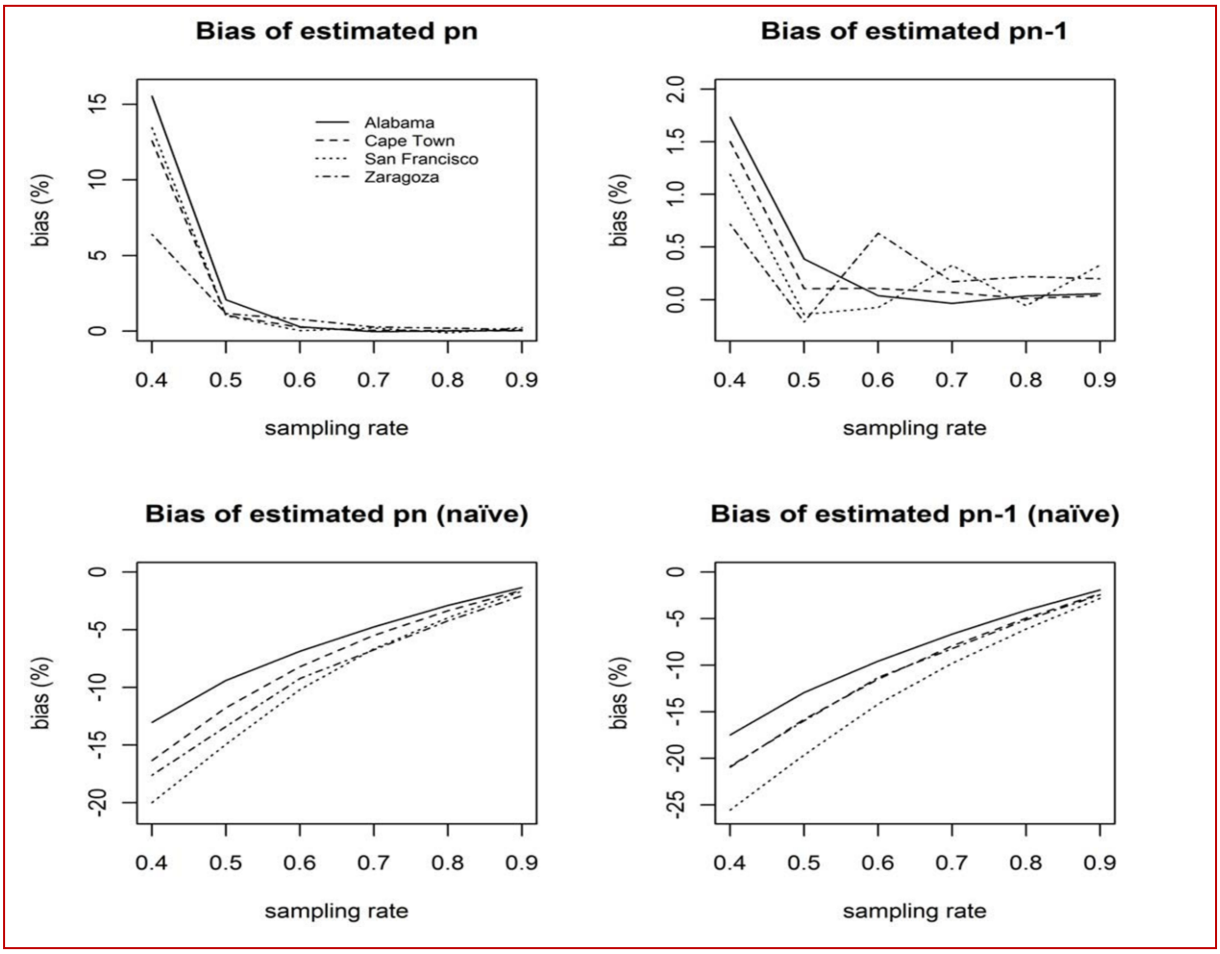

Figure 3. Bias in proportions using the naıve method, or the P matrix inversion method when the vector S is truncated at size 6. Simulated populations are used from which typical samples approximate the published data.