**Protocol S2. Model-based geostatistical framework for predicting G6PDd prevalence maps**

This protocol provides information about the geostatistical model used to map the prevalence of G6PDd. The general principles of model-based geostatistics have been previously described by Diggle and Ribeiro [[1](#_ENREF_1)]. The general framework of the model used here has been fully described by Piel et al. in application to HbS mapping [[2](#_ENREF_2)], though specific adaptations to that model were made to reflect the G6PD gene’s sex-linked inheritance mechanism.

**S2.1 Model requirements in relation to G6PD genetics**

The G6PD gene is carried on the X-chromosome, meaning that males inherit only a single copy of the gene (“hemizygotes”). In contrast, females have two copies so may carry two wild-type alleles or two deficient alleles (“homozygotes”), or a combination of one wild-type and one deficient (“heterozygotes”). The relative frequencies of each of these genotypes can be predicted if populations are assumed to be in Hardy-Weinberg equilibrium [[3](#_ENREF_3),[4](#_ENREF_4)], where represents the frequency of deficient alleles, making the allele frequency of normal G6PD expression. The overall population allele frequency corresponds to the frequency of male hemizygotes. As a consequence, the frequency of female homozygous deficiency expression corresponds to , and in heterozygotes. However, for reasons discussed in Protocol S5, the heterozygous female genotype is often not expressed as phenotypically deficient, so only a proportion of genetically heterozygote females are identified in surveys and reported as deficient cases. There is therefore a genotype-phenotype disjunction between frequencies of observed deficient females and expected deficient females based upon Hardy-Weinberg derivations from frequencies of deficiency in males.

The model requirements were therefore to incorporate prevalence data relating to both males and females; and from this input dataset to generate continuous frequency estimates across malaria endemic countries (MECs; Protocol S1) with quantified uncertainty measures for rates of deficiency in (i) males, (ii) homozygous females, and (iii) all females – this final category being the combination of all homozygotes and the proportion of heterozygotes expected to be diagnosed as phenotypically deficient.

**S2.2 The model**

In this section, we describe our Bayesian spatial model for the G6PDd allele frequency surface . takes as its argument an arbitrary location on the Earth’s surface. The posterior induces a posterior for the homozygous frequency surface . We computed summaries of and , such as the mean and the variance ,at each location , to produce the maps relating to G6PDd allele frequency. Predictions of expected genetic heterozygotes were adjusted to reflect the observation that many genetic heterozygotes are not phenotypically deficient. A deviance term, , was determined by the model from the input data to represent the predicted deviation from the expected rates of genetic heterozygotes (see Likelihood). National and regional level prevalence of G6PDd in males and females was then computed from the model posterior distributions, as described in Protocol S4.

Prior

We model as a nonlinear transformation of a Gaussian random field [[5](#_ENREF_5),[6](#_ENREF_6)] , plus a random field that associates an independent normally distributed value with each location on the earth’s surface. Specifically,

The link function maps the random variable , which can be any real number, to the interval (0,1), so can be used as a probability or prevalence. We used a non-standard link function, which is described below.

The prior for is parameterized by the constant mean function , and the standard exponential covariance function with amplitude parameterand range parameter . The distance function gave the great-circle distance between and , unless was on a different side of the Atlantic ocean to , in which case it returned . This modification prevented data in Africa from unduly influencing the east coast of south America. Suitable priors were assigned to the scalar parameters *,* and :

GP indicates a Gaussian process. The units of *,*  and are earth radii, and and are unitless. The unstructured component is modelled as normally distributed with unknown variance :

The distribution parameters for , the deviance term from expected Hardy-Weinberg deficiency rates in heterozygous females were defined by and , where:

Likelihood

Separate likelihoods were defined for males and females, in accordance with their different inheritance mechanism of the G6PD gene.

