Supplementary Material

for the manuscript
'The Bayesian Spatial Bradley–Terry Model:
Urban Deprivation Modelling in Tanzania'

1 A 1-d Study

We construct a one dimensional 'city', where the areas which are compared are points on the line. Although this is a simple toy example, it allows us to visualise the city and level of deprivation in each area. We draw the locations for 100 areas from a Laplace distribution with mean 0 and variance 8, as this gives a region near 0 which is densely packed with relatively small areas, and fewer but larger areas in the outlying parts of the city. We specify the level of deprivation in each area by the piecewise function

$$g(x) = \begin{cases} \sin(x) + x + \pi & \text{if } x < \pi, \\ \sin(4x) & \text{if } \pi \leq x \leq \pi, \\ \sin(x) - x + \pi & \text{if } x > \pi. \end{cases}$$

This gives a level of deprivation which changes quickly in the city centre compared to the outskirts.

We simulate the comparisons according to the model in equation (2), choosing pairs of areas uniformly at random to compare. We simulate data sets of various sizes to mimic real data collection. The sizes of simulated data sets used in this paper are shown in Table 1. These are the same as in the 2-d study in Section 3 of the main text, but with the larger sizes excluded since here there are many fewer areas and we find that both models reach data saturation by 9,000 comparisons.

We fit the BSBT model and run the MCMC algorithm for 500,000 iterations, removing the first 100,000 as a burn-in period. We estimate the quality of each area and learn plausible values of the variance hyperparameter α_λ^2. As in the 2-d study, we fix the tuning parameter $\delta = 0.01$, based on initial runs of the algorithm, and fix $\chi = \omega = 0.1$ in the prior

Table 1: Data set sizes used in the simulations studies, using 180 comparison per judge hour.

Judge hours	1	2	5	10	25	50
Comparisons	180	360	900	1,800	4,500	9,000

distribution on α_λ^2. The top row of Figure 1 gives results for the BSBT model using 900 and 9,000 comparisons; it shows the true deprivation levels, the location of the areas and the posterior median deprivation with a 95% credible interval for each area. We see two main effects from increasing the number of comparisons. The first is increasing accuracy of inference, with better estimates and less uncertainty when using 9,000 comparisons. The second is the model's ability to deal with extreme levels of deprivation, either very deprived or very affluent areas. The areas on the outskirts of the synthetic city make this a challenging data set for the BSBT model, since they are extreme both spatially and in terms of deprivation. When using smaller data sets, inferred deprivation levels in the BSBT model are pulled towards 0 by the prior distribution. As we do not have sufficient data to estimate the extent of the deprivation in the most extreme areas, we also underestimate the variance parameter. Although we do not accurately estimate the extent of the extremes, we do successfully identify which areas have extremes of deprivation.

Corresponding results from fitting the standard BT model to the same data sets are given in the bottom row of Figure 1. These plots show the same detail of the true deprivation and locations of areas, but here show MLEs and corresponding intervals based on quasi-variances. In the smallest data set there are some areas which feature in very few comparisons, so estimates for their level of deprivation are highly uncertain. This shows in the lower-left plot as intervals spanning all shown deprivation values and/or the point estimate not being visible on the scale used. Many estimates are also quite poor compared to the BSBT results. Figure 1 also shows that when using the 9,000 comparisons the standard and BSBT models perform fairly similarly. This is expected since the data set is large enough to estimate the model parameters well using either model.

To assess the model fit, we compute the Mean Absolute Error (MAE) for the result of each set of comparisons, which is given by

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\lambda_i - \hat{\lambda}_i|,$$

where $\hat{\lambda}_i$ is the estimate corresponding to the MLE or posterior median for area i. Figure 2 shows the log MAE for each data set. The BSBT model performs better than the standard

