ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2015)

Mapping the spatial distribution of a disease-transmitting insect in the presence of surveillance error and missing data

Andrew E. Hong, Corentin M. Barbu, Dylan S. Small and Michael Z. Levy

University of Pennsylvania, Philadelphia, USA

and the Chagas Disease Working Group in Arequipa

Arequipa, Peru

[Received August 2013. Final revision May 2014]

Summary. Maps of the distribution of epidemiological data often ignore surveillance error or possible correlations between missing information and outcomes. We analyse presence–absence data at the household level (12050 points) of a disease-carrying insect in Mariano Melgar, Peru, collected as part of the Arequipan Ministry of Health's efforts to control Chagas disease. We construct a Bayesian hierarchical model to locate regions that are vulnerable to under-reporting due to surveillance error, accounting for variability in participation due to infestation status. The spatial correlation in the data allows us to identify relative inspector sensitivity and to elucidate the relationship between participation and infestation. We show that naive estimates of prevalence would be biased by surveillance error and missingness at random assumptions. We validate our results through simulations and observe how randomized inspector assignments may improve prevalence estimates. Our results suggests that bias due to imperfect observations and missingness at random can be assessed and corrected in prevalence estimates of spatially auto-correlated binary variables.

Keywords: Bayesian hierarchical modelling; Spatial analysis; Statistical epidemiology; Surveillance error

1. Introduction

The increasing risk of vector-borne disease epidemics has accompanied the rise of urban environments in the developing part of the world. The use of spatial analysis in public health campaigns for disease control is documented in dengue (Teixeira *et al*., 2002; Lars *et al*., 2009; Sanchez *et al*., 2010), malaria (Trape *et al*., 1992; Wang *et al*., 2005; Matthys *et al*., 2006) and Chagas disease (Corrasco *et al*., 2005). Chagas disease is a tropical parasitic disease, affecting millions in central and south America. The disease agent is *Trypanosoma cruzi*, which is a parasite that is transmitted by the *Triatoma infestans* (*T. infestans*) insect vector. The policy for Chagas disease control has focused on the elimination of this vector (Dias *et al*., 2002). Although initiatives to control *T. infestans* have been active for decades (Guhl, 2007), the insect is a continually re-emergent threat in Peru (Levy *et al*., 2006). Because of the strain on public resources that is created by these recurring epidemics and the risk of emergence of insecticide resistance due to repeated treatment (Germano *et al*., 2010; Lardeux *et al*., 2010), there is

Address for correspondence: Andrew E. Hong, Department of Statistics, Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104, USA.
E-mail: andrew.e.hong@gmail.com

0964–1998/15/178000

interest in applying statistical methods to guide the application of insecticide to urban areas (Levy *et al.*, 2010; Barbu *et al.*, 2011).

This study was done in co-ordination with the efforts of the Peruvian Ministry of Health to control an epidemic of *Trypanosoma cruzi* infections in the city of Arequipa, Peru (Levy *et al.*, 2006). Insecticide treatment was preceded by a household level survey, which identified household infestations of *T. infestans*. The results of the survey, which was conducted in the district of Mariano Melgar, are shown in Fig. 1. At the time of the survey, the policy was to prioritize the treatment of households in district localities where the rate of household infestations exceeds 10%. Ecological surveys for observing infestation are conducted by human inspectors and are subject to under-reporting of presence. Inspectors are heterogeneous, differing in their ability to identify infestations. Although previous work on triatomine infestation detection has used spatial techniques, acknowledging the imperfect inspection process, this work has not accounted for heterogeneity in the inspectors' skills (Barbu *et al.*, 2011). Another concern of policy makers in this survey is the large proportion of missing information (34% of the records). Correlation between missingness and the studied outcome has been shown to lead to serious bias in clinical trials, raising serious concerns of the validity of research findings (Little *et al.*, 2012). In the case of spatially auto-correlated outcomes, Bayesian hierarchical models have been used extensively to obtain point estimates at missing locations under the assumption that missingness is uncorrelated with the outcome (Le *et al.*, 1997; Faes *et al.*, 2011; Kang and Cressie, 2011).

Here, we propose to adjust the risk mapping of triatomine infestations for surveillance error,

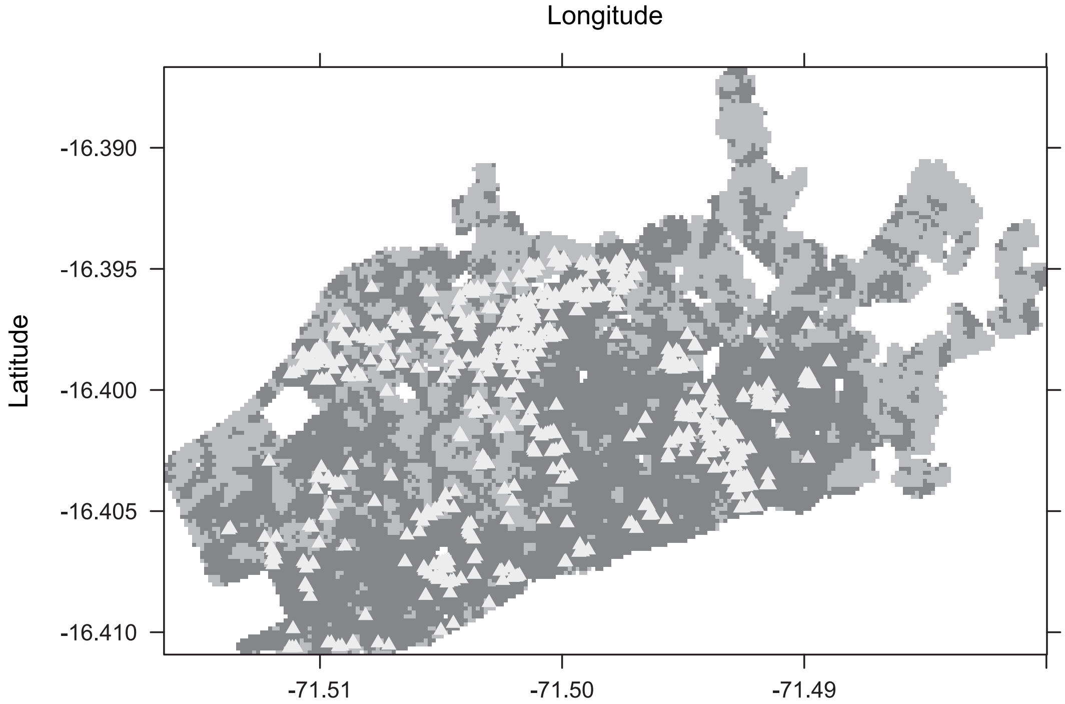


Fig. 1. Transect data from the 2011 *T. infestans* survey in Mariano Melgar, Arequipa, Peru, that were collected by the Ministry of Health (this area is 3838 m × 2664 m and contains 12050 total households; the survey identified 608 positive households and contains 4098 non-participating households): □, infested; ■, uninfested; ▨, 'not applicable'

caused by the lack of sensitivity on the part of inspectors and missingness not at random, caused by different inclination to participate in the surveys depending on the infestation status. Explicitly, we construct a Bayesian hierarchical model that jointly assesses

(a) the probability that households participate depending on their infestation status,
(b) the individual sensitivities of the inspectors and
(c) the prevalence of infestation accounting for (a) and (b).

