Trans R Soc Trop Med Hyg 2021; **115**: 208–210
doi:10.1093/trstmh/trab020 Advance Access publication 15 February 2021

COMMENTARY

Rethinking neglected tropical disease prevalence survey design and analysis: a geospatial paradigm

Peter J. Diggle [a,b,*], Benjamin Amoah[a], Claudio Fronterre [a], Emanuele Giorgi[a], and Olatunji Johnson[a]

[a]*Medical School, Lancaster University, Lancaster, LA1 4YF, UK;* [b]*Health Data Research, 215 Euston Road, London, NW1 2BE, UK*

*Corresponding author: Tel: +44 796 382 9999; E-mail: p.diggle@lancaster.ac.uk

Received 9 December 2020; editorial decision 13 January 2021; accepted 3 February 2021

Current methods for the design and analysis of neglected tropical disease prevalence surveys largely rely on classical survey sampling ideas that treat prevalence data from different locations as an independent random sample from the probability distribution induced by a random sampling design. We set out an alternative, explicitly geospatial paradigm that can deliver much more precise estimates of the geospatial variation in prevalence over a country or region of interest. We describe the advantages of this approach under three headings: streamlining, whereby more precise results can be obtained with smaller sample sizes; integrating, whereby a joint analysis of data from two or more diseases can bring further gains in precision; and adapting, whereby the choice of future sampling location is informed by past data.

Keywords: elimination surveys, geospatial methods, predictive inference, prevalence mapping

Introduction

The conventional approach to the design and analysis of neglected tropical disease (NTD) prevalence surveys is to select a representative set of communities in the area of interest and apply a diagnostic test to a sample of individuals in each selected community. The resulting data are then analysed by an agreed, context-dependent protocol. To provide a specific focus, we consider the transmission assessment survey (TAS) protocol for establishing whether an evaluation unit (EU) has achieved elimination of lymphatic filariasis (LF). Under the TAS protocol, for each EU the total number of sampled individuals who test positive for LF is compared with a tabulated cut-off value depending on the numbers of communities and individuals sampled, and elimination or non-elimination is declared accordingly.[1]

This approach suffers from several shortcomings. First, it delivers an unqualified yes/no answer rather than stating how likely it is that EU-level prevalence exceeds any prespecified threshold. Second, by ignoring spatial context, it delivers an unnecessarily imprecise estimate of prevalence. Third, its 'one-disease-at-a-time' approach misses an opportunity to make better use of the limited resources by jointly collecting and analysing data on two or more coendemic diseases.

Here, we argue that model-based geostatistics (MBG)[2] offers a better approach to the design and analysis of georeferenced prevalence surveys.[3] In what follows, we expand on the advantages of MBG as a paradigm for the design and analysis of tropical disease prevalence surveys under three broad headings: streamlining, integrating and adapting.

Streamlining

Figure 1 shows a hypothetical prevalence surface along with a set of locations at which prevalence, y say, has been measured. A classical approach to predicting the area-wide average prevalence uses the sample mean, $\bar{y}$, the sample variance, s^2 and the sample size, n, to calculate an approximate 95% prediction interval, $\bar{y} \pm 1.96\sqrt{(s^2/n)}$. The result, in this example, is 41.3 ± 8.6. If the measurement locations are chosen at random, this result is valid, but also needlessly imprecise because it takes no account of the spatial context. Inspection of Figure 1 suggests that each measured value not only tells you what the value of the prevalence surface is at that location, it also gives you a good idea of the prevalence at nearby locations. In other words, the prevalence surface exhibits spatial correlation, with the correlation between a pair of measured values decreasing as the distance between their locations increases. MBG does not presume this behaviour, but estimates it from the data and exploits it; for our example, MBG produces a 95% prediction interval of 41.6 ± 1.4.

We have found comparable gains in precision in case studies of district-level prevalence of LF in Ghana[4] and of trachomatous trichiasis prevalence in Ethiopia.[5]

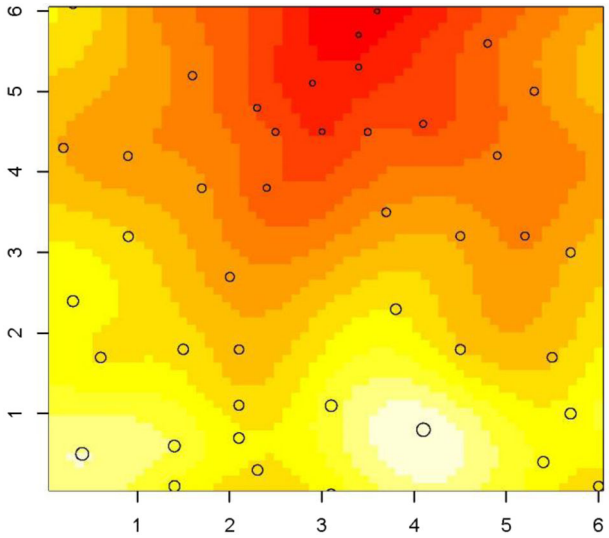

Figure 1. A hypothetical prevalence surface, colour-coded from red (low) through orange and yellow to white (high). Data locations are indicated by open circles whose radii are proportional to the true prevalence at each location.

Integrating

The key to the success of MBG is that measured values of prevalence are correlated with unmeasured values. The same applies to studies of two or more coendemic diseases, or of a single disease using multiple diagnostic instruments. If measurements of two different things are correlated, data on one will help to predict both.

Amoah et al.[6] developed geostatistical models for the joint distribution of two or more spatial prevalence surfaces and used these to map Loa loa parasitological prevalence using a combination of parasitological prevalence data and a lower-cost questionnaire-based alternative (RAPLOA).[7] The practical advantage of a joint analysis is that the parasitological method is more accurate, but the lower cost of the RAPLOA method allows its implementation at many more locations.

