Adam Howes
Email: ath19@ic.ac.uk

August, 2024

Dear Dr Paciorek and Dr Sykulski,

Thank you for taking the time to read and examine my thesis “Bayesian spatio-temporal methods for small-area estimation of HIV indicators”. I attach my response to the minor corrections you have suggested.

With kind regards,

Adam Howes

Thesis corrections for "Bayesian spatio-temporal methods for small-area estimation of HIV indicators"

Adam Howes (ath19@ic.ac.uk)

Contents

1 Dr. Christopher Paciorek

Thank you for the thorough discussion of the thesis. Both during the defense and in the provided corrections. I have addressed your suggested corrections point by point, as follows.

1.1 General comments

1.1.1 Chapter 4

> I'd like to see some more context relating the potential shortcomings to the public health setting (you have a bit of this in Section 4.1.3). For a public health analyst, when might they be most concerned about using the Besag model? What kinds of areal arrangements/neighborhood structures/types of data might be most prone to concern? E.g., one might be concerned about cases like Canadian provinces where their populations are so concentrated right near neighboring US states and most of the provincial area is sparsely populated.

In Section 4.1.3. I have modified the text as follows:

> The Besag model was originally proposed by Besag, York, and Mollié (1991) for use in image analysis. In this setting, areas correspond to pixels arranged in a regular lattice structure. In an image, the data point at each pixel can be thought of as an average of the intensity or colour over the space represented by the pixel.
>
> Since it's original proposal, the Besag model has seen wider use. However, for small-area estimation of HIV, the spatial structure corresponds to administrative units. These administrative units may have a more irregular spatial structure than a lattice. Furthermore, data points may not come about by uniform averaging over a space. For example, population density may vary across the area.

> And as we discussed, I'd like for you to see if you can drill down into the localized results of the simulation to give some insight into where the smoothing is sub-optimal in the simulations. Relatedly would you expect the features highlighted by your vignettes to occur in reality in public health settings?

Not yet resolved.

1.1.2 Chapter 5

> I'd like to see the chapter initially clearly lay out the overall goal, the quantitative representation of that, the various pieces of the analysis and how they fit together, and the data available, as well as what components you can estimate uncertainty for. In particular the notion of "reaching" the population needs to be clearly spelled out initially. And as we discussed, please make clear how prevalence is needed.

I have updated Section 5.1 giving the background for Chapter 5 as follows:

> In this chapter, I used a Bayesian spatio-temporal model (Section 5.3) of behavioural data from household surveys (Section 5.2) to estimate HIV risk group proportions. To then estimate risk group specific HIV prevalence and HIV incidences (Section 5.4), I combined the proportion estimates with population size, HIV prevalence and HIV incidence estimates, as well as risk group specific HIV incidence rate ratios, and HIV prevalence rate ratios. Finally, by ordering district, age, risk group strata by HIV incidence, I estimated an upper bound for the number of new HIV infections that could be averted under different risk prioritisation strategies (Section 5.4.3).

Model 5.11 omits various interactions. Focusing on the category-area-age interaction, which seems like the omitted interaction most likely to have substantial variation in reality, some effort to come up with some "residual" type diagnostic to assess model mis-specification in this regard would be helpful (e.g., perhaps some sort of variogram type analysis of some sort of age-group specific "working residuals" to borrow a GLM framing). Or you mentioned fitting the model with the interaction for one country. If that is not too burdensome that would also be a reasonable approach here.

The linear predictor used for the multinomial logistic regression model was

$$\eta_{ita} = \theta_{ita} + \beta_k + \zeta_{c[i]k} + \alpha_{ac[i]k} + u_{ik} + \gamma_{tk}.$$

This equation does contain age-country-category interactions $\alpha_{ac[i]k}$, but you are right to point out that age-district-category interactions are omitted. A model containing the effects α_{aik} is likely to cause computational difficulties.