We then discuss the findings of our model in Mariano Melgar that guided the Ministry of Health's 2011 campaign in Arequipa.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Inspection and data collection

The *T. infestans* survey was conducted by the Ministry of Health in Mariano Melgar, which is a district of 12050 households. We mapped the locations of these households by determining their relative position to city blocks and comparing field maps of these blocks with satellite images from Google Earth™ (`http://earth.google.com`). Inspectors requested the participation of residents before searching households for *T. infestans*. The outcome of

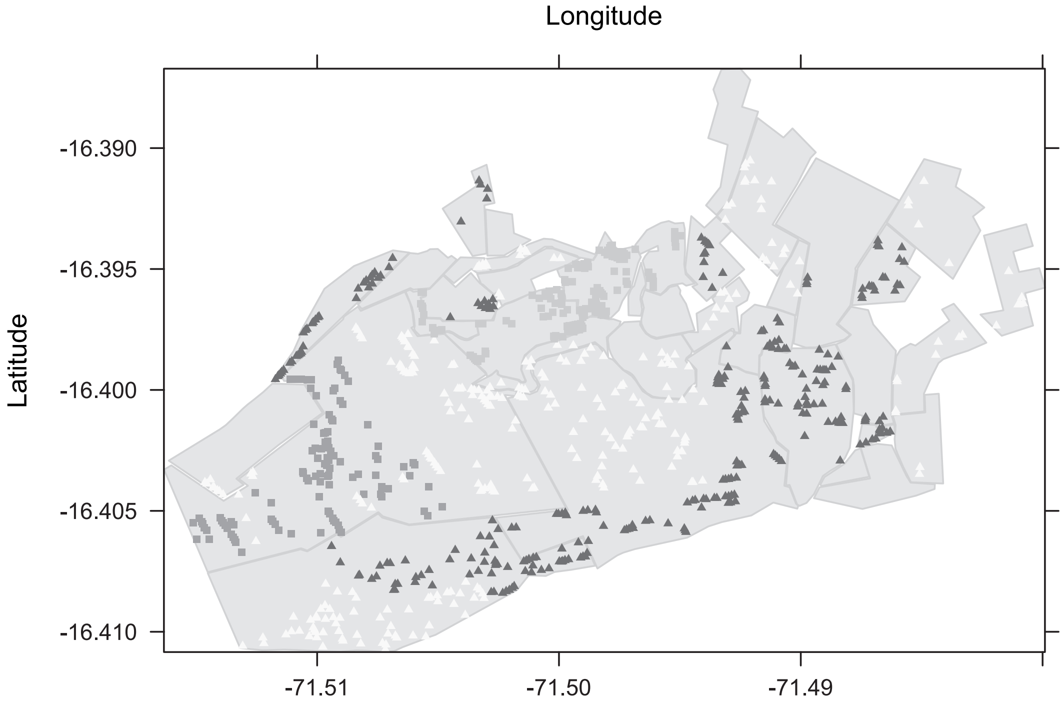


Fig. 2. Examples of the household assignments of four inspectors across the 35 localities of Mariano Melgar: the number and distribution of assigned households inspected varies by individual; the assignments of inspectors A (△) (301 total households) and B (▲) (291 total households) span multiple localities across the district, whereas the assignments of inspectors C (■) (107 total households) and D (□) (137 total households) are highly localized

each inspection was either the successful collection of insect samples, the failure to locate samples or non-participation of the household. Each entry of the data consists of a pair of co-ordinates denoting the location of the household, a presence–absence status and the identifier or labelling of the inspector. 4098 households (around 34%) opted not to participate in the survey. In this study, missingness appears to aggregate spatially in regions with lower rates of infestation and is therefore missingness not at random. We model the relationship between the true infestation and the point pattern of missingness to analyse this claim in Section 3.2.

Separate data identifying the sensitivity of the 40 inspectors who were involved in this study were unavailable. However, validated data from previous treatment campaigns suggested the general sensitivity of human inspectors, which informed our prior specification. We rely on knowledge of the co-ordinate locations of the households to infer the distribution of the *T. infestans* infestation and inspectors' assignments. The household assignments of four inspectors are shown in Fig. 2. During the survey, inspectors were assigned to households, on the basis of staffing constraints, which resulted in subgroups of inspectors inspecting an entire region. Because of these aggregated assignments, the surveillance error in this study is spatially correlated. This confounding between infestation distribution and inspector sensitivity through geographic location potentially biases the estimation of the presence of insects. For a geographic region, it becomes difficult to disassociate the severity of the infestation apart from the sensitivity of the inspectors. For this reason, we study in Section 5 how prevalence estimates may be subject to confounding by inspectors' assignment.

3. Model specification

The primary interest in this study is to infer the true household infestation status, which we modelled as a binary outcome: infested or uninfested. In the context of our application, this quantity is not directly observable through the observation of inspectors, who are subject to surveillance error. We model the true binary infestation status by using a generalized linear model with the probit link function. A complete diagram of the hierarchical components is shown in Fig. 3. Each of the 12050 households is included in the model and indexed by i. The ith household is given by its easting and northing co-ordinates by using the universal transverse Mercator map projection. For household i, the variable y_i is the true infestation status, where $\{y_i = 1\}$ indicates that household i is infested. We model the probability that the ith household is infested by

$$\Phi^{-1}\{\mathbb{P}(y_i = 1 | u_i, t)\} = u_i + t \tag{1}$$

where Φ is the cumulative distribution function of the standard Gaussian distribution, u_i is a continuous, household level effect, capturing how at risk the ith household is for being infested, and t is an intercept term for the entire district. The covariance structure for $u = [u_i]_{i=1}^n$ among the households is defined by their geographic locations and will ensure that the infestation statuses are spatially correlated. The usual convention is to place a diffuse $N(0, \sigma^2)$ prior on t. Similar approaches for spatial modelling of categorical data may be found in Banerjee *et al.* (2003, 2004) for other public health settings.

3.1. Spatial effect

We used a conditionally auto-regressive Gaussian model for u to capture the spatial similarity of the infestation. The model for u, which was popularized by Besag *et al.* (1991), is a centred Gaussian distribution with a precision matrix $\mathbf{\Lambda}$. The entries of the model precision matrix are

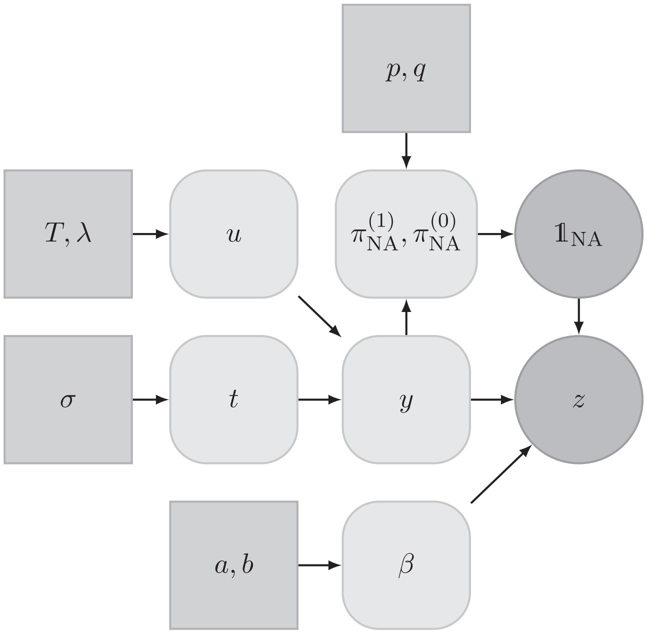

p, q
T, λ
u
$\pi_{\mathrm{NA}}^{(1)}, \pi_{\mathrm{NA}}^{(0)}$
$\mathbb{1}_{\mathrm{NA}}$
σ
t
y
z
a, b
β