*In males*: The frequency of G6PDd in males corresponds to the population allele frequency of deficiency, . If male individuals are sampled at the th observation location , the probability distribution for the number of copies of the G6PDd allele that will be found is binomial, with probability :

*In females*: Under the assumptions of Hardy-Weinberg equilibrium [[3](#_ENREF_3),[4](#_ENREF_4)], if female individuals are sampled at the th observation location (for a total of chromosomes), the probability distribution for the number of G6PDd females is binomial, with probability :

Flexible link function and empirical Bayesian analysis

The link function for binomial data is usually taken to be the inverse logit function:

Piel et al. [[7](#_ENREF_7)] employed this model. Applying the change of variables formula, the induced prior for is:

Note that this is essentially a two-parameter family of probability distributions, since and appear only in the sum.

Under this model, in areas where data points are highly clustered the best fitting values of and result in long right-hand tails for the predictive distribution of prevalence in the next observation at , thus skewing the summary statistics to give implausibly high predicted median values. To overcome this problem, we used an alternative flexible link function:

We were unable to infer jointly with the other model parameters in a fully Bayesian manner due to poor MCMC mixing, so we adopted an empirical fitting approach inspired by data pre-processing steps employed in classical geostatistics. This improved the fitting of the model to the data and is described in more detail below.

Empirical Bayesian approach to fitting the polynomial coefficients

For each observation of males tested and deficient males identified, we first obtained the posterior expectation of the gene pool-wide prevalence of G6PDd with uniform prior density on [0,1]:

We discarded values for which was below 25. Then, we inferred the parameters and of the non-spatial Bayesian model:

We then plotted the posterior predictive cumulative distribution function (CDF) of against its empirical CDF, and fitted the coefficients *Ȼ* of the cubic polynomial function to the points using least squares, subject to the constraint that must be invertible (or, equivalently, monotone).

The polynomial coefficients for such a function are specific to the dataset. The set of coefficients used, corresponding to an invertible function and fitting the empirical CDF was:

In the Bayesian analysis of the full spatial model, the fitted values of Ȼ were taken as known and fixed. Although this empirical procedure is admittedly informal, the resulting non-standard link function did substantially improve the fit of the model to the data.

Prior predictive constraint

The G6PDd database was largely composed of opportunistic community surveys. As such, data points were unevenly distributed across the malaria endemic region (Figure 2.1A), and the model had to be adapted to this. Inclusion of the highly clustered datasets, particularly in the Philippines, did not appear to alter predictions across the rest of the map. In contrast, high and low latitude areas which were sparse in data (such as northern China and Argentina) were very hard for the model to predict, generating implausibly high population estimates of G6PDd prevalence. Based on our knowledge of the distribution of G6PDd from existing maps [[8-10](#_ENREF_8)], it seemed reasonable to assume that G6PDd was at low prevalence in these areas, supporting the principle that most surveys will have been conducted where G6PDd was expected to be found. To introduce this into the model, we decided to constrain , and in such a way that the prior predictive distribution of , before the data are incorporated, puts probability mass of 1×10-4 or less on values in excess of 0.0001. In other words, we constrained 99.99% of the prior predictive probability mass to be between allele frequencies of 0% and 0.01%.This constraint arguably induces a lack of fit by forcing to depart from its prior mean by many standard deviations in areas where G6PDd allele frequency is known to be high; but it does remedy the implausibly high predictive values in the data-sparse edges of the map, and does not seem to adversely affect the fit in areas of non-zero allele frequency. A detailed explanation of this process is given by Piel et al. [[2](#_ENREF_2)].

**S2.3 Model implementation**

As previously described [[11](#_ENREF_11)], implementation of the modelling and mapping procedure was divided into two computational tasks: (i) the Bayesian inference stage was implemented using the Markov chain Monte Carlo (MCMC) algorithm [[12](#_ENREF_12),[13](#_ENREF_13)] and was used to generate samples from the posterior predictive distribution (PPD) of the parameter set and the spatial random fields at the data locations; and (ii) a prediction stage in which samples were generated from the PPD of G6PDd frequencies at each pixel on the 5×5 km grid, as described below in Protocol 2.5.