The distinct differences between the CPO and information criteria (IC) results (and the very structured pattern in the surprising IC results) suggest the possibility of a bug somewhere, as we discussed. Getting the observation-specific values from INLA might help to better understand this.

I agree.

1.1.3 Chapter 6

Chapter 6 extends standard INLA computation in two ways. For the first, I'd like to see more clarity in how this differs from the Stringer et al. (2022) approach (i.e., that you go beyond the Gaussian mixture over the quadrature points, as we discussed in the defense) and the details of the software implementation (e.g., giving an overview in the chapter describing what someone would need to do to make use of your code/approach).

The Stringer, Brown, and Stafford (2022) approach is that of Section 6.1.3.1. It is similar to

the `inla::inla` with `method = "gaussian"`. The novel approach implemented in Section 6.2 is similar to `inla::inla` with `method = "laplace"`.

I have clarified this point in the following text:

> First, a universally applicable implementation of INLA with Laplace marginals, where automatic differentiation via `TMB` is used to obtain the derivatives required for the Laplace approximation. For users of `R-INLA`, the Stringer, Brown, and Stafford (2022) approach is analogous to `method = "gaussian"`, while the approach newly implemented in this chapter is analogous to `method = "laplace"`. Section 6.2 demonstrates the implementation using two examples, one compatible with `R-INLA` and one incompatible.

As we discussed in the defense, I'm concerned about any case where one draws from marginals, implicitly assuming no dependence, either at the hyperparameter level or the latent process level, and then does inference on a derived quantity that depends on more than one input. You should be clear anytime you do this that this is problematic (and try to avoid as much as possible).

Not yet resolved.

Relatedly, assuming I'm understanding correctly, there is an important tradeoff between using Laplace marginals for improved accuracy for latent marginals and using the Gaussian mixture over the quadrature points, which allows one to make draws and do inference on any derived quantity in a way that takes account of posterior dependence between and amongst hyperparameters and latent process values. If that's the case, I think it's worth pointing this out and discussing when one can use the Laplace marginals in a public health context and when one might need to use the Gaussian mixture.

This is an interesting observation. If you're correct, then I agree it worth pointing out, and it would shift my preference towards using the mixture of Gaussians approach. The way in which I think that it might not be correct is that the INLA approach fundamentally targets marginals, even when using the mixture of Gaussians. That is, the mixture of Gaussians does not "take account of posterior dependence between and amongst hyperparameters and latent process values" as suggested. I think this in part because there is work extending INLA to the joint setting – which I don't think would need to exist if this question were correct. To follow up here.

1.2 Minor comments

1.2.1 Chapter 2

4: "develop into a stage" -> "Infection with HIV can"

Changed to:

> If untreated, infection with HIV can develop into a more advanced stage known as acquired immunodeficiency syndrome (AIDS).

6: "to result a reduction"

Changed to:

> found complete surgical removal of the foreskin to result in a reduction

11: "Both DHS and PHIA surveys collecting"

Changed to:

> Both DHS and PHIA surveys collect demographic, behavioural, and clinical information

12: individual disclosure: error may come from them not knowing status

Good point! I have added the sentence:

> Furthermore, individuals may be unaware of their HIV status.

14: "UNAIDS process"

This stub has been been corrected as follows:

> Indeed, careful validation of data by country teams is a crucial part of the yearly UNAIDS HIV estimates process.