Fig. 3. Model diagram: the variables z, the reported infestation status, and $\mathbb{1}_{\mathrm{NA}}$, the non-response indicator, are the observed data; z is generated by y, the true infestation status, and β, the inspector sensitivity; $\mathbb{1}_{\mathrm{NA}}$ is generated by y, the true infestation status, and $(\pi_{\mathrm{NA}}^{(0)}, \pi_{\mathrm{NA}}^{(1)})$, the probability of participation depending on the true infestation status; the main parameter of interest is the infestation status of the households, y, which is spatially correlated through the random field u

based on the pairwise Euclidean distance between households $d_{i,j}$, a scaling parameter k_u and a threshold T:

$$\mathbf{\Lambda}_{i,j} = \begin{cases} k_u \sum\limits_{\{k: d_{i,k} < T\}} d_{i,k}^{-1}, & \text{if } i = j, \\ -k_u d_{i,j}^{-1} \mathbb{1}_{d_{i,j} \leqslant T}, & \text{if } i \neq j. \end{cases} \tag{2}$$

The effect of the scaling and threshold parameters is most evident on the marginal distributions of u_i conditional on the rest of the households u_{-i} (Rue and Held, 2005):

$$u_i | u_{-i} \sim N \left(\frac{\sum\limits_{\{j: d_{i,j} < T\}} d_{i,j}^{-1} u_j}{\sum\limits_{\{j: d_{i,j} < T\}} d_{i,j}^{-1}}, \frac{1}{k_u \sum\limits_{\{j: d_{i,j} < T\}} d_{i,j}^{-1}} \right). \tag{3}$$

For the conditional marginal, $u_i | u_{-i}$ is centred at a weighted sum of neighbouring values within the threshold radius. Households whose distance to i exceeds T have no effect on the conditional distribution of u_i. Within the threshold radius T, households that are closer to i are given more weight proportionally to their inverse distance. The scaling parameter k_u determines the variation of u_i around this centre.

Previously, we had analysed a neighbouring district in Arequipa and determined that the correlation in infestation statuses is negligible for households that are separated by 50 m or more (Barbu *et al.*, 2013). We fix our threshold T at 50 m. For the prior on k_u we follow the usual practice of placing a conjugate $\exp(\lambda)$ prior on k_u (Paciorek, 2007).

3.2. Missing data

We treat the missingness point pattern, which we denote as $\mathbb{1}_{\text{NA}}$, where 'NA' denotes 'not applicable', as a series of Bernoulli outcomes, depending on the true infestation status of the household. The probability of household participation is modelled separately for infested $\pi_{\text{NA}}^{(1)}$ and uninfested households $\pi_{\text{NA}}^{(0)}$ allowing for differential participation according to the infestation. Further, because of the marked differences of socio-economic status between localities, we allow this relationship to vary across localities:

$$\pi_{\text{NA}\,j}^{(1)} = \mathbb{P}(\mathbb{1}_{\text{NA}i} = 0 | y_i = 1), \tag{4}$$

$$\pi_{\text{NA}\,j}^{(0)} = \mathbb{P}(\mathbb{1}_{\text{NA}i} = 0 | y_i = 0) \tag{5}$$

where the household indexed by i is in locality j. We used identical beta distribution priors $B(p, q)$ for both parameters in each of the localities.

3.3. Surveillance process

To account for the human error in surveillance, we model the sensitivity of each inspector as the probability $\beta \in [0, 1]$ that an inspector locates *T. infestans* in the household, when the insect is present. Then, the reported outcome infested or uninfested follows a Bernoulli distribution in observed households. If $\beta_{j(i)}$ is the sensitivity of the inspector who inspected household i, the distribution of the reported outcome is

$$\mathbb{P}(z_i = 1 | y_i, \beta_{j(i)}, \mathbb{1}_{\text{NA}i}) = \beta_{j(i)} y_i, \qquad \text{if } \mathbb{1}_{\text{NA}i} = 0. \tag{6}$$

The sensitivity $\beta_{j(i)}$ is therefore relevant only if the true infestation status of the household, y_i, is positive. This surveillance model is the individual inspector sensitivity model, where each inspector in the study has his or her own sensitivity parameter β_j. In contrast with this individual inspector model, a simpler model is the group inspector model, where the sensitivity is identically β for all the households. We use a beta distribution $B(a, b)$ as the prior for the sensitivity parameters either for each inspector separately or for the common sensitivity of all the inspectors in the simpler model.

4. Results for the 2011 Mariano Melgar survey

Infestation by *T. infestans* in Mariano Melgar, Arequipa, was strongly clustered in space. On the basis of the raw surveys (the proportion of infested among the surveyed households), only four localities fitted the 10% prevalence criterion for inclusion in blanket insecticide treatment, which is the uniform application of insecticide to all households in the locality. Our role in this study was to apply the model that was proposed in Section 3 to adjust these estimates and possibly to identify additional localities at risk for major *T. infestans* infestations.

4.1. Full model results

To perform the analysis on the Mariano Melgar survey, we placed priors on the following parameters: the intercept of the linear model, t; the precision parameter of the Gaussian spatial effect, k_u; the inspector sensitivities $\{\beta_j\}_j$. For the first two parameters, we used a diffuse Gaussian prior $N(0, \sigma^{-2} = 1 \times 10^{-8})$ and a diffuse exponential prior $\exp(\lambda = 1 \times 10^{-4})$. Our estimates of the presence–absence values are dependent on the specification of the inspector sensitivity priors. The sensitivity of our analysis to this prior is displayed in Table 1. Cross-sectional pretreatment and post-treatment data from previous spraying campaigns showed that human inspectors are

Table 1. Pairwise correlations between estimated inspector rankings in the Mariano Melgar survey against various prior specifications for surveillance error†

Prior	*Correlations between estimated rankings for the following priors:*			
	B(6.5, 2)	*B(5,5)*	*B(1,1)*	*B($\frac{1}{2}$,$\frac{1}{2}$)*
$B(6.5,2)$		0.9600	0.9392	0.7356
$B(5,5)$	0.9600		0.9765	0.7921
$B(1,1)$	0.9392	0.9765		0.7471
$B(\frac{1}{2},\frac{1}{2})$	0.7356	0.7921	0.7471	

†Different prior specifications for the sensitivity parameters produce different estimates of each inspector's sensitivity when analysing the Mariano Melgar survey. However, when ranking inspectors on the basis of their estimated sensitivity by using these priors, we found that the rankings were fairly consistent

accurate roughly 76% of the time. After consulting local experts, we agreed on the use of a beta prior $B(6.5,2)$ for inspectors' sensitivities. We used a relatively weak beta prior $B(7,3)$ for the missingness probabilities to reflect the overall rate of missing households in the data (around 0.34) and found that the choice of prior had little influence on the estimates for the Mariano Melgar data.

We implemented the model by using a Gibbs sampler, which is outlined in Appendix A, and estimated the posterior probability of infestation, $\hat{\mathbb{P}}(y_i=1|z)$, for each of the households. Averaging these household level estimates by locality, we produced the localitywide infestation estimates that are shown in Table 2. With our informative prior for inspector sensitivity, we found that the locality estimates for two additional localities, 11 and 37, exceeded the 10% mark. We mapped the infestation estimates at the block level in Fig. 4 to provide guidance on insecticide application in areas where blanket treatment (insecticide application to all participating households) is not warranted.

