PREDICTING MALARIA INFECTION IN GAMBIAN CHILDREN FROM SATELLITE DATA AND BED NET USE SURVEYS: THE IMPORTANCE OF SPATIAL CORRELATION IN THE INTERPRETATION OF RESULTS

MADELEINE C. THOMSON, STEPHEN J. CONNOR, UMBERTO D’ALESSANDRO, BARRY ROWLINGSON, PETER DIGGLE, MARK CRESSWELL, AND BRIAN GREENWOOD
MALSAT Research Group, Liverpool School of Tropical Medicine, Liverpool, United Kingdom; Department of Parasitology, Prince Leopold Institute of Tropical Medicine, Antwerp, Belgium; Medical Statistics Unit, Lancaster University, Lancaster, United Kingdom; London School of Hygiene and Tropical Medicine, London, United Kingdom

Abstract. In line with the renewed World Health Organization Global Malaria Control Strategy, we have advocated the use of satellite imagery by control services to provide environmental information for malaria stratification, monitoring, and early warning. To achieve this operationally, appropriate methodologies must be developed for integrating environmental and epidemiologic data into models that can be used by decision-makers for improved resource allocation. Using methodologies developed for the Famine Early Warning Systems and spatial statistics, we show a significant association between age-related malaria infection in Gambian children and the amount of seasonal environmental greenness as measured using the normalized difference vegetation index derived from satellite data. The resulting model is used to predict changes in malaria prevalence rates in children resulting from different bed net control scenarios.

Malaria is an environmental disease. Anopheles mosquitoes transmit the causative agent, Plasmodium spp., when the environmental parameters (such as water availability, temperature, and humidity) permit. For example, in many parts of the world where temperature is not a limiting factor, malaria transmission is highly seasonal, with peak transmission following the period of peak rainfall.

Understanding how malaria varies in the community as a result of seasonal or year-to-year changes in environmental factors is important for the planning of national malaria control programs since it may allow interventions to be adapted to specific sites or times of year. This is essential for the effective control of the disease. Traditional methods used to describe the level of malaria endemicity in a particular area include calculating the percentage of children with an enlarged spleen or with parasitemia. Assessing the relationship between environmental parameters and such malarialographic indices in a quantitative manner is fraught with difficulties since the prevalence of malaria may vary considerably within a small area, and the data collected at a limited number of points are not necessarily applicable to a broader region.

Furthermore, prevalence data collected during a limited period of time cannot describe the seasonal variations that occur even in areas of high endemicity. There are also important logistical and financial constraints to the organization of widespread, often repeated, malaria parasite prevalence surveys.

Consequently, the assessment of malaria endemicity in a particularly locality must be extrapolated from a limited number of surveys that are constrained in space (often to a few villages) and in time (perhaps once at the end of the rainy season for one or two years).

Recent developments in the use of geographic information systems (GISs) for malaria risk mapping have resulted in an ambitious plan to map malaria endemicity throughout Africa using age-related prevalence rates. The Mapping Malaria Risk in Africa (Atlas du Risk de Malaria) initiative is using both epidemiologic and entomologic data from all available sources to determine the different epidemiologic situations occurring throughout Africa and their respective malaria risk. For areas where such information is unavailable, environmental parameters, such as rainfall and temperature, have been used to predict transmission levels.

Large international spatial databases of environmental factors in digital format are increasingly available and provide information relevant to predicting the distribution of disease vectors. Such databases include both archived and near real-time environmental data obtained from meteorologic satellites.

Meteorologic satellites, such as the polar orbiting and geostationary satellites operated by the National Ocean and Atmospheric Administration of the United States (NOAA series) and the European Organization for the Exploration of Meteorological Satellites (Meteosat series), observe large areas of the earth’s surface (for Meteosat this includes the entire African continent) with an optimal spatial resolution of 1.1 km and 5 km, respectively. Their high temporal resolution (twice a day and 30 min, respectively), spectral characteristics (which include both visible and thermal channels), and high contrast capability means that they are ideally suited to global and repetitive studies of the environment. Consequently, they are used widely for monitoring seasonal and inter-annual changes in environmental phenomena such as rainfall, temperature, plant phenology, and biomass. Environmental proxies are derived empirically by comparing land-based phenomena (e.g., meteorologic data or crop growth characteristics) with post-processed satellite data.

Widely used environmental proxies include rainfall estimates from cold cloud duration, vegetation indices (such as the normalized difference vegetation index [NDVI]) and surface temperatures (including sea surface temperatures [SST] and land surface temperatures [LST]). Recent work has indicated the potential use of Meteosat thermal infrared data for estimating ambient air temperatures.