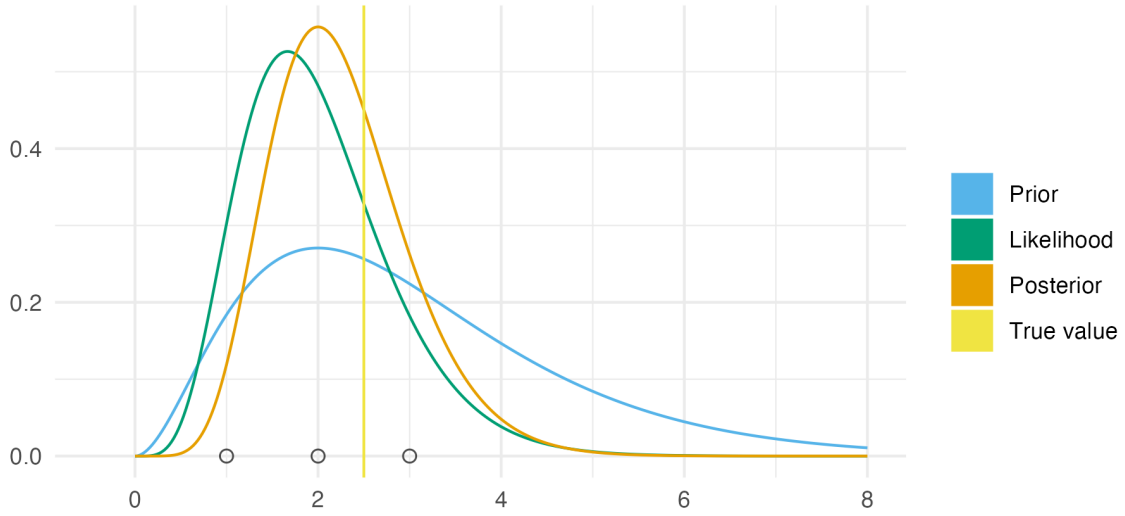
1.2.2 Chapter 3

16 (and elsewhere): Please look up usage of "that" vs. "which" so you can join me in the grammar police. "Models which do not produce" -> "Models that do not produce"

Thank you for the pointer! I have fixed this issue and look forward to joining the good fight.

Fig 3.1: I suggest that you also show the likelihood.

I have now included the likelihood in Figure 3.1 as follows. Figure 6.1, demonstrating the Laplace approximation, has also been updated to include the likelihood.



> 17: Beyond just p(y) even if you know the full form of p(phi|y) what do you do with it in non-trivial dimensions? You have to be able to either draw from it or estimate expectations of interest. the issue is rather broader than just the unknown normalizing constant.

Good point! I have added the sentence:

> Further, even given a closed form expression for the posterior distribution, if ϕ is of moderate to high dimension, then it is not obvious how to evaluate expressions of interest, which usually themselves are integrals, or expectations, with respect to the posterior distribution.

> 19. You haven't defined 'convergence' when you dive into diagnostics.

Good point! I have altered the text to read:

> After running an MCMC sampler, it is important that diagnostic checks are used to evaluate whether the Markov chain has reached its stationary distribution. If so, the Markov chain is said to have converged, and its samples may be used to compute posterior quantities. Though it is possible to check poor convergence in some cases, we may never be sure that a Markov chain has converged, and thus that results computed from MCMC will be accurate.

21. I'd frame this as deterministic approximations need to focus on approximating expectations of interest. I think of Laplace as approximating an integral over part of parameter space (often > random effects') to be able to work with a smaller-dimensional space, such as for maximization.

Thank you for the comment. In Chapter 6, I refer to a Laplace approximation over part of the parameter space as the marginal Laplace approximation.

23. "data is" -> "data are" (also p 66 and perhaps elsewhere)

I have corrected to "data are" in this instance and elsewhere in the thesis.

30. You distinguish ELGM from LGM with having defined eta for LGM or been explicit about 1:1 relationship of x and y.

Good point. In an LGM, it is that there is a one-to-one relationship between $\mathbf{y}$ and $\boldsymbol{\eta}$. I have added the sentence:

> In an LGM, like the more general GLMM case as given in Equation (3.6), there is a one-to-one correspondence between observations y_i and elements of the linear predictor η_i.

Sec 3.4: worth commenting on additivity of these measures that treat each obs as a unit of information given you are in a spatial setting.