By introducing dependence between the missingness point pattern and the infestation, the data are not missing at random. For the Mariano Melgar survey, we allowed this dependence to vary across localities. This model inferred that survey participation rates were higher for infested households compared with uninfested households, consistently across localities. Across almost all localities, we found that the estimated participation rates, displayed in Table 3, were higher for infested households (except for locality 9). Because the model linked underparticipation to lower rates of infestation, it is likely that estimates of prevalence made under the missingness at random assumption would overestimate the infestation.

Fig. 5 displays the posterior distributions of the least sensitive, the 10th, 20th, 30th and most sensitive inspectors, ranked by posterior mean. Although these posterior distributions vary from the prior, the group average of all the inspectors' posterior means was 0.75601, which was close to the prior mean of 0.7647. Similarly, when using $B(\frac{1}{2},\frac{1}{2})$ and $B(1,1)$ priors of mean 0.5 we found that the group averages were 0.5663 and 0.5507 respectively. Although the overall levels of estimated infestation were sensitive to the prior, we found that the rankings of inspectors remained consistent across prior specifications; see Table 1. This consistency suggests that information is present in the data to identify the relative sensitivity of inspectors.

Table 2. Estimated prevalence of *T. infestans* in localities of the district of Mariano Melgar†

	Number of units	*Number of infestations*	*Number of NAs*	*Proportion of infected participants*	*Infestation estimate (missingness at random)*	*Infestation estimate (missingness not at random)*
1	294	0	173	0	0.0041	0.0025
2	2605	142	998	0.0884	0.1092	0.0821
3	271	5	125	0.0342	0.0755	0.0559
4	170	0	60	0	0.0056	0.0051
5	82	0	49	0	0.0133	0.0032
6	73	0	45	0	0.0101	0.0043
7	82	0	51	0	0.0994	0.0402
8	37	0	12	0	0.0103	0.0076
9	108	16	1	0.1495	0.2016	0.1916
10	147	17	77	0.2429	0.3077	0.2163
11	132	8	38	0.0851	0.1200	0.1104
12	113	18	56	0.3158	0.3784	0.2880
13	604	145	50	0.2617	0.3548	0.3378
14	147	0	60	0	0.0147	0.0154
15	273	16	78	0.0821	0.1027	0.0885
16	134	3	57	0.0390	0.0339	0.0310
17	113	0	53	0	0.0260	0.0134
18	180	0	101	0	0.0023	0.0009
19	169	0	100	0	0.0005	0.0005
21	374	0	203	0	0.0003	0.0003
22	108	0	64	0	0.0014	0.0019
23	176	1	59	0.0085	0.0198	0.0217
24	225	1	101	0.0081	0.0125	0.0089
25	166	0	83	0	0.0038	0.0008
26	226	0	147	0	0.0055	0.0012
28	143	0	82	0	0.0072	0.0000
30	213	0	122	0	0.0018	0.0010
31	106	0	62	0	0.0040	0.0022
32	33	0	12	0	0.0031	0.0016
33	82	0	57	0	0.0074	0.0010
34	50	0	22	0	0.0000	0.0000
35	160	0	87	0	0.0087	0.0026
36	534	20	96	0.0457	0.0572	0.0569
37	2022	137	373	0.0831	0.1086	0.1000
38	1698	79	344	0.0583	0.0770	0.0706

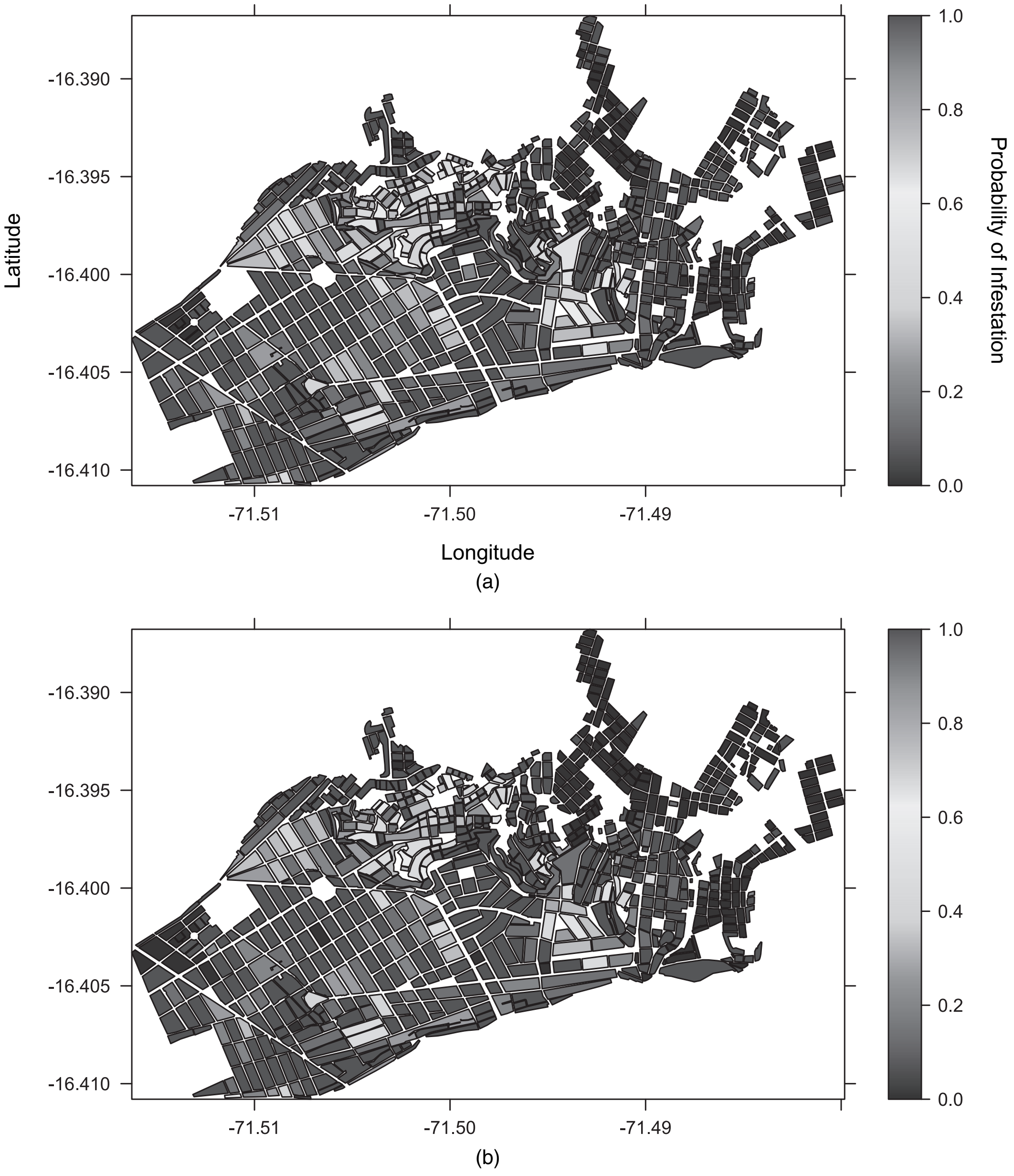
†Displayed are the total number of households, the number of positively identified households for infestations, the number of non-participating households and the average probability of infestation by locality. We used these estimates to identify two additional at-risk localities: 11 and 37 that were later treated with insecticide in the spring of 2011.