Numerous studies have been undertaken using NOAA-advanced very high resolution radiometer (AVHRR) data to predict the distribution and abundance of a range of veterinary and medical disease vectors. Examples include Culicoides imicola (African horse sickness virus), Phlebotomus papatasii (cutaneous leishmaniasis), Phlebotomus orientalis (visceral leishmaniasis), and Ixodes scapularis (Lyme disease).
The potential for using meteorologic satellite data in the study of malaria transmission dynamics has been reviewed in an earlier paper. Using epidemiologic and entomologic data from The Gambia, we have shown that even simple analysis of proxy ecologic variables derived from satellite data can indicate variation in environmental factors affecting malaria transmission indices. In particular, we have highlighted the potential for using information on the seasonal changes in vegetation growth and senescence as indicators of the length and intensity of the malaria transmission season since these processes are closely allied to rainfall and humidity.

This has been supported in analysis of temporal changes in seasonal vegetation status estimated using the NDVI and seasonal variation in the number of parasitologically diagnosed malaria cases in both The Gambia and Kenya albeit with a lag period of one month. Further studies have shown that the NDVI is associated with the occurrence of different proportions of the Mopti karyotypes in populations of An. gambiae in Mali, which are known to vary according to different ecologic settings. Based on our preliminary results and further work (Thomson MC, Connor SJ, unpublished data), we have advocated the use of satellite imagery in providing environmental information in near real time for malaria stratification, monitoring, and early warning. For this potential to be realized, appropriate methodologies must be developed for integrating environmental and community-based epidemiologic data into models that can be used by decision-makers for improved resource allocation.

The study described in this paper has addressed this issue using methodologies developed by the United Nations Food and Agricultural Organization Global Information and Early Warning System. It provides for the first time statistical evidence of the association between age-related prevalence rates in the community and the seasonal amount of environmental greenness measured using the NDVI.

Statistical modeling of the relationship between disease prevalence and environmental data is complicated by spatial relationships, which typically result in positive correlations between observations from spatially close sampling units. These correlations in turn negate the basic assumptions underlying standard linear or generalized linear regression analysis. Failure to allow for spatial correlation typically leads to spuriously small standard errors of regression parameter estimates, and corresponding over-statement of the significance of regression effects. However, few studies relating vector distribution to satellite data have addressed the problem of spatial correlation in the data sets. A rare exception is a study undertaken by Kitron and others, who undertook a spatial analysis of tsetse fly distribution in the Lambwe Valley of Kenya using Landsat Thematic Mapper (TM) imagery and a GIS. They found that using multiple regression analysis, they could explain 87% of the variance in fly density using several TM bands. However, when they applied spatial filtering using Moran’s I measure of spatial correlation, a significant positive link was observed among all trap data. They therefore concluded that the positive correlation between spectral data and fly abundance was largely due to determinants not included in the analysis.

In our study, we have used an adaptation of a method of Liang and Zeger to make inferences that account for spatial correlation. Results from the model developed demonstrate the relevance of satellite data to studies of malaria transmission and indicate how remote sensing, GIS, and spatial analysis can be used to indicate village-based variation in the effectiveness of an intervention. We have used the model to predict changes in prevalence rates in children according to different levels of bed net usage.

**MATERIALS AND METHODS**

**Study area.** The Gambia is a small country situated on the west coast of Africa. It extends eastwards from the Atlantic Ocean on either side of the Gambia River. The geographic determinants of malaria transmission in the country have been described in detail elsewhere. Briefly, the country consists of flat, woodland savannah with swamps bordering the river, which is saline to a distance of approximately 150 km from the coast. The climate is typical Sahelian (Sudano-Sahel) with a short rainy season that lasts from June–July to October followed by a long dry season covering the remaining months. Minimum and maximum mean monthly temperatures range between approximately 26°C and 33°C with highest temperatures in the eastern part of the country. Therefore, there are no temperature restrictions on malaria transmission in The Gambia, but transmission is restricted largely to the rainy season (rainfall for 12 meteorologic stations in 1992 ranged from 400 to 850 mm) when temporary breeding sites and an environment suitable for adult mosquito survival is created.

Considerable differences in malaria endemicity have been found within The Gambia (i.e., village-based parasite prevalence rates ranging from 1% to 89% in children 1–4 years of age) and these have been related to ecologic differences affecting the vector (species and population density) and the human population (use of bed nets).

