Multi-Level Spatial Comparative Judgement Models To Map Deprivation

Seymour, Rowland; Sirl, David; Preston, Simon; Goulding, James

DOI:
10.5281/zenodo.8314257

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Seymour, R, Sirl, D, Preston, S & Goulding, J 2023, Multi-Level Spatial Comparative Judgement Models To Map Deprivation. in *Proceedings of the Joint Statistical Meetings 2023.* Zenodo, Joint Statistical Meetings 2023, Toronto, Ontario, Canada, 5/08/23. https://doi.org/10.5281/zenodo.8314257

Multi-Level Spatial Comparative Judgement Models To Map Deprivation

Rowland G. Seymour* David Sirl† Simon Preston ‡ James Goulding§

Abstract

While current comparative judgement models provide strong algorithmic efficiency, they remain data inefficient, often requiring days or weeks of extensive data collection to provide sufficient pairwise comparisons for stable and accurate parameter estimation. This disparity between data and algorithm efficiency is preventing widespread adoption, especially so in challenging data-collection environments such as mapping human rights abuses. We address the data inefficiency challenge by introducing the finite element Gaussian process Bradley–Terry mixture model, an approach that significantly reduces the number of pairwise comparisons required by comparative judgement models. This is achieved via integration of prior spatial assumptions, encoded as a mixture of functions, each function introducing a spatial smoothness constraint at a specific resolution. These functions are modelled nonparametrically, through Gaussian process prior distributions. We use our method to map deprivation in the city of Dar es Salaam, Tanzania and locate slums in the city where poverty reduction measures can be carried out.

Key Words: Bradley–Terry, Preference Learning, Bayesian Computation, Gaussian Processes

1. Introduction

Comparative judgement models, such as the Bradley–Terry model Bradley and Terry (1952), rank and estimate the features of a group of objects by modelling a set of pairwise comparisons made between them. The use of the approach now proliferates across a range of domains, due to the fact that, while attribution of a single, consistent value to some object can be an extremely difficult task, assessing an object's relative value in comparison with another can be trivial - whether in reflecting the superiority of one sports team over another Cattelan et al. (2012); the relevance of a document compared to rivals in a set of search results Radlinski and Joachims (2007); or even the ability of one animal to outfight a competitor for a mate Stuart-Fox et al. (2006). More recently, comparative judgement studies have been used in social good applications, such as mapping deprivation (Seymour et al., 2022b) and violence against women and girls (Seymour et al., 2023).

Bradley-Terry models are of increasing interest to the machine learning community, as we seek to generate large, accurately labelled data sets to underpin modelling processes. In many contexts, traditional labelling methods are impractical, from assessment of disease risk in challenging clinical environments, to the collection of UN Sustainable Development Goal (SDG) indicators, such as levels of poverty, modern slavery and gender-violence, in countries with limited infrastructure. In such conditions standard data-collection techniques can be logistically challenging, prohibitively expensive and rapidly out of date. While comparative judgement models offer an attractive alternative, current models remain *data inefficient*, often requiring days or weeks of extensive data collection to provide sufficient pairwise comparisons for stable and accurate parameter estimation. This disparity between data and algorithmic efficiency is hampering widespread adoption - and much practical pressure remains to minimise the number of comparisons required to accurately

*School of Mathematics, University of Birmingham, UK
†Mathematical Sciences, University of Nottingham, UK
‡Mathematical Sciences, University of Nottingham, UK
§N/LAB, University of Nottingham, UK

recover ground-truth qualities for the objects of concern. Furthermore, as comparative judgement models describe the relative difference in qualities of objects, novel methods are required to provide results which are readily interpretable by non-scientific practitioners.

Addressing these issues, we introduce a finite element Gaussian process Bradley–Terry mixture model (GP-BT), applicable to any context where spatial correlations exist. Data efficiency improvements are achieved by adding structure into the model via explanatory variables, and leveraging prior assumptions about them encoded as a mixture of functions. These covariates can be of any dimension, but we focus on the spatial case due to its wide applicability, with each function introducing spatial smoothness constraints into the model at a specific resolution. These functions are modelled nonparametrically, placing a Gaussian process (GP) prior distribution on each and developing an efficient Markov chain Monte Carlo algorithm to learn function structure. Incorporating spatial constraints into the model allows the features of each object in the set to depend on the features of nearby objects, greatly reducing the amount of data that needs to be collected. As the model uses multi-level spatial constraints, we are able to model the spatial pattern in features at different resolutions and tailor the results to the needs of the end user. Practical implications of this new technique are demonstrated via two real-world case-studies: modelling deprivation patterns across the UK and providing the highest-resolution predictions to date of poverty levels across the 452 subwards of Dar es Salaam, Tanzania