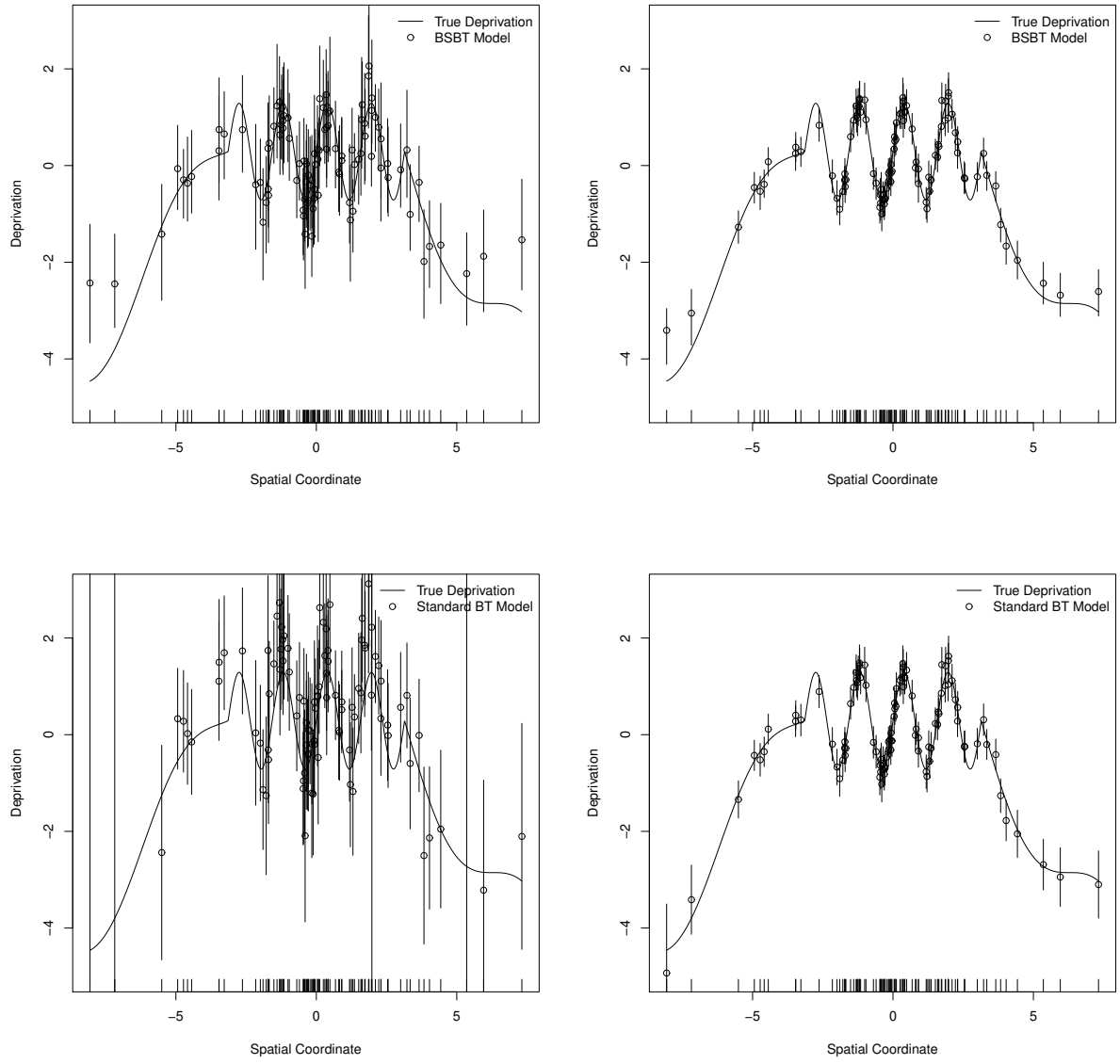

Figure 1: Results for the 1-d simulation study using the BSBT and standard BT models. The top row shows the results of the BSBT model and the bottom row shows the results of the standard BT model. The left column shows the models fitted with 900 comparisons and the right column with 9,000 comparisons. The same scales are used on all plots for ease of comparison, but this means that in the bottom row some intervals are larger than can be displayed. Tick marks on the horizontal axes show the locations of the areas in the 1-d city.