**Epidemiologic data.** Data were collected during the evaluation of the Gambian National Impregnated Bednet Program, whose objective was to treat with insecticide (permethrin) all bed nets found in all villages covered by the National Primary Health Care (PHC) Program over a period of 2–3 years. The results of this evaluation have been reported elsewhere. Prior to the study, the aim of the program was explained to communities involved and verbal consent was obtained from the mothers of children included in the survey. This study was reviewed and approved by the Scientific Coordinating Committee of the Medical Research Council Laboratories, The Gambia and the Ethical Committee the Gambian Government.

A cross-sectional malaria morbidity survey of 2,276 children (1–4 years of age) from 65 villages from 5 ecologically different areas of The Gambia was carried out at the end of the 1992 rainy season (Figure 1). Each child was given a clinical examination and their age and weight/height was recorded. Details from their health card were taken if a health card was available. Finally, a blood sample was collected by fingerprick for thick and thin blood films for de-
FIGURE 1. A, distribution of the 65 study villages (dots) in The Gambia in 5 areas superimposed on the spatial distribution of the 10 seasonal classes of the normalized difference vegetation index (NDVI). The colors correspond to seasonal profiles in Figure 1B. Yellow predominates as one moves northeastwards towards the Sahel while dark green predominates as one moves southwestwards. B, Seasonal profiles of the NDVI obtained using ADDAPIX analytical software. Ten classes were chosen. Classes 1 and 2 are blue to indicate the fact that their low values are associated with the presence of water. Dark green represents the longest seasonal profile while pale yellow indicates a short vegetation cycle. The NDVI measure (called NDVLS) used as a proxy for length of the malaria transmission season = the area under the curve for the period first dekad of May until 3 dekads prior to the morbidity survey. A lag of one month between NDVI values and malaria cases has been shown in an early study on Gambian data. This value was extracted for each village coordinate. A second variable, the quadratic value of NDVLS (NDVLS^2), was computed and included in the analysis.

termination of malaria parasitemia. Approximately two-thirds of the children came from PHC villages, half of which had their bed nets treated with permethrin in 1992, a few months before the survey. Entomologic studies were also undertaken during the same period in three of the five study areas.

have been processed to produce a global archive of NDVI images at a temporal resolution of one dekad (10 days) and a spatial resolution of approximately 1.1 km. These data are held by the National Aeronautics and Space Administration (NASA) PATHFINDER 1 km project and can be accessed over the Internet (http://edcwww.cr.usga.gov/landdaac/1KM/comp10d.html and Eindenshink J, Faunderen J, 1997. The 1-km avhrr global land data set: first stages in implementation, http://edcwww.cr.usga.gov/landdaac/1KM/paper.html).

Thirty-six dekad images from May 1992 to April 1993 of The Gambia were extracted from the PATHFINDER active 1.1 km AVHRR archives. The images, in Goodes Homologine Projection, were processed using the Image Display and Analysis Software/Map and Image Display and Analysis Software (IDA/WINDISP)(http://ag.arizona.edu/replfman/windisp3.html). One image was incomplete and was therefore recreated using the mean of the previous and subsequent images. The images were then entered into a spatial statistical analysis program, ADDAPIX, that is designed for monitoring seasonal and interannual vegetation growth using satellite imagery (Grigualo S, University of Venice, Venice, Italy). This program uses the IDA image format as the basic input file and submits the series of images to a principle component analysis. This is followed by a non-hierarchical clustering procedure with the aim of grouping pixels (image squares) that have a similar spatial and temporal pattern. After a period of experiment, 10 classes were chosen as a number large enough to indicate the ecological variation within The Gambia but small enough for a number of classes to contain sufficient villages for comparison of malaria endemicity (Figure 1). The seasonal profiles from the 10 classes indicate clearly that there is considerable difference in the length and greenness of the vegetation cycle in The Gambia despite its relatively small size and uncomplicated topography.

We were able to extract the associated NDVI class for each village by overlaying the distribution of the study villages as a boundary file in IDA format. The NDVI measure (called NDVLS) used as a proxy for length of the malaria transmission season was the area under the curve for the period of the first dekad of May until 3 dekads (one month) prior to the morbidity survey. A lag of one month between NDVI values and malaria cases has been shown in an early study on Gambian data. A second variable, the quadratic value of NDVLS (NDVLS^2), was computed and included in the analysis. A number of children were excluded from the current analysis because either their age was not known, their net usage was not known, or no vegetation information was obtained from their village environment due to persistent clouds. After these exclusions, 2,039 children were included in the analysis. Children were classified into three groups: 1) children who slept without a net (591, 29%); 2) children who slept under an untreated net (886, 43.5%); and 3) children who slept under an insecticide-treated net (562, 27.5%). A total of 728 (35.7%) children were found to be positive for Plasmodium falciparum parasites during the morbidity survey (Table 1).