2. Background

The Bradley–Terry (BT) model was first outlined in Bradley and Terry (1952) and has been widely used for analysing such comparative data (other examples including analysis the ability of major league baseball teams Phelan and Whelan (2018), ranking chess players Caron and Doucet (2012) and in educational assessment Pollit (2012)). The model describes the probability one object of a set 'beats' another when the two are compared. It does this by assigning a quality to each element of the set and describing the probability of one element beating another as a function of the qualities of the two elements being compared. In the standard BT model the quality of each object is a separate parameter, i.e. each object is independent of all other objects. Inference for the model parameters has typically been done by maximising the likelihood function (e.g. Davidson (1970), Hunter (2004)), but more recently attention has turned to Bayesian methods. As the posterior density is intractable, the authors of Adams (2005) proposed a Metropolis-Hastings algorithm to explore the density. In Guiver and Snelson (2009), the authors proposed an expectation-propagation method, which has computational advantages for large data sets. Expectation maximisation methods, based on minorization-maximisation described in Hunter (2004), have also been developed Caron and Doucet (2012). Bayesian hierarchical models which infer the model parameters and hyperparameters are developed in Phelan and Whelan (2018). As far as Bayesian nonparametric methods are concerned, using a Dirichlet Process prior distribution to infer the model parameters has been proposed, and a GP method has been developed for a preference learning model Chu and Ghahramani (2005).

Current methods for including structure in the BT model have been almost exclusively parametric, mainly through linear predictors, e.g. Springall (1973), Stern (2011). For contexts where objects naturally have some spatial association, these methods are unsuitable as linear predictors cannot describe complex spatial structure. We instead take a Bayesian nonparametric approach and use a mixture of GP prior distributions to model the spatial structure in Euclidean space and at a number of resolutions. This provides a flexible frame-

work which can model non-linear spatial structures, which reduces the amount of data required for accurate estimates. By decomposing the quality of the objects at different resolutions, we can provide results which can be meaningfully interpreted.

We evidence both the effectiveness of this approach in reducing the data collection burden and demonstrate the interpretability of the output via two studies, set in the context of urban modelling. The GP-BT model is applicable to domains where data is both noisy and hard to collect, and therefore highly relevant to data such as UN SDG indicators. In this vein we go on to demonstrate the applicability of the technique via two real-world case-studies estimating deprivation levels. In the first, we analyse spatial trends in deprivation levels in England. As the deprivation levels are known, we evidence how we can considerably reduce the amount of data compared to the standard model without comprising the quality of the results. In the second study, we estimate the deprivation levels across the 452 subwards of Dar es Salaam, Tanzania; using significantly less data than would previously have been required and providing more insight to the results than comparable methods.

The structure of the remainder of the paper is as follows. In §3 we describe the standard BT model, then describe the mixture model framework and the Bayesian nonparametric approach that we use to incorporate spatial structure. We apply our model to two data sets. In §4, we investigate deprivation in local authorities areas in England, where the true deprivation level is known. We find several trends in deprivation levels at various spatial resolutions and compare the efficiency of our model to existing methods. We then describe the results of our model applied to a comparative judgement data set for deprivation in Dar es Salaam, Tanzania in §5. We show that our Bayesian nonparametric spatial structure considerably reduces the amount of data that needs to be collected and gives insight into geographic trends in deprivation. Finally, we make concluding remarks and discuss possible extensions for this work.

3. Model

Our underlying framework, from a mathematical viewpoint, is that we have a set of N objects whose relative qualities $\lambda_i \in \mathbb{R}$ $(i = 1, \ldots, N)$ we wish to infer from the outcomes of a set of pairwise comparisons. We first review the standard BT model which assumes that the qualities λ_i are independent quantities for each object, then describe our method for incorporating geographically-induced correlations between physically proximate objects into the modelling before concluding with a summary of our model fitting procedure.

3.1 The Bradley–Terry Model

For a comparison between object i and object j, the outcome is modelled as

$$Y \sim \text{Bernoulli}(\pi_{ij}); \tag{1}$$

in which $Y = 1$ indicates that i beats j and $Y = 0$ indicates that j beats i; and π_{ij} is the probability, dependent on the qualities λ_i and λ_j of objects i and j respectively, that i will beat j. The Bradley–Terry model assumes

$$\pi_{ij} = \frac{\exp(\lambda_i)}{\exp(\lambda_i) + \exp(\lambda_j)} \quad \Longleftrightarrow \quad \text{logit}(\pi_{ij}) = \lambda_i - \lambda_j, \tag{2}$$

where $\text{logit}(\pi) = \log(\pi) - \log(1 - \pi)$. The data set consists of K pairwise comparisons of these objects: we write (i_k, j_k) $(k = 1, \ldots, K)$ for the objects compared in the k-th comparison and y_k to denote the outcome. For identifiability of the $\{\lambda_i\}$, it is necessary to impose a single linear constraint such as $\sum_{i=1}^{N} \lambda_i = 0$, and for the experimental design —

i.e. the particular choice of which pairwise comparisons (i_k, j_k) are made — to be such that if the objects are regarded as nodes of a graph and y_k denotes a directed edge from i_k to j_k, then there is a path from i to j for all i and j Hunter (2004). We discuss the experimental design further in §4 and §5.