4.2. Model comparison

We compared the Bernoulli inspection model in Section 3.3 with a linear regression model. We model the reported infestation outcome (binary) z as

$$\Phi^{-1}\{\mathbb{P}(z_i = 1 | u_i, \beta_{j(i)}, \mathbb{1}_{\mathrm{NA}i}, t)\} = u_i + \beta_{j(i)} \mathbb{1}_{\mathrm{NA}i} + t \tag{7}$$

where i denotes the location and j denotes the inspector. Because the outcome is usually taken as observed unambiguously, we make the comparison between the two approaches on the basis of their infestation estimate for the missing households in the study. We treat the non-participating households identically to the inspected households in the study, except for the fact that these



(a)

(b)

Fig. 4. Maps of the estimated prevalence of infestation across the city blocks of Mariano Melgar, Arequipa, Peru, before insecticide treatment in 2011: (a) estimated infestation prevalence of each household averaged by city block, when survey participation was assumed to be missing at random; (b) estimates produced by our model under the assumption that data were missing not at random

households have a fixed inspector effect equal to 0 (the effect of an average inspector with an effect equal to the prior mean). We used the same models as described in Section 3 for the spatial effect x and the intercept t. In contrast with Section 3.3, the inspector effect β is a continuous unbounded variable. In practice, we placed diffuse, centred Gaussian priors on these parameters, unlike the beta priors that were used for the Bernoulli model.

Table 3. Estimated participation rates in entomological surveys among residents of infested and uninfested households by locality in the district of Mariano Melgar†

Locality	π_{NA} *for infested units*	π_{NA} *for uninfested units*
1	0.7903	0.4267
2	0.8978	0.5934
3	0.7681	0.5364
4	0.7814	0.6510
5	0.7945	0.4509
6	0.7827	0.4252
7	0.6753	0.4404
8	0.7917	0.7075
9	0.9255	0.9716
10	0.7953	0.4339
11	0.8355	0.6959
12	0.7522	0.4578
13	0.9692	0.8832
14	0.7250	0.6029
15	0.8675	0.7000
16	0.8288	0.5841
17	0.7588	0.5552
18	0.8047	0.4538
19	0.7962	0.4269
21	0.8060	0.4692
22	0.7856	0.4389
23	0.7271	0.6790
24	0.7889	0.5621
25	0.7864	0.5188
26	0.7953	0.3686
28	0.7952	0.4562
30	0.8085	0.4446
31	0.7875	0.4427
32	0.7994	0.6623
33	0.8025	0.3591
34	0.7955	0.6110
35	0.7840	0.4783
36	0.8546	0.8170
37	0.9409	0.8015
38	0.9001	0.7872

†By introducing dependence between the missingness point pattern and the infestation, we move to a more realistic setting, where data are no longer missing at random. For the Mariano Melgar survey, we allowed this dependence to vary across localities. This model inferred that the participation rates for the survey were higher for infested households compared with uninfested households, consistently across localities.

For the Mariano Melgar study, the estimates for the non-participating households by using this linear regression model were similar to the estimates that were produced by our model. The estimates are particularly similar between linear regression and our model when households are assumed to participate at random, demonstrated in Table 4. This similarity is expected as non-participating households are treated as locations to be interpolated and not used to inform the fit of the model. However, the linear regression approach does not estimate the probability of false negative households depending on the inspector.

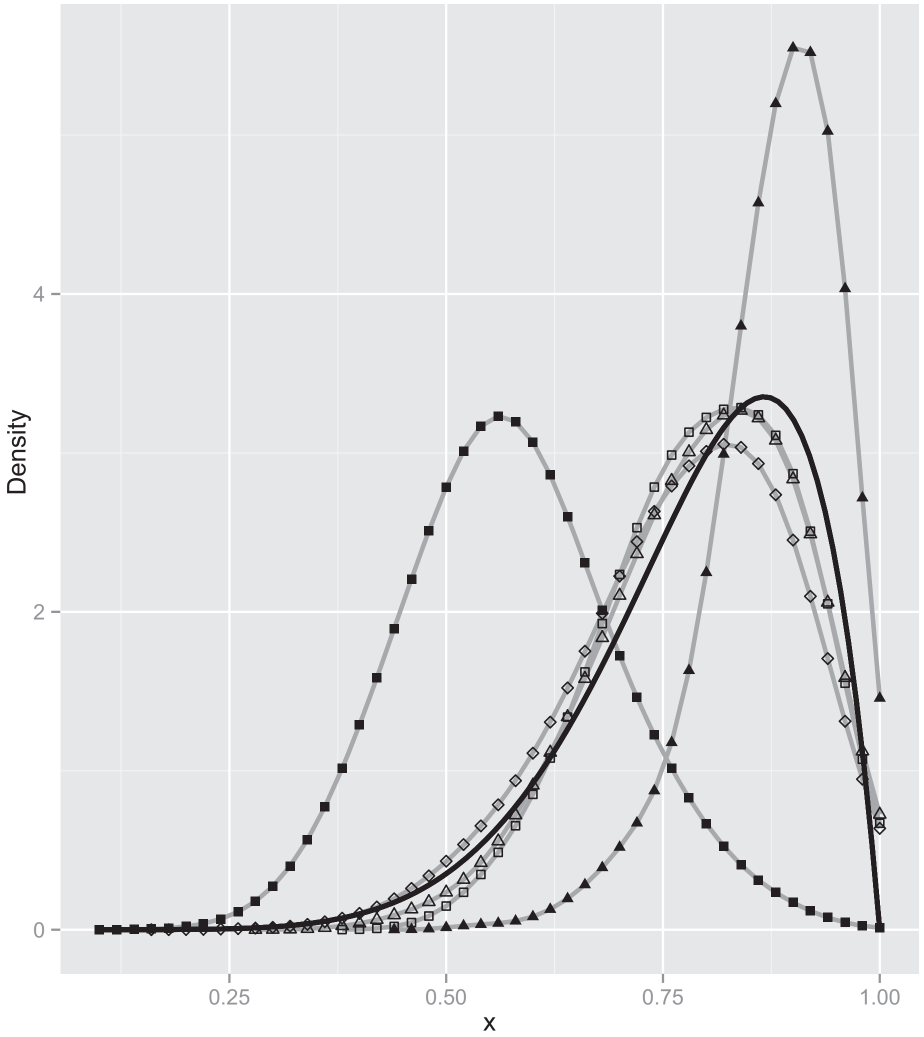


Fig. 5. Posterior distributions of the inspector sensitivity parameters β: displayed here are the distributions of the least sensitive (■), the 10th (□), 20th (◇), 30th (△) and the most sensitive (▲) inspectors from the Mariano Melgar survey, ranked according to their mean sensitivity; these posteriors deviate from the $B(6.5, 2)$ prior for inspector sensitivities (———) that was used for the analysis

5. Simulation study: household assignments and confounding

Confounding, in the context of this study, is the correlation across households between the risk of infestation and the surveillance error of the surveying inspector. Such confounding may potentially bias estimates of inspector sensitivity and infestation in our model. Confounding may be avoided if inspectors were randomly assigned to locations. However, randomized assignment is difficult to implement in practice as it is costly for each inspector to inspect geographically dispersed locations. Inspectors in Mariano Melgar were not randomly assigned to locations; instead, the locations that were inspected by an inspector tended to be geographically aggregated; see Fig. 2. Barring careful assignment of inspectors, it is difficult to ascertain whether or to what degree a particular study is affected by confounding. We conduct a simulation study to