Spatial analysis. Modeling the relationship between environmental data and malarial indices is complicated by the spatial correlation between observations, which invalidates the inferences associated with standard regression calculations. In particular, while standard regression modeling applied to spatially correlated data gives satisfactory point estimates of regression parameters, the nominal standard errors associated with these parameter estimates are typically too small, and this in turn leads to an exaggerated impression of the significance of regression relationships. We therefore undertook the following analysis.

First, we used a logistic regression model to estimate the probability of the presence of malarial parasites in each child as a function of NDVLS, NDVLS^2, and age (in days), and adjusted for the effects of nets (treated, untreated, or absent), PHC (yes or no), (health card) (yes or no), and area (a five-level factor) identifying the five ecologically different areas. Second, we computed the variogram of the standardized residuals from the fitted logistic regression model to estimate the spatial correlation in the data. Based on the appearance of the variogram, we assumed that the correlation between a pair of measurements was an exponentially declining function of distance, i.e., \( p(d) = \exp(-\alpha d) \) for some positive value of \( \alpha \).

Third, from the fitted exponential correlation model, we adjusted the nominal standard errors of the logistic regression parameter estimates to allow for the effects of spatial correlation, using the method of generalized estimating equations. This allows us to reassess the nominal significance of particular terms in the regression model, and to simplify the model accordingly.

Having fitted the model, we can use it to compute the estimated probability of infection for any combination of values of the explanatory variables, and so predict the prevalence rates for children in any particular village with a given set of characteristics. We can then repeat this process under any chosen set of circumstances to quantify the effect on prevalence rates of any change in behavior, e.g., increased use of mosquito nets.

RESULTS

The results of the non-spatial logistic regression analyses of the presence/absence of malaria in each child (non-spatial model) are given in Table 2. Age (in days) and NDVLS^2 are both positively correlated with the presence of parasites (nominal \( P < 0.001 \)), while NDVLS, the use of nets (treated or untreated), living in a PHC village, and possessing a health card are all negatively correlated with the presence of parasites (nominal \( P < 0.001 \)). As noted earlier, quoted nom-

<table>
<thead>
<tr>
<th>Type of village</th>
<th>% Children with health card</th>
<th>% Children with Plasmodium falciparum infection</th>
<th>No. of children in survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Non-PHC</td>
<td>0.69</td>
<td>0.47</td>
<td>218</td>
</tr>
<tr>
<td>A PHC</td>
<td>0.6</td>
<td>0.49</td>
<td>373</td>
</tr>
<tr>
<td>B Non-PHC</td>
<td>0.77</td>
<td>0.38</td>
<td>425</td>
</tr>
<tr>
<td>B PHC</td>
<td>0.78</td>
<td>0.29</td>
<td>461</td>
</tr>
<tr>
<td>C PHC</td>
<td>0.73</td>
<td>0.27</td>
<td>562</td>
</tr>
</tbody>
</table>

* \( A = \) no net; \( B = \) untreated net. \( C = \) treated net. 
† PHC = primary health care.
inal $P$ values are likely to overstate the true significance of these results.

The results of the logistic regression analyses of the presence/absence of malaria in each child corrected for the spatial correlation in the data (spatial model) are also presented in Table 2. The second model has the same estimated regression parameters, but because the standard errors have been corrected for spatial correlation, the significance of the covariates has changed accordingly. In the spatial model, the $t$-value for age remains approximately the same as in the non-spatial model, indicating (as we would expect) that there is no spatial difference in age structure of 1–4-year-old children in The Gambia. The $t$-value for NDVLS and NDVLS$^2$ changed considerably from $-4.408$ to $-2.139$ and from $4.507$ to $2.169$, respectively, but both remained significant at the 5% level. This indicates that some but not all of the association observed between presence of parasites and length of transmission is due to spatial correlation. The $t$-values of the variables nets (untreated) and nets (treated) are also reduced in the spatial model when compared with the non-spatial model but remain highly significant ($P < 0.001$). Neither living in a PHC/non-PHC village nor possessing a health card remain significant ($P > 0.05$) once spatial effects are taken into account.

### Predicting prevalence rates under different scenarios.

Using the model presented in Table 2, we predicted the number of children infected with *P. falciparum* under differing net usage scenarios (Table 3). These scenarios included if no children slept under a net, if all children slept under a net, if all children slept under a treated net, and if the current net status prevailed but none were treated.