Given K independent judgements, the likelihood function for the qualities $(\lambda_1, ..., \lambda_N)$ is the product

$$\pi(\boldsymbol{y} \mid \lambda_1, \ldots \lambda_N) = \prod_{k=1}^{K} \pi_{i_k,j_k}^{y_k} \left(1 - \pi_{i_k,j_k}\right)^{1-y_k}, \tag{3}$$

where $\boldsymbol{y} = (y_1, ..., y_K)$.

The principal contribution of this paper is to incorporate spatial structure into the Bradley–Terry model by assuming $\lambda_i = f(\boldsymbol{x}_i)$, i.e. the quality of an object is a function of its spatial position $\boldsymbol{x}_i \in \mathbb{R}^d$. This is in principle a completely arbitrary function and for the purposes of inference we will assume a finite element mixture framework, which we now describe.

3.2 Finite Element Mixture Framework

To allow for the quality of each object to be modelled on a number of spatial resolutions, we construct a finite element mixture framework Gelman et al. (2013). This is valuable as it allows us to leverage the different kinds of spatial structures that exist in the domain in which the objects reside. Different resolutions of spatial structure might stem from local neighbourhood relationships in a city, perhaps, to trends that run across districts, through to city wide structures such as a north-south divide. It is the leveraging of such patterns that will allow us to vastly reduce the number of comparison required to recover ground truth qualities. We provide this facility by associating each comparison with one of M functions; thus the overall structure is given by

$$\lambda_i = p_1 f_1(\boldsymbol{x}_i) + \ldots + p_M f_M(\boldsymbol{x}_i),$$

where $p_1, \ldots, p_M \geq 0$ are mixture weights subject to the constraint $\sum_{m=1}^{M} p_m = 1$ and $f_1, \ldots, f_M$ are arbitrary functions chosen so that the mixture class is suitably rich. In our application we will take $f_1, \ldots, f_M$ to be functions characterising spatial variation on different length scales; see §3.3.

For each of the K comparisons, we introduce a latent variable z_k whose value determines to which component of the mixture comparison k corresponds. The distribution of z_k is given by $P(z_k = m) = p_m$ $(m = 1, \ldots, M)$, so p_m is the probability that comparison k corresponds to the m^{th} component of the mixture. For subsequent use we define $\boldsymbol{p} = \{p_1, \ldots, p_M\}$. The result of comparison k thus has the conditional distribution

$$(y_k \mid z_k = m) \sim \text{Bernoulli}\left(\pi_{i_k j_k}^{(m)}\right); \qquad \pi_{ij}^{(m)} = \frac{\exp(f_m(\boldsymbol{x}_i))}{\exp(f_m(\boldsymbol{x}_i)) + \exp(f_m(\boldsymbol{x}_j))}$$

being the probability of i beating j when f_m is used for the comparison. Augmenting the model in equation (3) to take into account the set of latent variables $\boldsymbol{z} = \{z_1, \ldots, z_K\}$, gives the mixture model likelihood function

$$\pi(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{p}, f_1, \ldots, f_M) = \prod_{k=1}^{K} \left(\pi_{i_k j_k}^{(z_k)}\right)^{y_k} \left(1 - \pi_{i_k j_k}^{(z_k)}\right)^{1-y_k}. \tag{4}$$

Although we could propose parametric forms for the functions f_m when fitting models, it is difficult to justify any particular parametric form. We instead use Bayesian nonparametric methods to model these functions.

3.3 The Bayesian Nonparametric Approach

Bayesian nonparametric models are a class of models which have an infinite-dimensional parameter space. We choose this parameter space to be the set of all possible solutions to the problem in question. For example, for a regression problem, we typically choose this to be the space all continuous functions. Bayesian nonparametric models allow us to make more general assumptions about the generating process, instead of strict parametric assumptions.

To model the functions f_m we use a GP. A GP distribution on a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is completely described by its mean function μ and covariance function k. Denoted

$$f \sim \mathcal{GP}(\mu(\cdot), k(\cdot,\cdot)),$$

a GP can be defined as an infinite collection of random variables, any finite subset of which follow a multivariate normal distribution Rasmussen and Williams (2006). That is, for a collection of points $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, then

$$\boldsymbol{f} \sim \text{MVN}(\boldsymbol{\mu}(\boldsymbol{X}), \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}))$$

where MVN denotes the multivariate normal distribution, $\boldsymbol{f} = (f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n))$, $\boldsymbol{\mu}(\boldsymbol{X}) = (\mu(\boldsymbol{x}_1), \cdots, \mu(\boldsymbol{x}_n))$ and $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})$ is an n-by-n covariance matrix with $(i,j)^{\text{th}}$ element equal to $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