The scalar parameters ,,and were updated jointly using Haario, Saksman and Tamminen’s adaptive Metropolis algorithm [[14](#_ENREF_14)], as implemented by PyMC’s AdaptiveMetropolis step method. Each value of *ε(oi)* at observation location was updated separately using the standard one-at-a-time Metropolis algorithm. The distribution of the Gaussian random field at the observation locations, , is conjugate to the distribution of, so we updated by sampling from its full conditional distribution. MCMC output parameter values are summarised in Table S2.1.

Convergence of the MCMC tracefile was judged by visual inspection; one million MCMC iterations were run, with 10% recorded in the output tracefile, the first 50,000 iterations of which were excluded from the mapping stages. During the mapping process, the posterior distributions were thinned by 100, resulting in 500 mapping iterations. MCMC dynamic traces are available on request.

The model code was written in Python programming language (<http://www.python.org>), and is freely available from the MAP’s code repository (<https://github.com/malaria-atlas-project> and <https://github.com/RosalindH/g6pd>). The MCMC algorithm was used from the open-source Bayesian analysis package PyMC [[15](#_ENREF_15)] (<http://code.google.com/p/pymc>).

**S2.4 Overview of mapping procedure**

The output algorithm from the MCMC was used to generate PPDs for each model output at all pixels in the MEC 5×5 km grids. The PPDs can be used to determine the most probable prevalence estimate at each pixel [[5](#_ENREF_5)], these were summarised for each pixel as median values, together with the PPD’s interquartile range (IQR). The median was determined to be more representative than the mean value, due to the PPD’s right-hand skew, inflating the mean values of the predictions. Maps were generated using Python and Fortran code, available from the MAP’s code repository (<http://github.com/malaria-atlas-project/generic-mbg>) and subsequently displayed in GIS software (ArcMap 10.0, ESRI Inc., Redlands, CA, USA).

**S2.5 Uncertainty**

Uncertainty metrics.

The MCMC and derived map predictions are directly informed by the evidence-base of surveys. However, there is usually some level of disparity within this database (i.e. in clustered areas) and between the database and the map predictions. The relative influence of each point’s influence on the model predictions is moderated according to the data points’ sample size (through a binomial model previously described [[11](#_ENREF_11)]), their spatial distribution and the level of heterogeneity in the frequencies identified between neighbouring data points. These three factors influence the model’s ability to predict frequencies, which are then reflected in the spread of the PPD. The precision of the PPD is an indicator of the representativeness of the summary statistic used (in this case median values) [[5](#_ENREF_5)] and is quantified here by the interquartile range (IQR) of the PPD, thus corresponding to a 50% probability class. However, because the values of the IQR are affected by the underlying G6PDd prevalence levels, with IQR generally increasing with prediction values. The IQR was therefore also standardised against the median map to produce an uncertainty index less affected by the underlying prevalence levels and more illustrative of relative model performance driven by data densities in different locations (). Maps of both uncertainty metrics are presented for comparison (Figure S2.1).

Uncertainty maps.

Greatest absolute variation in the predictions (Figure S2.1B) was found from areas of relatively high predicted frequencies and low input data availability (Figure S2.1A), specifically southern Pakistan, the central Sahel region across Chad and Sudan, and southern central Africa (Democratic Republic of Congo (DRC), Zambia, Malawi, southern Tanzania and Mozambique) and Madagascar. Model predictions for these areas should be interpreted with caution; additional surveys from these areas would enable more robust predictions.