Good point. I have added the paragraph:

> Equation (3.11) is additive, and treats each observation as an independent unit of information. Special care is therefore required in applying cross-validation techniques to dependent (Section 3.2.1.2) spatio-temporal data. For example, Bürkner, Gabry, and Vehtari (2020) and Cooper et al. (2024) use "leave-future-out" (LFO) cross-validation in the time-series context. Similarly, in Chapter 4 I apply a spatial-leave-one-out (SLOO) cross-validation scheme.

35. (3.30) should be for π_{2hj}.

Thank you for spotting this! Corrected.

37. "difficultly"

Corrected to:

> problem difficulty

37. "arrived at using by"

Corrected to:

> arrived at by estimating the variance of

1.2.3 Chapter 4

41. Might be worth including the variance piece raised to the power ‘n-c‘.

I have added the factor $\tau_u^{\frac{n-n_c}{2}}$ to Equation 4.3. Initially I excluded this factor as the primary intention was to demonstrate that $p(\mathbf{u})$ is a function of the pairwise differences and thus improper.

41. "recommended against": passive, and by who?

I have altered the text to read:

> Directly using the Besag model as described in Section 4.1.1 has several practical limitations in applied settings. To overcome these limitations, Freni-Sterrantino, Ventrucci, and Rue (2018) recommend three best practices:

42. Why is unit variance correct?

Yes that's a good point, what I mean to say is that the singletons have unit variance in the "structure matrix" sense. I have corrected the text to read that $p(u_i) \sim \mathcal{N}(0, \tau_u^{-1})$.

47. τ_v and τ_w are not orthogonal - what does this mean?

I use this terminology to mean that the posteriors for τ_v and τ_w are likely to be correlated. See Figure C.8 and Figure C.9 in Appendix C for an example, showing that the BYM2 parameterisation overcomes this issue. This appears to be inline with use of the term by e.g. Cox and Reid (1987).

48. Is convolution the right term here?

I have altered the text to use the term "convolved random effects" following use of this terminology by Morris et al. (2019).

55: what is meant by "model is implemented in arealutils"?

I have updated the text to read:

in the `arealutils` R package (Howes 2023)

> 56: need citation for v being hard to estimate

Although I found references (Zhang 2004; Williams and Rasmussen 2006; Anderes 2010; Karvonen and Oates 2023) supporting difficulties estimating one or multiple hyperparameters for spatial models (including using the Matérn kernel) I did not find a reference to directly support ν being difficult to estimate. For example, Williams and Rasmussen (2006) note that above $\nu = 7/2$ (i.e. quite smooth) then it's difficult to estimate ν from "finite noisy training samples". Hence I have updated the text to read:

> We fixed the smoothness hyperparameter ν to 3/2 to avoid concerns regarding the joint identifiability of the smoothness and lengthscale hyperparameters.

My intuition is that the observations in Chapter 4 being from a binomial distribution further supports that ν would be difficult to estimate.

> 56: have Li vary with size?

The L_i did not vary with the size of the area. A fixed number were used for each area.

> 56: effect -> affect

Change made, thank you!

> 57: "and the calibration"

I have altered the text to read:

> and the probability integral transform (PIT; Dawid (1984)) values

> 57: What parameter is shown in Figs 4.7-4.9 - it's not clear you're assessing the latent process values. And in that case you should be clear the CRPS is averaged over locations.

Good points! I have altered these three captions. As an example, for Figure 4.7 the caption now reads:

> The mean CRPS in estimating ρ_i and its standard error for each inferential model and simulation model on the grid geometry (Panel 4.6E). The mean value averages over both areas and simulation runs.

> 5 9: Explain that mean CRPS is mean over the simulations.

As above, this is a good point! I have added the sentence:

The mean values are an average over both the number of areas in each geometry and the number of simulations run.

63: Table 4.4 has no standard errors.

(Adding standard errors to Table 4.4. would cause it not to fit onto the page.)