The GP can be viewed as a prior distribution over the space of all plausible functions which satisfy the (weak) assumptions specified via the covariance function. In this article, our choice of covariance function is the squared exponential

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j; \alpha, l) = \alpha^2 \exp\left(-l^{-2} \left\|\boldsymbol{x}_i - \boldsymbol{x_j}\right\|_2^2\right).$$

With this covariance function, samples drawn from the GP prior distribution are continuous and smooth. The squared exponential covariance function has two hyperparameters: α^2, which specifies the signal variance; and the length scale parameter l, which can be loosely interpreted as the number of units we must move in the input space to see appreciable change in the function value. We follow the common practice of taking $\mu(\cdot)$ to be identically zero Rasmussen and Williams (2006).

3.4 Fitting our GP-BT Model

Having developed our model in the preceding subsections we now describe how we fit it in our Bayesian framework to infer posterior distributions for the quality of each object or subward, given the pairwise comparison data $\boldsymbol{y}$. The parameters we infer are the functions $f_1, \ldots, f_M$ at the points $\boldsymbol{x}$, the GP variance hyperparameters $\boldsymbol{\alpha}^2 = \{\alpha_1^2, \ldots, \alpha_M^2\}$, the mixture weights $\boldsymbol{p}$ and the latent variables $\boldsymbol{z}$. The posterior distribution of these parameters is given by

$$\begin{aligned}\pi(\boldsymbol{p}, \boldsymbol{z}, f_1, \ldots, f_M, \boldsymbol{\alpha}^2 \mid \boldsymbol{y}) \propto{}& \pi(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{p}, f_1, \ldots, f_M)\pi(\boldsymbol{z})\pi(\boldsymbol{p}) \\ &\times \prod_{m=1}^{M} \pi(f_m \mid \alpha_m^2)\pi(\alpha_m^2). \end{aligned} \tag{5}$$

The first term on the right hand side of this formula is the likelihood function (4). The remaining terms are the prior distributions on the model parameters. When applying our algorithm we use independent uniform (i.e. uninformative) prior distributions on $\boldsymbol{z}$, and the

prior distribution on $\boldsymbol{p}$ is uniform on its simplex of possible values. We place independent GP prior distributions on $f_1, \ldots, f_M$: $f_m \sim \mathcal{GP}\left(0,\, k(\cdot,\cdot;\alpha, l_m)\right)$ and $l_1 < l_2 < \ldots < l_M$ are chosen to reflect appropriate, context-dependent, spatial resolutions. We place independent inverse-Gamma prior distributions on the variance hyperparameters, $\boldsymbol{\alpha}^2$.

The resulting posterior distribution has a non-standard form and we generate samples from it using an MCMC algorithm. We now describe the individual steps in the MCMC algorithm, and the full algorithm is given in Algorithm 1. As we are using a conjugate prior distribution for the component weights $\boldsymbol{p}$, the full conditional distribution is a Dirichlet distribution

$$\boldsymbol{p} \mid \boldsymbol{z} \sim \text{Dir}(\chi_1 + n_1, \ldots, \chi_M + n_M),$$

where n_m is the number of comparisons associated with component m, and χ_i are parameters from the prior distribution. As we use a uniform prior distribution, we set $\chi_1 = \ldots \chi_P = 1$. We can sample from this distribution directly using a Gibbs sampler. The prior distributions for the variance parameters are also conjugate and the full conditional distribution for the m^{th} variance hyperparameter is

$$\alpha_m^2 \mid \boldsymbol{f}_m \sim \text{inv-}\Gamma(\psi_0 + n_m/2,\, \omega_0 + \boldsymbol{f}_m^T \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}; 1, l_m)^{-1} \boldsymbol{f}_m/2).$$

The parameters ψ_0 and ω_0 are the rate and scale of the prior distribution. We follow Gelman (2006) and set $\psi_0 = \omega_0 = 0.1$, as this results in a prior distribution that is somewhat uninformative. Regarding the latent variables $\boldsymbol{z}_k$, the full conditional probability that comparison k is associated with component m of the mixture is

$$\pi(z_k = m \mid y_k, \boldsymbol{p}, f_1, \ldots, f_M) = \frac{\pi(y_k \mid z_k = m)}{\sum_{m'=1}^{M} \pi(y_k \mid z_k = m')}.$$

The full conditional distribution for f_m, the deprivation at spatial resolution l_m is

$$\pi(f_m \mid \boldsymbol{p}, \boldsymbol{z}, \alpha_m^2) \propto \pi(f_m \mid \alpha_m^2) \prod_{k; z_k = m} \left(\pi_{i_k j_k}^{(z_k)}\right)^{y_k} \left(1 - \pi_{i_k j_k}^{(z_k)}\right)^{1 - y_k}.$$

To generate samples for each function f_m, we use an under-relaxed proposal mechanism in a Metropolis-Hastings algorithm Neal (1998), as this allows us to update each function as a block and reduce the computational complexity.