BT model in all cases, though as discussed above we approach data saturation for the largest data sets. Figure 2 also shows that when we fix the number of comparisons, the error in the BSBT model is always lower than in the standard model, and especially for smaller data sets offers a notable improvement. For example, when using 900 comparisons, MAE in the BSBT model (0.418) is less than half that in the standard model (0.975). The level of error in the standard model with 1,800 comparisons is of the same order as the BSBT model with 900 comparisons, demonstrating that with the BSBT model we can collect appreciably less data without compromising the quality of the estimates. In small data sets we cannot compute the MLE for areas which are not featured in any comparisons. As such, when using 180 comparisons we are unable to compute the MAE for the standard BT model. The BSBT model does not suffer from this issue since it uses a correlated prior distribution, so we still get an estimate of deprivation for these areas, albeit with large uncertainty.

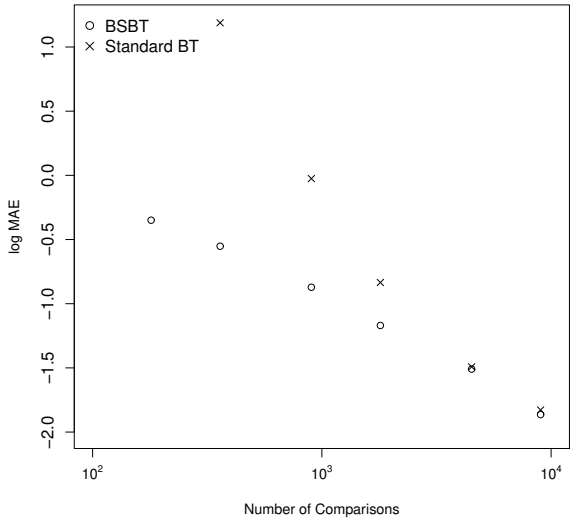

Figure 2: Log MAE for the simulation study comparing performance of the standard BT and the BSBT models in terms of error as a function of the number of comparisons.

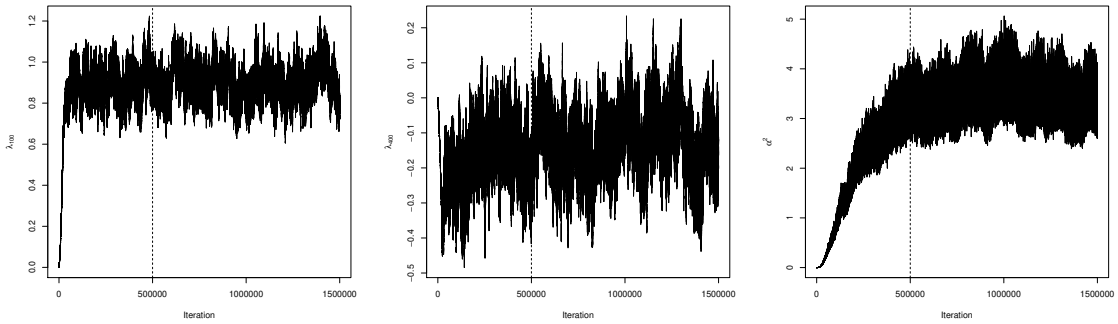

Figure 3: Trace plots for λ_{100} (left), λ_{400} (middle) and α_λ^2 (right). The dashed line is at 500,000 iterations and marks the end of the burn-in period.

2 BSBT diagnostics for the Dar es Salaam data set

We fit the BSBT model to the Dar es Salaam data set and produce the results shown in Section 4.1 of the main text. Trace plots for λ_{100} and λ_{400} and α_λ^2 are shown in Figure 3. These show that the deprivation parameters converge quickly, but the variance hyperparameter is slower to converge. Based on the diagnostic plots, we consider the first 500,000 iterations as a burn-in period, and compute the posterior distributions for the model parameters from the remaining 1,000,0000 iterations. The trace plots show the Markov chain is mixing well.