Adjusting the IQR values to the predicted median frequency values gives a representation of the relative variability in the PPD (Figure S2.2B). The greatest relative prediction uncertainty is in peripheral transition regions, such as the Horn of Africa and southern Africa – where prevalence drops relative to surrounding areas. The high IQR region across DRC and neighbouring areas is not so pronounced in the proportional map. Greater uncertainty is shown across the Americas, previously masked in the raw IQR map (Figure S2.2A). The large proportion of northern China which is uninformed by data has very high relative uncertainty, reflecting an uncertainty in the model about where exactly to bring down the prevalence predictions to the very low levels predicted across the north of the country. An equivalent increase in uncertainty and associated data absence is seen on Borneo.

These two uncertainty metrics complement each other, allowing interpretation of confidence in the map predictions and can be used to identify areas where additional data would be most informative towards improving our understanding of G6PDd prevalence.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **Symbol** | **Mean** | **Median** | **St. Dev.** | **IQR** | **95% BCI** |
| Nugget variance |  | 0.239 | 0.226 | 0.080 | 0.037 | 0.190 |
| Amplitude (or partial sill) |  | 3.422 | 3.426 | 0.202 | 0.213 | 0.919 |
| Scale (or range) |  | 0.498 | 0.499 | 0.003 | 0.002 | 0.007 |

**Table S2.1. MCMC output parameter values**. Summary values of the model parameters, as fully described in Protocol S2. Summary statistics of the MCMC output include mean, median, standard deviation (St. Dev.), interquartile range (IQR) and the 95% Bayesian credible interval (95% BCI). Scale is measured in units of Earth radii; other parameters are unitless.

 **Figure S2.1. Prediction uncertainty metrics.** Panel A shows the input surveys database. Panel B is the interquartile-range (IQR) of the PPD of male G6PDd prevalence, representing absolute values of map uncertainty. The map in Panel C was derived directly from the IQR map, adjusted to the median values.

**References**

1. Diggle PJ, Ribeiro PJ, Jr. (2007) Model-based Geostatistics: Springer.

2. Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, et al. Global estimates of sickle haemoglobin in newborns. The Lancet: In press.

3. Hardy GH (1908) Mendelian Proportions in a Mixed Population. Science 28: 49-50.

4. Weinberg W (1908) Über den nachweis der vererbung beim menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg 64: 368-382.

5. Patil AP, Gething PW, Piel FB, Hay SI (2011) Bayesian geostatistics in health cartography: the perspective of malaria. Trends Parasitol 27: 246-253.

6. Banerjee S, Carlin BP, Gelfand AE (2004) Hierachical modeling and analysis for spatial data. Monographs on Statistics and Applied Probability 101. Boca Randon, Florida, U.S.A.: Chapman & Hall / CRC Press LLC.

7. Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, et al. (2010) Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. Nat Commun 1: 104.

8. Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The History and Geography of Human Genes. Princeton, New Jersey: Princeton University Press. 1088 p.

9. WHO Working Group (1989) Glucose-6-phosphate dehydrogenase deficiency. Bull World Health Organ 67: 601-611.

10. Nkhoma ET, Poole C, Vannappagari V, Hall SA, Beutler E (2009) The global prevalence of glucose-6-phosphate dehydrogenase deficiency: a systematic review and meta-analysis. Blood Cells Mol Dis 42: 267-278.

11. Hay SI, Guerra CA, Gething PW, Patil AP, Tatem AJ, et al. (2009) A world malaria map: *Plasmodium falciparum* endemicity in 2007. PLoS Med 6: e1000048.

12. Gilks WR, Richardson S, Spiegelhalter D (1995) Markov chain Monte Carlo in Practice: Chapmand & Hall/CRC.

13. Gelman A, Carlin JB, Stern HS (2003) Bayesian Data Analysis. Texts in Statistical Science. Boca Raton, Florida, U.S.A.: Chapman & Hall / CRC Press LLC.

14. Haario H, Saksman E, Tamminen J (2001) An Adaptive Metropolis Algorithm. Bernoulli 7: 223-242.

15. Patil A, Huard D, Fonnesbeck CJ (2010) PyMC: Bayesian Stochastic Modelling in Python. J Stat Softw 35: 1-81.