The code to implement this algorithm is available at `github.com/123anon/gpbt`. The time taken to execute the MCMC algorithm largely depends on the number of comparisons K. The computational and memory requirements for the GPs are negligible, since their size only depends on the number of objects N.

4. Deprivation in England

We now carry out a simulation study based on real deprivation data, as this allows us to better explore the applicability of the GP-BT model. The UK Ministry for Housing, Communities and Local Government publish an Index of Multiple Deprivation (IMD) for each of the 317 local authority areas in England McLennan et al. (2019). In this example, we generate simulated comparative judgement data sets of various sizes using the IMD for each area. To mimic real world data collection, we assume a judge takes 20 seconds to make a single comparison, which equates to 180 comparisons per hour. To represent fieldwork lasting $\{1, 2, 5, 10, 20, 30\}$ hours, we simulate $\{180, 360, 900, 1800, 3600, 5400\}$ comparisons. To generate the comparisons, we simulate from model (3) and choose pairs of areas uniformly at random from the list of all possible pairs. We fit the GP-BT

Algorithm 1 MCMC Algorithm for the GP-BT Model

Initialise the chain with values $f_1^{(0)}, \ldots, f_M^{(0)}, \boldsymbol{\alpha}^{2(0)}\boldsymbol{p}^{(0)}$ and $\boldsymbol{z}^{(0)}$
On iteration i of the MCMC algorithm do
for $m \leftarrow 1, M$ **do**
  Propose $f'_m = \sqrt{1-\delta^2} f_m^{(i)} + \delta\nu_m$, where $\nu_m \sim \mathcal{GP}(0,\, k(\cdot,\cdot;\alpha_m^{2(i)}, l_m))$
  Accept with probability $p_{acc} = \frac{\pi(\boldsymbol{y}|\boldsymbol{z}^{(i)},\boldsymbol{p}^{(i)},f_1^{(i+1)},\ldots,f'_m,\ldots f_M^{(i)})}{\pi(\boldsymbol{y}|\boldsymbol{z}^{(i)},\boldsymbol{p}^{(i)},f_1^{(i+1)},\ldots,f_m^{(i)},\ldots,f_M^{(i)})}$
end for
for $m \leftarrow 1, M$ **do**
  Sample $\alpha_m^{2(i+1)} \mid \boldsymbol{f}_m \sim \text{inv-}\Gamma\left(\psi_0 + \frac{n_m}{2},\, \omega_0 + \frac{1}{2}\boldsymbol{f}_m^{(i)^T}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X};1,l_m)^{-1}\boldsymbol{f}_m^{(i)}\right)$
end for
for $m \leftarrow 1, M$ **do**
  Sample $p_m^{(i+1)}$ from $\pi(p_m^{(i+1)} \mid \boldsymbol{z}^{(i)})$ using a Gibbs step
end for
for $k \leftarrow 1, K$ **do**
  for $m \leftarrow 1, M$ **do**
    Compute $\pi(z_k^{(i+1)} = m \mid y_k, \boldsymbol{p}^{(i+1)}, f_1^{(i+1)}, \ldots, f_M^{(i+1)})$
  end for
  Sample $z_k^{(i+1)}$ according to these weights
end for

model to the data sets modelling the patterns in deprivation at three spatial resolutions. We set the length scale parameters to 12, 50, and 100km, as these are the 1^{st}, 10^{th} and 25^{th} percentiles of the distances between each pair of local authority areas. This will help us identify trends within cities, across urban conurbations and at regional levels. We run the MCMC algorithm for 1,000,000 iterations, removing the first 500,000 as a burn-in period. Model fit is computed via the mean absolute error

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

where $\hat{\lambda}_i$ is the estimated deprivation in area i; specifically the MLE in the standard BT model and the posterior median in the other models. We compare the GP-BT model to the standard BT model, for which we use the `BradleyTerry2` R package Turner and Firth (2012). For small data sets that do not feature all the areas, we cannot compute MLEs for the standard BT model and the corresponding MAE is undefined. The results are shown in Table 1.

Judge hours	# Comparisons	Standard BT	GP-BT
1	180	–	**0.709**
2	360	–	**0.684**
5	900	–	**0.610**
10	1,800	1.562	**0.485**
20	3,600	0.409	**0.341**
30	5,400	0.369	**0.311**

Table 1: MAE for the standard BT and GP-BT models for the six English local authority area data sets. The model with the smallest error is shown in bold.

Results show that use of the GP-BT model significantly reduce the amount of data we require to achieve comparable results to the standard model. For example, in order to achieve an MAE of around 0.35 with the standard model we would need to collect 5,400 comparisons - but with the GP-BT model we can reduce this by a third, to only 3,600 comparisons. This is equivalent to reducing the fieldwork from 30 judge hours to 20 judges hours, making logistical costs of employing the GP-BT model cheaper, quicker and easier than the standard model.