According to the model, removing nets from those children who currently use them would result in an increase of parasite prevalence from 36% to 45%. Providing net treatment without increasing net coverage would result in a slight decrease in parasite prevalence from 36% to 35%. A more significant decrease from 36% to 33% would result if net coverage were increased to include all children even when the nets were not treated. Not surprisingly, the best possible result would be obtained if all children slept under a treated net (prevalence rate decrease from 36% to 30%).

### DISCUSSION

In our study, when taking into account spatial correlation, the significance of all associations between malaria prevalence and human (age, untreated nets, treated nets, health card) and village (PHC and area) factors and environmental satellite data (NDVLS and NDVLS$^2$) are reduced with the exception of child age. However, the associations with bed net use (treated or untreated) and with the satellite data (NDVLS and NDVLS$^2$) are still significant at the 1% and 5% levels, respectively, even when spatial effects are taken into account. This suggests that these associations have a more direct link and may be determined by a biological mechanism. These results confirm the widely published non-spatial analysis of the effectiveness of insecticide-treated bed nets in reducing infection with malaria parasites. They also reaffirm the importance of traditional bed nets (not treated with insecticide) in reducing exposure of children to malaria. Indeed, the study indicates that in The Gambia, increasing the use of bed nets might be as effective in reducing the rate of infection among children as providing insecticide to treat them. However, insecticide-treated bed nets are thought to be more effective than untreated nets at preventing infections characterized by high-density parasitemia, and this might partially explain their significant impact on childhood mortality. Further studies are required to estab-
The association of the NDVI with a suitable environment (NDVI) in 1992 in Jahally, The Gambia. Dekadal NDVI values from Addapix class 8 represents the seasonal cycle of vegetation at Jahally, Area 3. Anopheles gambiae s.l. occur in large numbers in early June, presumably in association with the local irrigated rice fields, but no mosquitoes infected with sporozoites are found until the last dekad of June. Both vector populations and the number of infected mosquitoes decreases rapidly once the NDVI values start to decrease.

In conclusion, this study demonstrates the importance of spatial effects in the inferences that can be drawn from epidemiologic models created from multiple regression analyses, which have highlighted the differences in infection rates between children living in PHC and non-PHC villages. According to our analyses, such results represent spatial differences in the siting of PHC and non-PHC villages rather than intrinsic qualities associated with the delivery of primary health care.

The model also suggests that environmental satellite data (NDVI), which indicates the length of the growing season, might be useful for predicting the levels of malaria endemicity in children once behavioral factors are taken into account. It provides empirical support for the use of satellite data to provide first-level stratification of regions or countries for malaria control activities.

The association of the NDVI with a suitable environment (NDVI) indicates the length of the growing season, which is correlated with saturation deficit in The Gambia. The NDVI has been shown to be highly correlated with spatial effects in the inferences that can be drawn from epidemiologic models created from multiple regression analysis. It confirms the effectiveness of insecticide-treated bed nets in reducing parasite prevalence but suggests that in the Gambian situation, the use of traditional untreated nets play a significant role in reducing malaria exposure in children. The model indicates that earlier studies suggesting that living in a PHC village provides protection from malaria infection is the result of spatial correlation. The model also suggests that there may be an a priori relationship between the NDVI and malaria infection. We suggest that this relationship is likely to be based on the fact that high NDVI values are indicative of a moist environment supporting both mosquito breeding and adult survival.

Acknowledgments: Eric Pfriman is thanked for adaptations to WINDISP. The Banjul Department of Water Resources is thanked for providing the meteorologic data. Data used in this study include data produced through funding from the Earth Observing System Pathfinder Program of NASA's Mission to Planet Earth in cooperation with NOAA. The data were provided by the Earth Observing System Data and Information System (EOSDIS) Distributed Active Archive Center at the Goddard Space Flight Center, which archives, manages, and distributes this data set.

Financial support: This work was supported by the United Kingdom’s Government Department for International Development (United Kingdom).

Disclaimer: The Department for International Development is not responsible for the views expressed in this report.

Authors’ addresses: Madeleine C. Thomson, Stephen J. Connor, and Mark Cresswell, MALSAT Research Group, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, United Kingdom. Umberto D’ Alessandro, Department of Parasitology, Prince Leopold Institute of Tropical Medicine, Antwerp, Belgium. Barry Rowlingson and Peter Diggle, Medical Statistics Unit, Lancaster University, Lancaster LA1 4YE, United Kingdom. Brian Greenwood, London School of Hygiene and Tropical Medicine, London WC1 7HT, United Kingdom.

REFERENCES