To determine the reliability of the judges and if they were carrying out the comparisons faithfully, we use a χ^2 style heuristic. For each judge, we see how the observed comparisons differ from the expected comparisons based on the fitted model. The value of the heuristic for each judge (j) is given by

$$X_j^2 = \frac{1}{N_j} \sum_{i=1}^{N_j} \frac{(O_{j,i} - E_{j,i})^2}{E_{j,i}},$$

where N_j is the number of comparisons judge j made, and $O_{j,i}$ and $E_{j,i}$ are the observed and expected outcomes of judge j's i-th comparison. The latter is calculated using equation (1) of the main text and posterior mean estimates of $\boldsymbol{\lambda}$. The values are shown in Figure 4, which shows that the judges are largely homogeneous by this measure. We investigate the comparisons for the worst five judges (by this measure) and find that they are nonetheless largely consistent with the consensus.

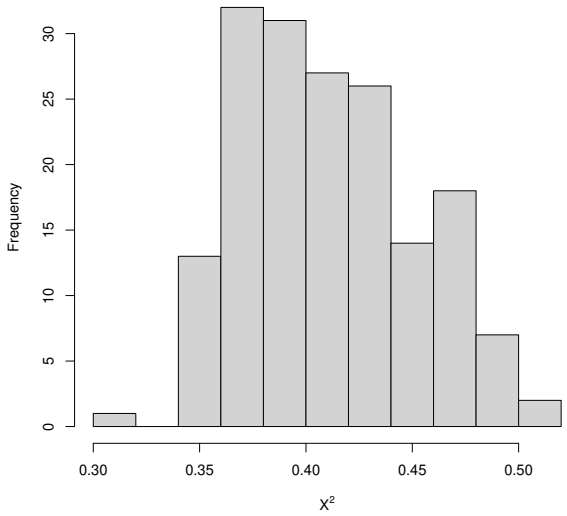

Figure 4: A histogram showing the judge reliability heuristic for each of the judges in the study.

3 Treatment of ties for the Dar es Salaam data set

Around one in seven of the comparisons are tied comparisons and their treatment affects the results. Throughout, we have considered a treatment where the winner of each tied comparison was randomly allocated. We carry out two sensitivity analyses to justify our choice of treatment. The first concerns the random random allocation of winners for the tied comparison and the second investigates other treatments of these ties.

3.1 Randomly allocating a winner

We investigate the effect of the random allocation of winners of each tied comparison on the results. We generate 20 new data sets, in each allocating a winner of each tied comparisons according to a different random seed. We fit the standard BSBT model to each data set using the same fitting procedure used in the main text (1,500,000 iterations with the first 500,000 removed as a burn-in period). We compute the posterior mean deprivation levels for each data set and compare these estimates to those reported in the main text using Spearman's rank correlation coefficient. This is given by

$$r_s = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)},$$

where d_i is the difference in ranks of λ_i stemming from the two data sets. Over all 20 data sets, the lowest correlation coefficient is 0.993 and the mean value of the correlation coefficients is 0.994, suggesting that the random allocation has little effect on the results. Figure 5 shows the values of the estimates for λ_{50}, λ_{150}, λ_{250} and λ_{350}, and in all four cases the range of estimates is small and the value reported in the main text is a representative estimate.

3.2 Other treatments

We consider two further treatments of the tied comparisons, which are:

- treating ties as half a win for both subwards, as described in Glickman (1999), effectively replacing y_{ij} in equation (2) of the main text by $y_{ij} + t_{ij}/2$, where t_{ij} is the number of tied comparisons of subwards i and j; and
- discarding the tied comparisons altogether.