In Figure 1 we show the results for the GP-BT model with 5,400 comparisons. The long length scale GP shows a north-south divide, with the south being generally less deprived than the north of the country. The medium length scale GP locates large conurbations, for example urban areas around Manchester, Liverpool and Newcastle. The short length scale GP identifies patterns in individual cities, for example areas in London or around Birmingham. We are able to accurately estimate the deprivation levels by combining the GPs. The standard BT model is more computationally simpler, taking 5 minutes compared to 5 hours. However, the burden of having to collect vastly more data for the standard BT model to achieve comparable accuracy to GP-BT is a considerable weakness, and besides, the 5 hour run time of the GP-BT model is tolerable in practice.

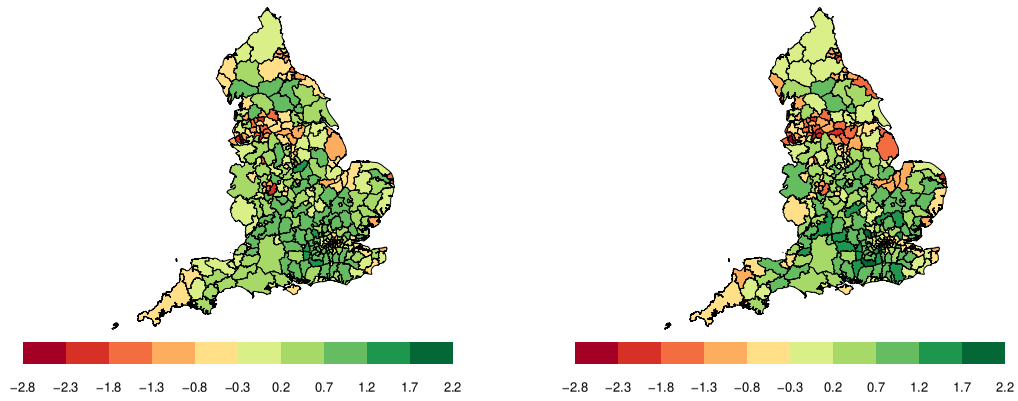
5. Deprivation in Dar es Salaam, Tanzania

We now demonstrate the effectiveness of the GP-BT model via a real-world case study. Our data, collected for this study, consists of 75,078 pairwise comparisons of subwards in the city of Dar es Salaam, Tanzania, a city with approximately 6.5 million people. The data for this project is in the `BSBT` R package (Seymour et al., 2022a) and was described in Seymour et al. (2022b). The city has almost doubled in size in the last 10 years United Nations Department of Economic and Social Affairs (2019), and the majority of citizens live in informal residences Limbumba and Ngware (2016). This has left official statistics concerning poverty rapidly out of date. With household surveying being logistically challenging and prohibitively expensive, the use of comparative judgement offers a potentially valuable alternative.

Data for the analysis was collected *in situ* over the course of two weeks in August 2018. 172 judges, all residents of the city itself, were recruited to take part in the project. Dar es Salaam consists of 452 subwards, and to estimate the deprivation of each, judges were shown pairs of subwards via a web interface. Subwards presented for comparison to each participant were chosen uniformly at random from sets of subwards that the participant had indicated prior familiarity with.

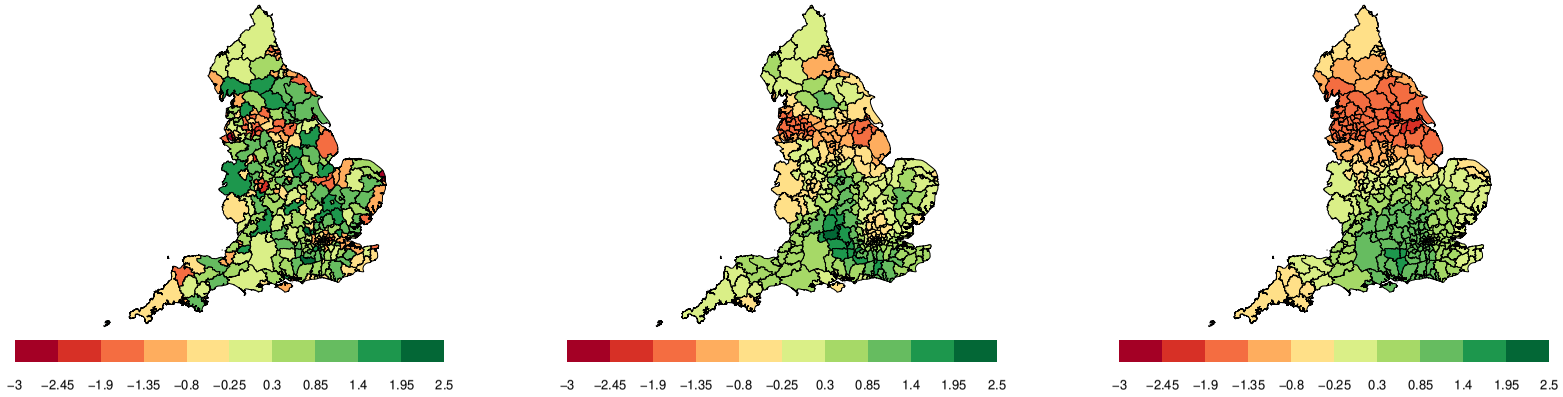
We run the MCMC algorithm for 1,500,000 iterations, removing the first 500,000 as a burn-in period. Using all 75,078 comparisons, this takes around 10 hours to run. The results are shown in Figure 2. The long length scale component identifies a city wide trend of higher affluence in the north of the city and towards the coast, with affluence decreasing further inland and towards the south of the city. The medium length scale function highlights several interesting areas in the centre of the city, which have extremely low levels of affluence, these areas are both slum areas with very low quality housing and little access to utilities. Immediately next to one slum area is the university, which the model seemingly correctly predicts is an affluent area. The medium length scale element of our model is flexible enough to capture the difference between these extreme areas. The short length scale function acts as a correction element, taking account of both the long and medium length functions.