Table 4. Comparison of the infestation estimates for missing households between the Bernoulli inspection error model and the linear regression model†

Locality	*Estimate, regression*	*Estimate, missingness at random*	*Estimate, missingness not at random*
1	0.0023	0.0055	0.0018
2	0.1020	0.0995	0.0223
3	0.0986	0.0800	0.0358
4	0.0171	0.0109	0.0101
5	0.0219	0.0164	0.0019
6	0.0205	0.0125	0.0043
7	0.1228	0.1428	0.0542
8	0.0217	0.0125	0.0117
9	0.0500	0.0300	0.1200
10	0.2510	0.2810	0.0930
11	0.0945	0.1019	0.0553
12	0.3388	0.3476	0.1638
13	0.2224	0.2880	0.0962
14	0.0334	0.0267	0.0226
15	0.0829	0.0816	0.0294
16	0.0055	0.0132	0.0055
17	0.0461	0.0412	0.0174
18	0.0003	0.0035	0.0008
19	0.0021	0.0008	0.0005
21	0.0024	0.0005	0.0004
22	0.0066	0.0021	0.0018
23	0.0345	0.0297	0.0319
24	0.0141	0.0139	0.0061
25	0.0032	0.0058	0.0008
26	0.0026	0.0076	0.0011
28	0.0025	0.0107	0.0000
30	0.0013	0.0026	0.0010
31	0.0076	0.0062	0.0028
32	0.0025	0.0025	0.0000
33	0.0041	0.0086	0.0004
34	0.0000	0.0000	0.0000
35	0.0137	0.0148	0.0037
36	0.0540	0.0503	0.0422
37	0.0869	0.0869	0.0294
38	0.0806	0.0687	0.0325

†The table compares the results produced by our model with a linear regression model, where inspectors are treated as fixed regression effects. We average the household estimates for the *non-participating households* by locality. The linear regression estimates were quite similar to the estimates produced by the Bernoulli model, when data are assumed to be missing at random.

understand how the inspector assignment in Mariano Melgar affects the estimation of the infestation and inspector sensitivities by comparison with estimation when inspectors are randomly assigned.

5.1. Methodology

40 inspectors collected the data in the Mariano Melgar survey, each assigned to a set of households in the survey. To create a randomized assignment, we simulate data by using the same number of inspectors and totals of households inspected by each inspector, but we select uni-

formly at random the locations of each inspector's assigned households. The interest of these simulations is the relationship between the spatial distribution of inspectors and our ability to infer the infestation accurately. For these simulations, for simplicity we do not simulate the missing data point pattern by using the model that was described in Section 3.2.

We simulate a mock infestation by using the posterior for true presence y from the survey data. To simplify the effect of prior specification, we simulate all inspector sensitivities from a common $B(6.5, 2)$ prior. We then generate two distinct data sets for the observed infestation by using the Mariano Melgar survey assignment and the randomized assignment. Finally, we estimate for each data set the household level infestation probability and the inspector sensitivity by using the Gibbs sampler (see Appendix A).

5.2. Factors influencing estimation performance

In addition to the effect of the inspector assignments, we are also interested in the effect of inspector sensitivity priors and inspector sensitivity models on the performance of our estimation. As a gold standard, usage of the $B(6.5, 2)$ generating prior should result in the lowest estimation error. We contrast the usage of this prior with the usage of the uniform $B(1, 1)$ prior and the centred $B(5, 5)$ prior. Results by using the $B(1, 1)$ prior demonstrate how effectively the inspector sensitivities may be learned from the data. Usage of the $B(5, 5)$ prior provides insight into the sensitivity of our model to a strongly misspecified prior.

5.3. Estimators and measures

Each round of the simulation produces a simulated observed infestation status z_{Sim}, given the inspector assignment. For each simulated data set, we then estimate the infestation and inspector accuracies for every combination of prior and model factors previously detailed.

The estimate for the infestation is the posterior probability of infestation for each household. The measure of accuracy for the infestation estimates is the squared distance between the simulated infestation data y_{Sim} and their estimate $\hat{y}$. As the former is a binary outcome and the latter is a probability forecast, the root-squared distance $\|y_{\text{Sim}} - \hat{y}\|_2$ is the Brier score (Brier, 1950). For the inspector sensitivities, the estimate is the posterior expectation $\hat{\beta}$. We again use the squared distance between this estimate and the generating parameter, $\|\hat{\beta} - \beta_{\text{Sim}}\|_2$, to measure the accuracy of our inspector sensitivity estimates.

5.4. Results of the simulation studies

The estimations from the 50-run simulation study are summarized for infestation in Table 5 and inspector sensitivity in Table 6. For every choice of prior, we attained better estimates of the infestation when inspectors were randomly assigned to households compared with when inspectors followed the Mariano Melgar survey assignment and missingness was ignored. However, the differences in Brier score between assignment type were significant but not extreme. In the worst case, when a $B(5, 5)$ prior was placed on the inspector sensitivities, we found that the mean Brier score was only 1.3% larger for the Mariano Melgar assignment compared with randomized assignments. The randomized assignment of inspectors is well guarded against confounding, where the assignment of inspectors to households is highly spatially correlated, but the performance of our model under the survey assignments was not significantly worse. These results demonstrate that the inspector assignment that was used to conduct the Mariano Melgar survey is not prone to excessive confounding error and lend credence to our findings in Section 4.

Table 5. Simulation mean (and standard deviations in parentheses) of Brier scores measuring the effect of factors on infestation estimation†

Prior	*Brier scores mean for infestation estimates* ($\|y_{\mathrm{Sim}} - \hat{y}\|_2$)			
	Survey assignment		*Randomized assignment*	
$B(1,1)$	21.2609	(0.2835)	21.0009	(0.2850)
$B(5,5)$	22.3966	(0.2542)	22.1084	(0.2473)
$B(6.5,2)$	20.9187	(0.2308)	20.7710	(0.3225)

†Sample size: 50. The mean and standard deviation across simulations of the Brier scores, measuring how accurately we could estimate the infestation, are shown. Each cell (pair) represents a different combination of factors, where lower Brier scores are indicative of better estimates of the infestation under these factors.

Table 6. Simulation mean (and standard deviations in parentheses) of squared norms measuring the effect of factors on the accuracy of sensitivity estimation

Prior	*Squared norm mean of inspector sensitivity estimates* ($\|\beta_{\mathrm{Sim}} - \hat{\beta}\|_2$)			
	Survey assignment		*Randomized assignment*	
$B(1,1)$	1.4122	(0.0806)	1.2454	(0.1289)
$B(5,5)$	1.8430	(0.0411)	1.7918	(0.0486)
$B(6.5,2)$	0.6613	(0.0479)	0.6848	(0.0561)

†Sample size: 50. The mean and standard deviation across simulations of the squared norm difference between our estimates of the inspector sensitivities and the generating inspector sensitivity parameters are shown.

These simulations also affirm the sensitivity of the surveillance error estimates to prior specification. For the actual survey assignment of inspectors, the mean estimation error for the inspector sensitivities by using the misspecified $B(5,5)$ prior was 79% larger than the mean by using the correct $B(6.5,2)$ prior. Similarly, the mean estimation error by using the weak $B(1,1)$ prior was 14% larger than the mean error by using the $B(6.5,2)$ prior.