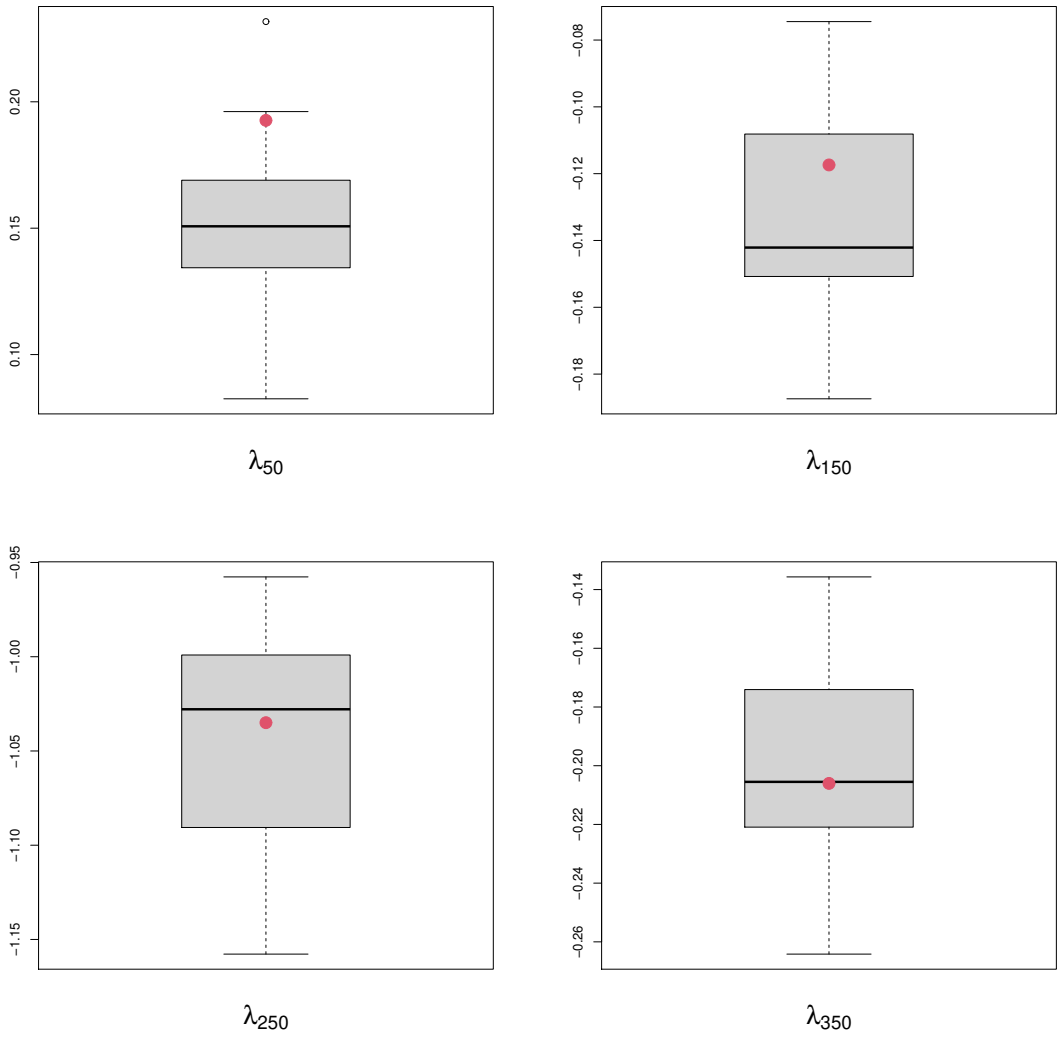

Figure 5: Box plots showing the posterior means for the 20 data sets using random allocations. The solid point shows the estimate reported in the main text.

We fit the BSBT model to the data using each of the treatments, running the MCMC algorithm for 1,500,000 iterations and removing the first 500,000 iterations as a burn-in period. This is the same model fitting procedure as used in the main text.

Figure 6 shows the posterior means (top row) and variances (bottom row) for the deprivation in each subward using the two treatments, compared to the results reported in the main text using a random allocation. When using either treatment, the rankings of the subwards are largely unchanged; the Spearman's rank correlation coefficient between randomly allocating a winner and treating ties as half a win is 0.993, and when discarding the tied comparisons is 0.995. Treating a tied comparison as half a win produces deprivation estimates that are broadly similar to allocating a winner at random, but suggests there is some slight shrinkage when using a random allocation. The variances using this treatment, however, are considerably smaller than using a random allocation. Finally, discarding the tied comparisons is the least attractive of the three treatments, but provides an upper bound for the level of uncertainty. Discarding the tied comparisons leads to the estimates of the deprivation levels in the most affluent and deprived subwards to be more extreme, as we are discarding data which will tend to make estimated affluences more similar to each other.

Overall we conclude that our results, using random allocation of a winner to break the ties, are robust to the random allocation used and are most conservative in terms of uncertainty of estimates amongst methods which use the ties in some way.