The estimates for $\boldsymbol{z}$, $\boldsymbol{p}$ and $\boldsymbol{\alpha}^2$ have little physical interpretation and the posterior estimates for the weights $\boldsymbol{p}$ and $\boldsymbol{\alpha}^2$ are given in table 2. The latent variables $\boldsymbol{z}$ correlate highly



(a) The true IMD values for local authority areas in England.

(b) The posterior median IMD using 5,400 comparative judgements.



(c) The posterior median of the short length scale GP.

(d) The posterior median of the medium length scale GP.

(e) The posterior median of the long length scale GP.

Figure 1: The results for the English local authority area simulation study with 5,400 comparisons. Figures (c)-(e) highlight the informative nature of having access to different spatial resolutions, with the UK north-south divide clearly visible in (e), but with district variations being interpretable from (d)

Parameter	Posterior Median	95% Credible Interval
p_{short}	0.398	(0.379, 0.415)
p_{medium}	0.218	(0.203, 0.233)
p_{long}	0.385	(0.402, 0.537)
α^2_{short}	0.502	(0.441, 0.574)
α^2_{medium}	0.857	(0.674, 1.05)
α^2_{long}	1.20	(0.976, 1.44)

Table 2: Estimates and credible intervals for the weights $\boldsymbol{p}$ and variance hyperparameters $\boldsymbol{\alpha}^2$.

with the estimates for $\boldsymbol{p}$.

6. Conclusion

In this paper we present a novel extension to the Bradley–Terry model for analysing comparative judgement data, allowing the incorporation of knowledge or assumptions about a very general spatial structure in the qualities of objects. The Bayesian nonparametric framework that we present allows us to build flexible spatial correlations into the model, letting the model learn about a subward from data on its neighbours, reducing the amount of fieldwork required. Our model's leveraging of spatial correlation can not only recover ground-truth qualities of objects, but provide more interpretable and relevant analysis for non-scientific practitioners. Our model is far more data efficient that the standard model, making our model attractive to practitioners working in unstable or difficult environments. Also, by decomposing the deprivation into processes on different length scales allows results from the GP-BT model can be interpreted in a more meaningful way.

There are a number of possible directions in which our model may be fruitfully extended and further explored. Another way to make the results interpretable would be to cluster objects both spatially and by quality. A Bayesian nonparametric method which would be suited to this is the Chinese Restaurant Process Pitman (2006). In the context of cities, we could cluster nearby subwards into neighbourhoods by deprivation. Another issue which is particularly pertinent for data of the type we consider is that of judge reliability; i.e. the possibility of identifying the extent to which different people making judgements are consistent with each other. There are natural ways to extend the present work to enable inference for model selection and judge reliability using data augmentation methods.

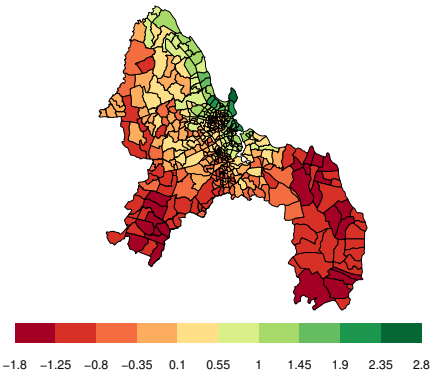
7. Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council [grant references EP/T003928/1 and EP/R513283/1].

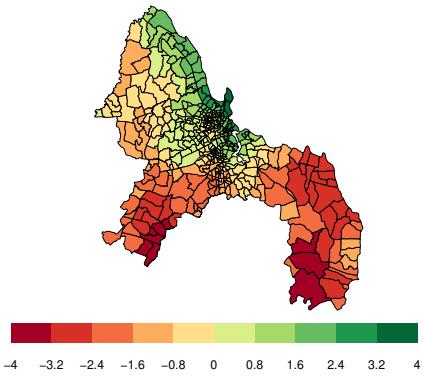
References

Adams, E. S. (2005). Bayesian analysis of linear dominance hierarchies. *Animal Behaviour*, 69(5):1191 – 1201.