64: "resulted wide"

Changed to:

resulted in wide

64: surprisingly

Changed, thank you!

67: "This chapter used of area-level models to for point-level data throughout". I can't parse this. You can only use point level model if have point level data.

I have altered the sentence as follows:

This chapter used area-level models to for data which arises by aggregation of point-level data.

67: "measures are disaggregated by area" - not sure of the point here.

I have altered the text to clarify the point:

Additionally, the measures used in this study were computed and presented by individual area. With refinements to the sample sizes used, these area-specific measures of performance could enable more nuanced conclusions about the use of spatial random effect models.

1.2.4 Chapter 5

71: FSW is not defined in Table 1 caption.

Changed to "female sex workers (FSW)".

71: In Table 1 why does High risk group IRR not vary with local incidence?

The conceptual model underpinning this decision was to frame the "High" risk group as a part of the general population at higher risk and the "Very High" risk group as a concentrated

epidemic subpopulation. In reality IRR is likely to vary by local incidence for the "Very High" risk group as well, and as such this is a fair critique. Practically speaking, the ALPHA network data used to inform the "High" IRR has quite little geographic variation.

> 71: Purpose of Table 1 is not clear. Nor how IRR is to be used.

I have included an additional clause that the IRR is "used to calculate HIV incidence" (by risk group).

> Tables sometimes appear earlier than they should (e.g., 5.1 and 5.2).

Not yet resolved.

> 77: Table 5.2: ϕ_{ik} should be u_{ik}.

Good spot! Thank you, fixed.

> 80: Mention country-specific vs single models earlier.

Not yet resolved.

> 82: I would say clearly that model structure for q_{ia} is discussed next.

I have altered the text to read:

> As all such surveys occurred in the years 2013-2018 (Figure 5.2) I assumed no dependence on time, hence omission of the index t. Model specification for the linear predictor η_{ia} is discussed in Section 5.3.2.1 to follow.

> 85: First paragraph of 5.3.3 is a bit hard to follow.

I have altered the text to read:

> Domain experts do not consider having had sex "in return for gifts, cash or anything else in the past 12 months" sufficient to constitute sex work. For this reason, I adjusted the estimates obtained based on the transactional sex survey question to match alternatively obtained age-country FSW population size estimates. Taking this approach retained subnational variation informed by the transactional sex survey question.

> 86: The bio-marker survey data and disaggregation model is unclear. How are risk groups known for individuals in the survey?

I have clarified this model as follows:

> To disaggregate HIV prevalence, I began by estimating HIV prevalence log odds ratios $\log(\text{OR}_k)$ relative to the general population. To do so, I began by calculating age, country, and risk group specific (as well as general population) HIV prevalence ρ_{cak} using bio-marker survey data from all 46 surveys included in the risk group model (Section 5.2.1).. I then fit a logistic regression model, with indicator functions for each risk group, and an indicator for being in the general population. The fitted regression coefficients in this model β_k correspond to log odds $\log \rho_k - \log(1 - \rho_k)$. The required log odds ratios may then be easily obtained by taking the difference in odds ratios.

88: Section 5.4.3 is hard to understand. I don't understand how it relates to 5.4.2. "Reach" is not clearly defined nor is it clearly discussed how it is quantified based on the various modeling pieces.

Section 5.4.3 relates to Section 5.4.2 because it uses I_{iak} which is calculated in Section 5.4.2. I have rewritten Section 5.4.3 to make this clearer and address these comments.

91: Not clear what the quantities are in the statement about "in most districts adolescent girls aged 15-19 were not sexually active". Is this an across-district or within-district quantity?

You're right that this sentence didn't make sense. I've rewritten it to clarify as follows:

> In the median district, 57.9% of adolescent girls 15-19 were not sexually active (95% credible interval [CI] at the district-level 27.7–79.7).