4 The standard BT model for the Dar es Salaam data set

We fit the standard BT model to the Dar es Salaam data set. To fit the model we use the `BradleyTerry2` R package. This took around 5 minutes on a 2019 iMac with a 3 GHz CPU. A map of the inferred deprivation parameters from the standard BT model are shown in Figure 7. This is broadly very similar to the map shown in the main text (Figure 5) of the results for the BSBT model, though the smoothing effect of the prior is apparent in a few places (mainly near the most affluent and most deprived areas). We directly compare the results of the two models in Figure 8. There is little difference between the results from the two models. As the uncertainty in the two models is measured in different ways, they are difficult to compare directly, however the uncertainties are of the same order and there is a clear pattern of subwards with relatively high/low uncertainty under one model also having relatively high/low uncertainty under the other model.

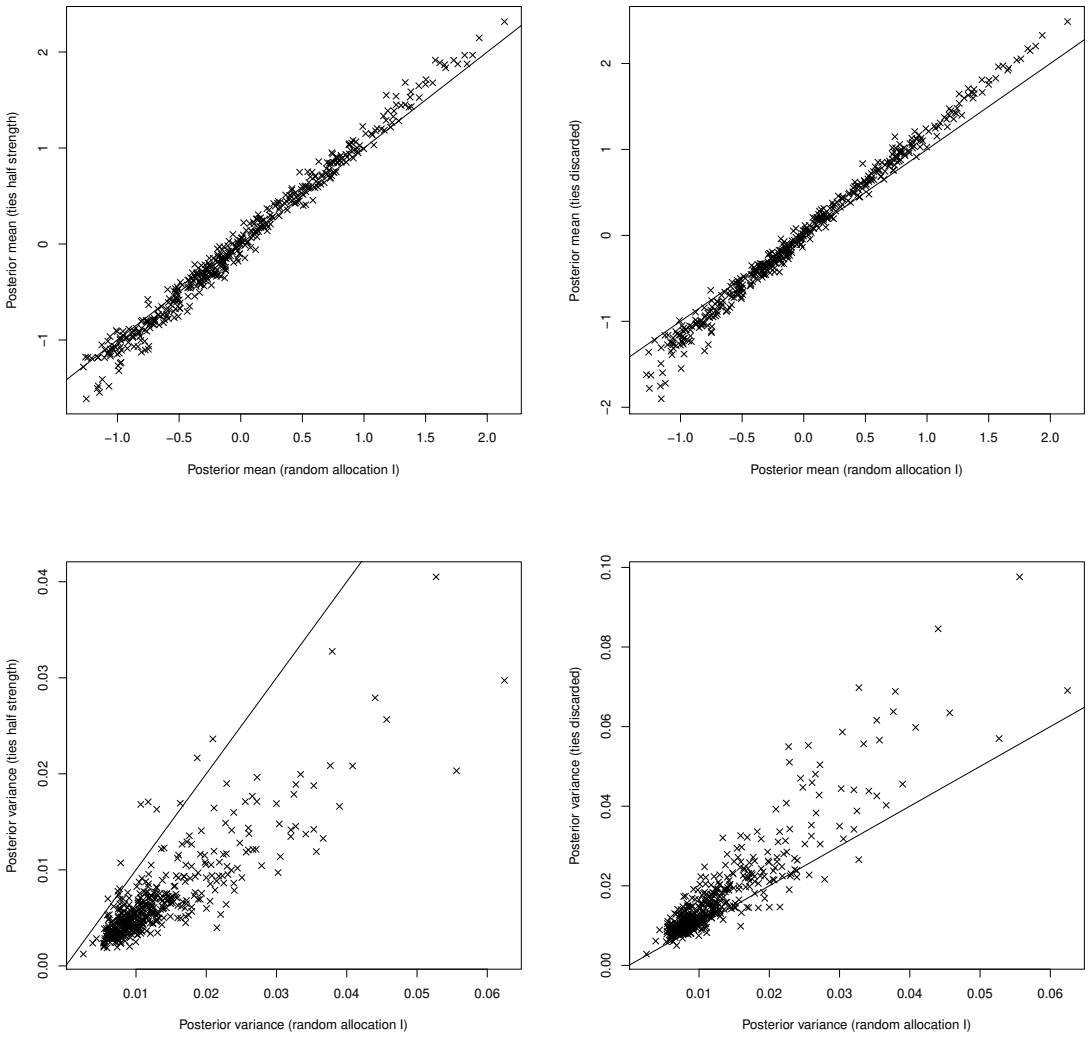

Figure 6: Posterior means and variances for the deprivation in each subward for the two treatments of ties, compared to the treatment used in the main text.