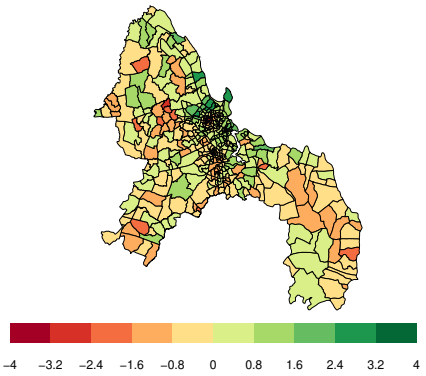
Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.



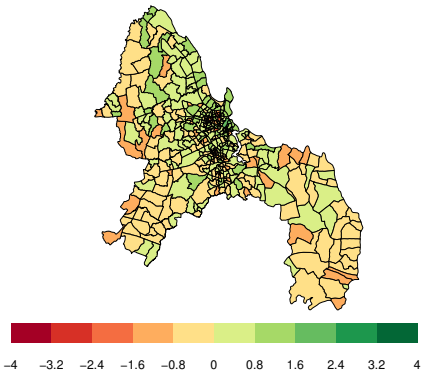
(a) Posterior median estimates for deprivation in each subward of Dar es Salaam



(b) Posterior median estimates for the long length scale function.



(c) Posterior median estimates for the medium length scale function.



(d) Posterior median estimates for the short length scale function.

Figure 2: The posterior median estimates for overall deprivation; and for the long, medium and short length scale components.

Caron, F. and Doucet, A. (2012). Efficient Bayesian inference for generalized Bradley–Terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196.

Cattelan, M., Varin, C., and Firth, D. (2012). Dynamic Bradley-Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C*, 62(1):135–150.

Chu, W. and Ghahramani, Z. (2005). Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*. ACM Press.

Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC.

Guiver, J. and Snelson, E. (2009). Bayesian inference for Plackett-Luce ranking models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 377–384, New York, NY, USA. ACM.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406.

Limbumba, T. M. and Ngware, N. (2016). Informal Housing Options and Locations for Poor Urban Dwellers in Dar es Salaam City. *The Journal of Social Sciences Research*, 2(5):93–99.

McLennan, D., Noble, S., Noble, M., Plunkett, E., Wright, G., and Gutacker, N. (2019). The English indices of deprivation 2019. Technical report, Ministry of Housing, Communities and Local Government, London, UK.

Neal, R. (1998). Regression and classification using Gaussian process priors. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 6*. Oxford Univeristy Press.

Phelan, G. C. and Whelan, J. T. (2018). Hierarchical Bayesian Bradley-Terry for applications in Major League Baseball. *Mathematics for Applications*, 7(1):71–84.

Pitman, J. (2006). *Combinatorial stochastic processes : Ecole d'été de probabilités de Saint-Flour XXXII, 2002*. Springer, Berlin New York.

Pollit, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19:281–300.

Radlinski, F. and Joachims, T. (2007). Active exploration for learning rankings from clickthrough data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 570–579.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts.

Seymour, R., Briant, J., and Zhang, Y. (2022a). *BSBT: The Bayesian Spatial Bradley–Terry Model*. R package version 1.2.1.

Seymour, R. G., Nyarko-Agyei, A., McCabe, H. R., Severn, K., Kypraios, T., Sirl, D., and Taylor, A. (2023). Comparative judgement modeling to map forced marriage at local levels.

Seymour, R. G., Sirl, D., Preston, S. P., Dryden, I. L., Ellis, M. J. A., Perrat, B., and Goulding, J. (2022b). The Bayesian Spatial Bradley–Terry Model: Urban Deprivation Modelling in Tanzania. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(2):288–308.

Springall, A. (1973). Response Surface Fitting Using a Generalization of the Bradley-Terry Paired Comparison Model. *Journal of the Royal Statistical Society Series C*, 22(1):59–68.

Stern, S. E. (2011). Moderated paired comparisons: a generalized Bradley-Terry model for continuous data using a discontinuous penalized likelihood function. *Journal of the Royal Statistical Society: Series C*, 60(3):397–415.

Stuart-Fox, D. M., Firth, D., Moussalli, A., and Whiting, M. J. (2006). Multiple signals in chameleon contests: designing and analysing animal contests as a tournament. *Animal Behaviour*, 71(6):1263–1271.

Turner, H. and Firth, D. (2012). Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software, Articles*, 48(9):1–21.

United Nations Department of Economic and Social Affairs (2019). *World urbanization prospects: the 2018 revision*. United Nations, New York.