95: does the approach presented allow identification of actual people or just targeting efforts to reach more such people collectively

The approach does not allow identification of actual people. This is an important limitation of this analysis, which I discuss in Section 5.6.1 in the paragraph starting "The efficiency of each stratified prevention strategy depends on the ability of programmes to identify and effectively reach those in each strata...".

96: "Accounting for the 0% of new infections"?

Good spot, thank you! I have edited this sentence to remove this mistaken figure:

> My analysis focused on females aged 15-29 years, and could be extended to consider optimisation of prevention more broadly, accounting for new infections among adults 15-49 which occur in women 30-49 and men 15-49.

1.2.5 Chapter 6

> 106: Not sure what you mean by "log p(y|x,theta) is small". This is the likelihood...

I based this sentence on Blangiardo et al. (2013), which I have now cited.

> 116: "in which, which"

Changed to:

> in which, similar to extended latent Gaussian models

> 122: "Method" in Table 6.1 a bit terse.

I have updated this column to be more specific about the first word referring to "latent field marginals" and the second being "over the hyperparameters" (apart from for NUTS, which is over the whole space).

> 122: Is "Gaussian, EB" the same as frequentist Laplace approx (up to hyperparameter prior)? If so, probably worth saying.

Yes I believe it is, but I am not an expert in frequentist procedures.

> 130: Somewhat unclear how the quadrature is implemented, wrapped around the TMB-based Laplace approximation. Is your code in R? (Sorry, this may be because I didn't have time to look through appendices.)

Yes my code is in R. All code for the thesis is available at `athowes/thesis` and for the analysis in Chapter 6 at `athowes/naomi-aghq`.

> 131: Using same number of iterations with stan (full posterior, including latent values) vs tmbstan (hyperparameters, much lower-dimensional space) seems odd.

Here I am using `tmbstan` with the default option `laplace = FALSE`. Hence the `tmbstan` sampler is operating over the full space, just as the `rstan` sampler. As such, it's reasonable to use the sample number of iterations with both samplers.

> 132: Fig 6.7 is just grid/AGHQ, not EB? If so, why present EB method?

Yes that's correct, and I have updated the caption of Figure 6.7 to make this more clear. (The focus of this section of the chapter was to validate that my approach for implementing Laplace marginals is correct (as compared to `R-INLA`). For that purpose, I focused on presenting

results for the methods with quadrature. Though you do raise a good point that the results for EB should also be similar. Part of the reason to include EB methods was for teaching purposes.)

> 132: Why surprising tmbstan faster than rstan - what are the different computations involved - having to compute Laplace vs doing HMC over higher dimensional space. I expect it would vary with I expect it would vary with hyperparameter and latent dimensions.

It is surprising as they are running the same algorithm: HMC over the full space.

> 136: "kridge" -> "krige"

Changed to `gstat::krige`.

> 137: "this" in "this difference" is unclear.

I have altered the text to read:

> As β_ϕ was fixed then differences in approximation accuracy between the Gaussian and Laplace approximations of $\phi(s)$ are due only to differences in estimation of $u(s)$.

> 143: "survey weighting increases variance" - what about effect of increasing precision in small strata? Are you talking about influence of complex survey design or somehow about weighting scheme?

I believe the point is that use of any weighting scheme (aside from all weights being equal) reduces the "effective sample size" of the data, thus all else equal increasing the variance of estimates.

> 149: INLA uses CCD for d>2, right? Would this not work for this setting?

Yes, `R-INLA` does use central composite design (CCD) for integration over moderate dimensions. I mention that `R-INLA` uses CCD for $m > 2$ in Section 6.1.4.1, illustrate CCD in Figure 6.4, and mention that it would be of interest to compare PCA-AGHQ to CCD in Section 6.6.3.1.

> 151: (6.97) has 'd' instead of 'm'

Agree! Fixed by swapping d to m.

> 154: "closet"

Changed to:

> posterior contraction was very close to zero.

154: Did you use MAP for theta when looking at Hessian eigenvalues?