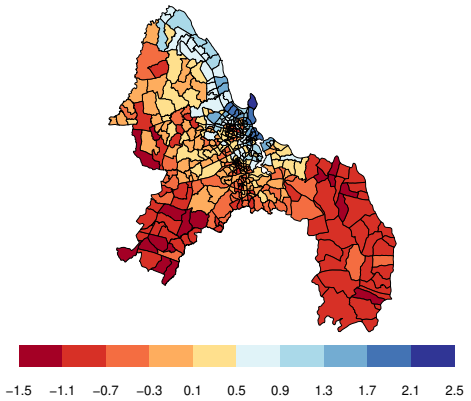

Figure 7: The results of the standard BT model applied to the full Dar es Salaam data set.

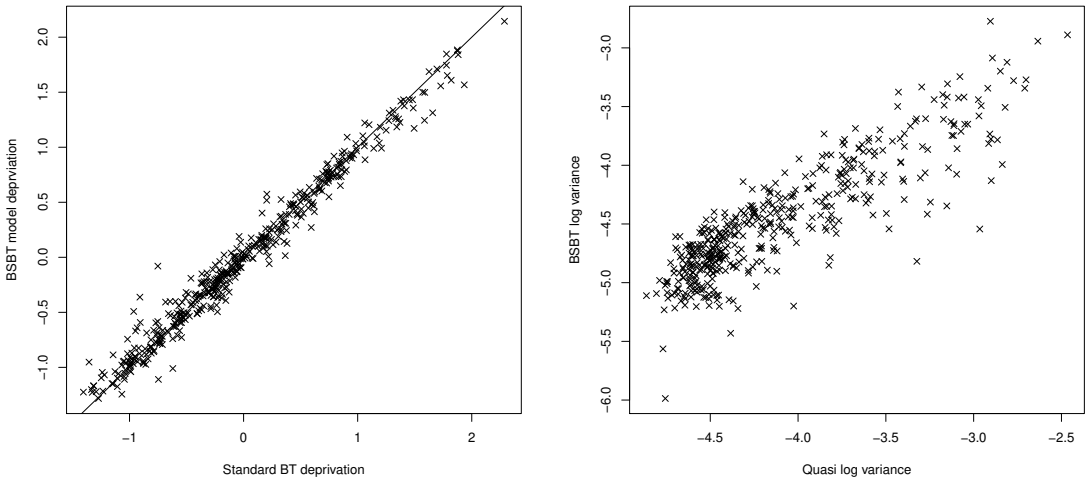

Figure 8: Left: The BSBT estimates plotted against the standard BT estimates. Right: The quasi variances for the BT model plotted against the variances of the posterior distributions for the BSBT model. Both are plotted on a log scale.

5 BSBT diagnostics for the Dar es Salaam data set with judge information

We fit the BSBT model with judge information to the Dar es Salaam data set and produce the results shown in Section 4.2. Trace plots for λ_{100}, the grand mean for subward 100, $\beta_{0,100}$, the difference between men and women's judgements in subward 100, α^2_λ and α^2_1 are shown in Figure 9. The same plots for students and non-students are shown in Figure 10. As we ran the MCMC algorithm for 5,000,000 iterations in both studies, we thin the results to reduce the required memory and store every 50th iteration. We take the 35,000th thinned iteration to be the end of the burn-in period. The mixing could be improved, particularly for α^2_1, which could be achieved by using adaptive MCMC and adapting the underrelaxed tuning parameter, δ.