We conclude on the basis of the significance of these results that there is insufficient information in the data to infer the absolute sensitivities of inspectors. Nevertheless, the consistency across priors of the rankings of inspector sensitivities in the Mariano Melgar survey (Table 1) suggests that the relative sensitivities of the inspectors may be learned from the observed infestation data, even in the absence of reliable prior information.

6. Discussion

We proposed a spatial model to analyse spatially clustered presence–absence data that quantifies the amount of under-reporting and accounts for data missing not at random. The model allows us to capture the heterogeneity of the surveillance errors across the individuals collecting the data. Applying our model to surveys for the presence of *T. infestans* in the district of Mariano Melgar in Arequipa, Peru, we identified two additional at-risk localities for treatment. Applying a simpler model, which did not account for the difference in participation between infested and non-infested households, we found four additional at-risk localities for treatment.

The willingness of infested households to participate affirms the Ministry's local community outreach efforts and the willingness to co-operate of communities that are severely affected by triatomine infestations. We identified these locality level differences in participation previously in Buttenheim *et al.* (2014). The Ministry of Health based its treatment decisions on estimates produced under the assumption of missingness at random. Because we found a link between underparticipation and lower rates of infestation, it is likely that these estimates were an overestimate of prevalence as evidenced by the estimates from our more complete missingness not at random model. Because of the strong entomological risk of reinfestation in triatomine insects, overestimates of prevalence may be of value to safeguard against the risk of reinfestation. However, because of the additional strain on public resources, these distinctions and assumptions should be clearly outlined to policy makers.

On simulated data, we showed that a hypothetical randomized assignment of inspectors only marginally improved the estimation of the infestation and inspector sensitivities. This similarity suggests that the assignment that was used to conduct the Mariano Melgar survey is not seriously susceptible to confounding issues in spite of some spatial correlation.

We found that our model produced interpolation of the risk of infestation in non-participating households that was similar to a more standard linear regression model. Because we modelled the effect of surveillance error by using an external Bernoulli random variable to the linear infestation risk, we believe that the value of our model is that it is more easily interpreted. In our model, the true unobservable phenomenon of interest is modelled separately from the observed data. Policy decisions can then be made on this distinct quantity and false negative probability is explicitly estimated.

Related work was done in the context of sociological surveys in Casas-Cordero *et al.* (2013), where researchers incorporated the covariates of the interviewers conducting the survey to help to explain the variation in the data. Casas-Cordero *et al.* (2013) found that questionnaires conducted regarding perception of social disorder in urban neighbourhoods were consistent across various interviewers. In this study, we found that posterior distributions of inspector detection abilities departed strongly from the prior distribution. Although the overall posterior probability of detection was dependent on the particular prior specification, we found that the rankings of inspector abilities were consistent across a variety of prior distributions.

There are limitations to this study. First, despite accounting for the influence of the infestation on the participation and variations in participations between localities, we may be overlooking other covariates, which may also be key in mitigating the confounding. In our previous work (Barbu *et al.*, 2013), we found that the importance of covariates such as livestock and household building materials was limited compared with the effect of the spatial correlation. In addition, our model for the spatial effect (equation (3)) is based heavily on a distance threshold and does not have an easily interpretable covariance (Pickard, 1977).

Individual surveillance error models are useful for incorporating inspector-to-household labelling data in prevalence estimates. The model that is proposed here not only quantifies the amount of under-reporting in survey data but also allows for the relative estimation of inspector quality. The infestation maps are produced efficiently at a fine resolution and account for non-participating households even missing not at random, which gives insight into the distribution of residual infestation post treatment.

Acknowledgements

The Chagas Disease Working Group in Arequipa includes Fernando Malaga Chavez, Karina Oppe Alvarez, Andy Catacora Rospigliossi, Claudia Patricia Mena Cornejo, Dr Juan Cornejo del Carpio, Javier Quintanilla Calderon, Kate Levy, Malwina Niemierko, Victor Quispe Machaca, Jenny Ancca, Katty Borrini, Renzo Salazar Sanchez, Katty Borrini-Mayori, Danitza Pamo, Giovanna Moscoso, Lina Mollesaca, Maria Luz Hancco and Renzo Salazar. We thank Dr Christopher Paciorek and Dr Tony Smith for insightful comments. We gratefully acknowledge the work of the following organizations that have organized and conducted the Chagas disease control programme in Arequipa: the Ministerio de Salud del Perú, the Dirección General de Salud de las Personas, the Estrategia Sanitaria Nacional de Prevención y Control de Enfermedades Metaxénicas y Otras Transmitidas por Vectores, the Dirección General de Salud Ambiental, the Gobierno Regional de Arequipa, the Gerencia Regional de Salud de Arequipa, the Pan American Health Organization and the Canadian International Development Agency. We thank the district of Mariano Melgar for its participation in this study.

Appendix A: Gibbs sampler

We now outline the Gibbs sampler for implementing the model. Using the parameter expansion that was popularized in Albert and Chib (1993) for the probit link, the closed forms for all the conditional distributions are known and in the form of common distributions. Modifications can be made for the logistic link by following Holmes and Held (2006). The parameter expansion for the binary outcome y is implemented by introducing the continuous variable y_0:

$$y_0 = u + t + \varepsilon, \tag{8}$$

$$y_{1,i} = \mathbb{1}_{\{y_{0,i}>0\}} \tag{9}$$

where u is the conditionally auto-regressive model with precision $N(0, k_u \mathbf{\Lambda})$, t is the intercept and ε is standard Gaussian error. If the prior on t is given by $N(\mu, \tau)$, where again τ is the precision, the prior on k_u is given by $\Gamma(k, \theta)$, where k and θ represent the scale and shape parameters, and the prior on each element of β is $B(a, b)$. The conditional distributions for the model parameters are then given by

$$(k_{\mathbf{u}}|\mathbf{u}) \sim \Gamma\left(\frac{n-1}{2} + k, \frac{1}{2} u^{\mathrm{T}} \mathbf{\Lambda} u\right),$$

$$((\mathbf{u}, t)^{\mathrm{T}}|k_u, y_0) \sim N\left\{\begin{pmatrix} k_u\mathbf{\Lambda}+\mathbf{I} & 1 \\ \mathbf{1}^{\mathrm{T}} & n+\tau \end{pmatrix}^{-1} \begin{pmatrix} y_0 \\ \mathbf{1}^{\mathrm{T}} y_0 + \mu + \tau \end{pmatrix}, \begin{pmatrix} k_u\mathbf{\Lambda}+\mathbf{I} & 1 \\ \mathbf{1}^{\mathrm{T}} & n+\tau \end{pmatrix}\right\},$$

$$(y_{0,i}|u_i, t, y_{1,i}) \sim \begin{cases} N(u_i + t, 1|y_{0,1} > 0) & \text{if } y_{1,i} = 1, \\ N(u_i + t, 1|y_{0,1} < 0) & \text{if } y_{1,i} = 0, \end{cases}$$

$$(y_{1,i}|u_i, \beta(i)) \sim \mathrm{Bern}\left[p_i = \begin{cases} \dfrac{\{1-\beta(i)\}\,\Phi(u_i+t)}{\{1-\beta(i)\}\Phi(u_i+t) + 1 - \Phi(u_i+t)} & \text{if } \mathbb{1}_{\mathrm{NA}i} = 0, \\ \Phi(u_i+t) & \text{if } \mathbb{1}_{\mathrm{NA}i} = 1 \end{cases}\right],$$

$$(\beta_i|y_{I_i}, z_{I_i}) \sim B\left\{\sum_{j\in I_i} y_j z_j + a, \sum_{j\in I_i} y_j(1-z_j) + b\right\}.$$

The Mariano Melgar analysis was performed in R (R Core Team, 2014) on an Intel Core i7 processor clocked at 2.8 GHz, where 1000 iterations of the Markov chain were performed in 299.77 s. The size of the precision matrix in this work was 12 050 × 12 050 with 0.18% sparsity. We found convergence to be consistent irrespectively of the starting point. We found the slowest mixing and most auto-correlated variable in the chain to be k_u, owing to the strong dependence between u and k_u in the sampling scheme used. On the basis of the samples of k_u produced by the chain, we recommend discarding the first 10 000 samples as burn-in. After burn-in, we recommend thinning the samples and retaining only every 10th sample.