I considered the spectral decomposition of the matrix $\hat{\mathbf{H}}_{\texttt{LA}}^{-1}$. I have updated the text to replace an incorrect usage of $\hat{\mathbf{H}}_{\texttt{LA}}(\boldsymbol{\theta}_{\texttt{LA}})^{-1}$, thank you for spotting this.

156: "Figure ??"

Fixed, thank you.

156: "far fewer than full 24" - is this a problem?

This is a problem in the sense that it would ideal for the quadrature nodes used to show some variability in all 24 dimensions. PCA-AGHQ does improve upon a naive product grid, but is still far from ideal.

156: "point estimates" "distributional quantities" - need "and"

Fixed, thank you.

157: Need caption to describe the green

Thank you, I have updated this caption to read:

> The grey histograms show the 24 hyperparameter marginal distributions obtained with NUTS. The green lines indicate the position of the 6561 PCA-AGHQ nodes projected onto each hyperparameter marginal. For some hyperparameters, the PCA-AGHQ nodes vary over the domain of the posterior marginal distribution, while for others they concentrate at the mode.

165: What went wrong with tmbstan?

I imagine that this is in relation to "Preliminary testing of this approach, using `tmbstan` and setting `laplace = TRUE`, did not show immediate success but likely could be worked on.". I do not know exactly what was going on, and think that further study of the use of the Laplace approximation within MCMC routines like HMC could be a fruitful direction.

2 Dr. Adam Sykulski

Thank you for providing a paper copy of the thesis annotated with suggested typographical changes. I have made these changes, and additionally thoroughly proofread the thesis as requested.

References

Anderes, Ethan. 2010. "On the Consistent Separation of Scale and Variance for Gaussian Random Fields."

Besag, Julian, Jeremy York, and Annie Mollié. 1991. "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics* 43 (1): 1–20.

Blangiardo, Marta, Michela Cameletti, Gianluca Baio, and Håvard Rue. 2013. "Spatial and spatio-temporal models with R-INLA." *Spatial and Spatio-Temporal Epidemiology* 4: 33–49.

Cox, David Roxbee, and Nancy Reid. 1987. "Parameter Orthogonality and Approximate Conditional Inference." *Journal of the Royal Statistical Society: Series B (Methodological)* 49 (1): 1–18.

Dawid, A Philip. 1984. "Present position and potential developments: Some personal views statistical theory the prequential approach." *Journal of the Royal Statistical Society: Series A (General)* 147 (2): 278–90.

Freni-Sterrantino, Anna, Massimo Ventrucci, and Håvard Rue. 2018. "A note on intrinsic conditional autoregressive models for disconnected graphs." *Spatial and Spatio-Temporal Epidemiology* 26: 25–34.

Howes, Adam. 2023. *arealutils: Utility functions for beyond-borders*.

Karvonen, Toni, and Chris J. Oates. 2023. "Maximum Likelihood Estimation in Gaussian Process Regression Is Ill-Posed." *Journal of Machine Learning Research* 24 (120): 1–47. http://jmlr.org/papers/v24/22-1153.html.

Morris, Mitzi, Katherine Wheeler-Martin, Dan Simpson, Stephen J. Mooney, Andrew Gelman, and Charles DiMaggio. 2019. "Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan." *Spatial and Spatio-Temporal Epidemiology* 31: 100301. https://doi.org/https://doi.org/10.1016/j.sste.2019.100301.

Stringer, Alex, Patrick Brown, and Jamie Stafford. 2022. "Fast, scalable approximations to posterior distributions in extended latent Gaussian models." *Journal of Computational and Graphical Statistics*, 1–15.

Williams, Christopher KI, and Carl Edward Rasmussen. 2006. *Gaussian Processes for Machine Learning*. Vol. 2. 3. MIT press Cambridge, MA.

Zhang, Hao. 2004. "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics." *Journal of the American Statistical Association* 99 (465): 250–61.