References

Glickman, M. E. (1999) Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **48**, 377–394. URL: `https://doi.org/10.1111/1467-9876.00159`.

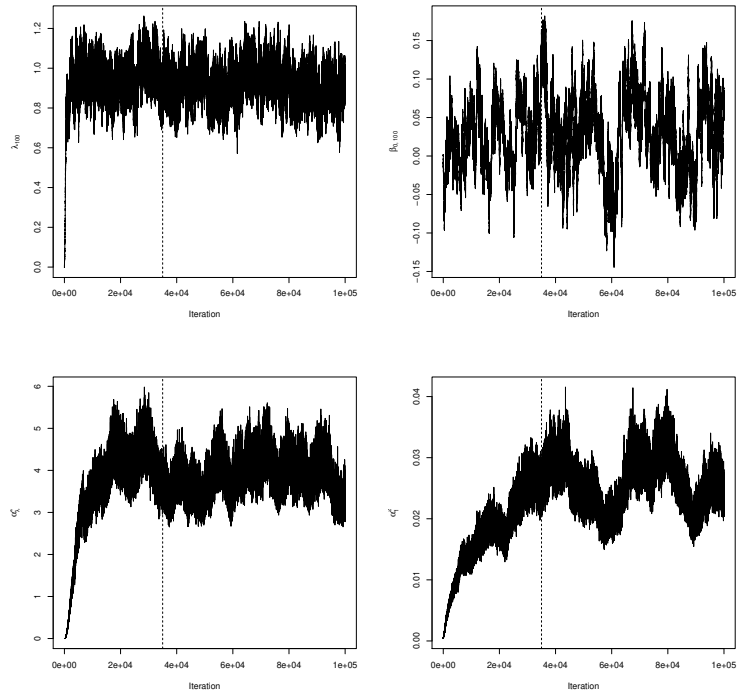

Figure 9: Men and women study. Trace plots for λ_{100} (top left), $\beta_{0,100}$ (top right), α_λ^2 (bottom left) and α_1^2 (bottom right). The 5 million iterations have been thinned to 100,000 and the dashed line at 35,000 iterations marks the end of the burn-in period.

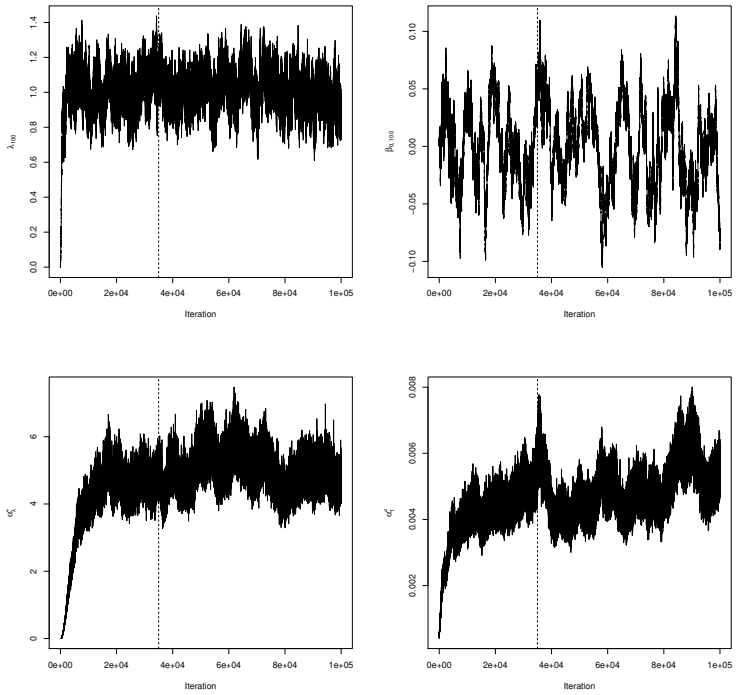

Figure 10: Student and non-student study. Trace plots for λ_{100} (top left), $\beta_{0,100}$ (top right), α_λ^2 (bottom left) and α_1^2 (bottom right). The 5 million iterations have been thinned to 100,000 and the dashed line at 35,000 iterations marks the end of the burn-in period.