A joint MBG analysis of prevalence data on two coendemic diseases can also lead to more efficient mapping of both, especially if the measurements for the two diseases are not colocated. The gains in precision from a joint analysis are likely to be modest if diagnostic tests for two or more diseases are applied to the same individuals. However, the reduction in the associated costs of fieldwork may be substantial, enabling the collection of larger samples, and hence greater precision, for fixed cost.

Adapting

Inspection of Figure 1 shows that the measurement locations are not distributed randomly, but are spaced somewhat regularly over the area. The advantage of this is that the existence of spatial correlation implies that measurements at a pair of very close locations will give essentially the same information as a single measurement. Random sampling allows this wasteful circumstance to occur by chance, whereas a spatially regulated sample prevents it.

Now imagine that measurements are recorded sequentially and, at each stage, this information can be used to decide where to place the next measurement location. Chipeta et al.[8] developed a formal strategy for conducting an adaptive sampling strategy of this kind. They envisaged collecting and analysing data in batches, with each stage of the process informing the selection of the next batch of measurement locations.

Adaptive sampling is particularly advantageous when not all ranges of prevalence are of equal importance. For example, in assessing whether a district has or has not achieved elimination of a particular disease, once we have established that prevalence in a particular EU is unequivocally above or below the elimination threshold, there is no need to take additional samples to predict its exact value. Kabaghe et al.[9] use adaptive sampling to find malaria hotspots in rural Malawi.

Adaptive sampling is not always practicable. However, one simple form of adaptive sampling that is both practical and statistically advantageous is in postelimination monitoring for recrudescence, where more intense sampling might well be conducted in areas of high historical prevalence. The same logic applies to the design and analysis of a sequence of prevalence surveys to assess the progress of a mass drug administration (MDA) programme, where survey designs appropriate to the early stages of a programme can be adapted in response to spatially heterogeneous changes in prevalence over consecutive rounds of MDA.

We emphasise that sampling from a set of locations, among which the probability of inclusion varies in a known manner, leads to correct inferences provided the analysis protocol respects the sampling design. By contrast, sampling subjectively at locations that are believed more likely to find positive cases and analysing the resulting data as if they had been randomised introduces bias.[10]

Obstacles to implementation

MBG methods will only be adopted widely if they can be implemented robustly by in-country teams. MBG has become well established through a process of peer review in the statistical and epidemiological literature. Case studies have demonstrated its usefulness in the context of tropical disease prevalence mapping. The methodology is implemented in open source software, namely the `R` package `PrevMap` (R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/).[11] We are developing a user-friendly, dashboard-style interface to allow statistical novices to run a limited set of MBG analyses with minimal supervision. We have run 2-d training courses in several countries aimed at researchers with some knowledge of statistics, and are developing a version of this training that is accessible to other staff engaged in practical public health decision-making.

Our vision is for a triple-layer ecosystem, in which trained, in-country champions can both advise statistically unqualified colleagues on a day-to-day basis and be advised by ourselves when they encounter problems that cannot be solved using prepackaged methods. This reflects our general philosophy of a symbiotic relationship between statistical theory and practice, whereby real-life public problems motivate the development of

novel methodology that is then available for future applications in an ever-improving virtuous cycle.

Authors' contributions: Diggle conceived the project and wrote the first draft. All authors contributed to the development of the underlying statistical methodology and commented on the first draft.

Funding: This work was supported through funding of the NTD Modelling Consortium by the Bill and Melinda Gates Foundation (OPP1184344) 'Moving towards elimination'.

Competing interests: None.

Ethical approval: Not applicable.

Data availability: Not applicable.

References

1 World Health Organization. Monitoring and Epidemiological Assessment of Mass Drug Administration in Global Programme to Eliminate Lymphatic Filariasis: A Manual for National Elimination Programmes. Geneva, Switzerland: World Health Organization; 2011.

2 Diggle PJ, Moyeed RA, Tawn JA. Model-based geostatistics (with Discussion). Appl Statist. 1998;47:299–350.

3 Diggle PJ, Giorgi E. Model-based Geostatistics: Methods and Applications in Global Public Health. Boca Raton, FL: CRC Press; 2019.

4 Fronterre C, Amoah B, Giorgi E, et al. Design and analysis of elimination surveys for neglected tropical diseases. J Inf Dis. 2020;221(Supplement_5):S554–60.

5 Amoah B, Fronterre C, Johnson O, et al. Geostatistical analysis can yield more precise estimates of neglected tropical diseases in elimination settings: an example of trachoma in Ethiopia. International journal of epidemiology. 2021; under review.

6 Amoah B, Giorgi E, Diggle PJ. A geostatistical framework for combining spatially referenced disease prevalence data from multiple diagnostics. Biometrics. 2020;76:158–70.

7 Takougang I, Meremikwu M, Wanji S, et al. Rapid assessment method for prevalence and intensity of *L. loa* infection. Bull World Health Org. 2002;80:852–8.

8 Chipeta M, Terlouw DJ, Phiri K, et al. Adaptive geostatistical design and analysis for sequential prevalence surveys. Spat Statist. 2016;15:70–84.

9 Kabaghe A, Chipeta MG, McCann RS, et al. Adaptive geostatistical sampling enables efficient identification of malaria hotspots in repeated cross-sectional surveys in rural Malawi. PLoS ONE. 2016;12:e0172266.

10 Diggle PJ, Menezes R, Su TL. Geostatistical analysis under preferential sampling (with Discussion). Appl Statist. 2010;59:191–232.

11 Giorgi E, Diggle PJ. PrevMap: an R package for prevalence mapping. J Statist Soft. 2017;78:1–29.