References

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*, 1st edn. Boca Raton: Chapman and Hall.

Banerjee, S., Wall, M. M. and Carlin, B. P. (2003) Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*, **4**, 123–142.

Barbu, C., Dumonteil, E. and Gourbière, S. (2011) Evaluation of spatially target strategies to control non-domiciliated triatoma dimidiata vector of Chagas disease. *PLOS Neglctd Trop. Dis.*, **5**, article e1045.

Barbu, C., Hong, A., Manne, J. M., Small, D., Calderón, J. E., Sethuraman, K., Quispe-Machaca, V., Ancca-Juárez, J., Cornejo del Carpio, J. G., Chavez, F. S., Náguira, C. and Levy, M. Z. (2013) The effects of city streets on an urban disease vector. *PLOS Computnl Biol.*, **9**, no. 1, article e1002801.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, **43**, 1–59.

Brier, G. (1950) Verification of forecasts expressed in terms of probability. *Mnthly Weath. Rev.*, **78**, 1–3.

Buttenheim, A. M., Paz-Soldan, V., Barbu, C., Skovira, C., Calderón, J. Q., Riveros, L. M. M., Cornejo, J. O., Small, D. S., Bicchieri, C., Naquira, C. and Levy, M. Z. (2014) Is participation contagious?: evidence from a household vector control campaign in urban Peru. *J. Epidem. Commty Hlth*, **68**, 103–109.

Casas-Cordero, C., Kreuter, F., Wang, Y. and Babey, S. (2013) Assessing the measurement error properties of interviewer observations of neighbourhood characteristics. *J. R. Statist. Soc.* A, **176**, 227–249.

Corrasco, H., Torrellas, A., García, C., Segovia, M. and Feliciangeli, M. (2005) Risk of Trypanosoma cruzi I (Kinetoplastida: Trypanosomatidae) transmission by Panstrongylus geniculatus (Hemiptera: Reduviidae) in Caracas (Metropolitan District) and neighboring States, Venezuela. *Int. J. Parasit.*, **35**, 1379–1384.

Dias, J., Silveira, A. and Schofield, C. (2002) The impact of Chagas disease control in Latin America: a review. *Mem. Inst. O. Cruz*, **97**, 603–612.

Faes, C., Ormerod, J. and Wand, M. (2011) Variational Bayesian inference for parametric and nonparametric regression with missing data. *J. Am. Statist. Ass.*, **106**, 959–971.

Germano, M. D., Acevedo, G. R., Cueto, G. A. M., Toloza, A. C., Vassena, C. V. and Picollo, M. I. (2010) New findings of insecticide resistance in Triatoma infestans (Heteroptera: Reduviidae) from the Gran Chaco. *J. Med. Entmol.*, **47**, 1077–1081.

Guhl, F. (2007) Chagas disease in Andean countries. *Mem. Inst. O. Cruz*, **102**, 29–38.

Holmes, C. and Held, L. (2006) Bayesian auxiliary variable models for binary and polychotomous regression. *Baysn Anal.*, **1**, 145–168.

Kang, E. and Cressie, N. (2011) Bayesian inference for the spatial random effects model. *J. Am. Statist. Ass.*, **106**, 972–983.

Lardeux, F., Depickère, S., Duchon, S. and Chavez, T. (2010) Insecticide resistance of triatoma infestans (hemiptera, reduviidae) vector of Chagas disease in Bolivia. *Trop. Med. Int. Hlth*, **15**, 1037–1048.

Lars, E., Beaty, B., Morrison, A. and Scott, T. (2009) Proactive vector control strategies and improved monitoring and evaluation practices for Dengue prevention. *J. Med. Entmol.*, **46**, 1245–1255.

Le, N. D., Sun, W. and Zidek, J. V. (1997) Bayesian multivariate spatial interpolation with data missing by design. *J. R. Statist. Soc.* B, **59**, 501–510.

Levy, M., Bowman, N., Kawai, V., Waller, L., Cornejo del Carpio, J., Cordova Benzaquen, E., Gilman, R. and Bern, C. (2006) Periurban Trypanosoma cruzi-infected Triatoma infestans, Arequipa, Peru. *Emergng Infect. Dis.*, **12**, 1345–1352.

Levy, M., Malaga Chavez, F., Cornejo Del Carpio, J., Vilhena, D., McKenzie, F. and Plotkin, J. (2010) Rational spatio-temporal strategies for controlling a Chagas disease vector in urban environments. *J. R. Soc. Interfce*, **7**, 1061–1070.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G. and Murphy, S. A. (2012) The prevention and treatment of missing data in clinical trials. *New Engl. J. Med.*, **367**, 1355–1360.

Matthys, B., Vounatsou, G., Tschannen, A., Becket, E., Gosoniu, L., Cissé, G., Tanner, M., N'goran, E. and Utzinger, J. (2006) Urban farming and malaria risk factors in a medium-sized town in Côte d'Ivoire. *Am. J. Trop. Med. Hyg.*, **75**, 1223–1231.

Paciorek, C. (2007) Computational techniques for spatial logistic regression with large data sets. *Computnl Statist. Data Anal.*, **51**, 3631–3653.

Pickard, D. (1977) Asymptotic inference for an Ising lattice: ii. *Adv. Appl. Probab.*, **9**, 476–501.

R Core Team (2014) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*, 1st edn. Boca Raton: Chapman and Hall.

Sanchez, L., Cortinas, J., Pelaez, O., Gutierrez, H., Concepcion, D. and Van der Stuyft, P. (2010) Breteau index threshold levels indicating risk for dengue transmission in areas with low aedes infestation. *Trop. Med. Int. Hlth*, **15**, 173–175.

Teixeira, M., Barreto, M., Costa, M., Ferreira, L., Vasconcelos, P. and Cairncross, S. (2002) Dynamics of dengue virus circulation: a silent epidemic in a complex urban area. *Trop. Med. Int. Hlth*, **7**, 757–762.

Trape, J., Lefebvre-Zante, E., Legros, F., Ndiaye, G., Bouganali, H., Druilhe, P. and Salem, G. (1992) Vector density gradients and the epidemiology of urban malaria in Dakar, Senegal. *Am. J. Trop. Med. Hyg.*, **47**, 181–189.

Wang, S., Lengeler, C., Smith, T., Vounatsou, P., Diadie, D., Pritroipa, X., Convelbo, N., Kientga, M. and Tanner, M. (2005) Rapid urban malaria appraisal (RUMA) I: epidemiology of urban malaria in Ouagadougou. *Malar. J.*, **4**, article 43.