Some thoughts on Hanks et al, *Environmetrics*, 2015, pp. 243-254.

Jim Hodges

Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota USA 55414
email: hodge003@umn.edu
October 13, 2015

It seems to me we disagree about two things. I'll talk about the big one first because whatever you think about it will determine what you think about the small one.

1 What does the random effect $\boldsymbol{\eta}$ mean?

I'm puzzled by this paper's interpretation of the spatial random effect $\boldsymbol{\eta}$. It seems to me the spatial random effect $\boldsymbol{\eta}$ is something *we choose* as a tool of data analysis and while this paper clearly sees $\boldsymbol{\eta}$ in that light, it also seems to treat $\boldsymbol{\eta}$, with equal conviction, as *being or emulating* the mechanism out there in the world that produced the data. These two roles of the random effect are quite different. Now you all obviously recognize this distinction, e.g., p. 253 line 3: "when the generating mechanism for spatially autocorrelated observations is assumed to be a spatially smooth missing covariate". It seems, however, that really embracing this distinction makes it hard to rationalize the SGLMM as a data-generating mechanism, in particular to take $\boldsymbol{\eta}$ literally as representing a component of the data-generating mechanism, out there in the world, with a distinct, indivisible existence, that deserves equal consideration with an explicit explanatory variable $\mathbf{X}$, rather than being just an error term, an expedient to absorb lack of fit of the fixed-effect part of the model.

Another way to say this (I think) is to recall that spatial analyses like this are necessarily observational, not experimental. As in any observational study, a spatial analysis faces the danger of mis-attributing variation in $\mathbf{y}$ to the explanatory variables $\mathbf{X}$ that we happen to know about and have in our dataset, when in fact some or all of that variation should, in a causal sense, be attributed to other explanatory variables that we don't have or don't know we should consider. An idea that seems common in spatial analyses — and I think it's the key idea here — is that simply because the data are spatially referenced, we can avoid this inherent problem of observational studies and adjust for missing covariates by including in the model the spatial random effect $\boldsymbol{\eta}$. We rationalize this belief by theory and simulations in which the data are generated by a mechanism that happens to be identical to the analytical tool (the spatial random effect) we propose to use to adjust for the missing covariates.

To the contrary, I think it's necessary to be excruciatingly clear about the distinction

between the analytic tool we choose for an analysis and the mechanism by which the data were generated. It hadn't occurred to me until now, but it seems this distinction doesn't matter if the purpose of a spatial analysis is prediction or interpolation, but it very much does matter when the purpose is to learn something about the data-generating mechanism, in particular about the importance of explicit predictors $\mathbf{X}$. Because clarity is important here, I hope you'll forgive me for being repetitive below about the data-generating mechanism (model) and the analysis model.

Focusing for now on the data-generating mechanism (model), I would ask: when you use a random effect $\boldsymbol{\eta}$ as a data-generating mechanism, what do you intend it to represent? In this paper, all I recall seeing on this subject is the line quoted above ("when the generating mechanism for spatially autocorrelated observations is assumed to be a spatially smooth missing covariate"), though I might have missed something. So for the moment let's allow two possibilities: in the data-generating model, the random effect $\boldsymbol{\eta}$ is a way to represent missing covariates with spatial structure; or in the data-generating model, the random effect $\boldsymbol{\eta}$ represents something else that induces near locations to be more similar than far. I don't know what "something else" could be but perhaps if we understand "missing covariates" in a limited sense as meaning columns in some design matrix, we may think there are processes too complicated to reduce to missing covariates in this sense. However, regarding this second possibility I would suggest that, as for a GP, once you've observed data at specific locations, such a process is simpler than the machinery makes it look, so that this second possibility can, in analyzing a given dataset, be reduced to missing covariates. Therefore, I'll assume that in the data-generating model, a spatial random effect $\boldsymbol{\eta}$ can *only* represent covariates with spatial structure that we don't have or don't know about, which are therefore missing from the point of view of the analysis model. [This is the crux of my argument. If you can make a compelling case that the data-generating mechanism can contain a spatially-correlated random effect that cannot be reduced to covariates that are unknown or unavailable and are thus necessarily missing from the analysis model, then my argument falls apart. But I don't think such a case can be made.]

If we start here, then, I suggest it is no longer meaningful to say that an analysis model that restricts the random effect $\boldsymbol{\eta}$ is misspecified while an analysis model that does not restrict $\boldsymbol{\eta}$ is correctly specified. Rather, *any* model is misspecified in that it does not include the correct explanatory variables: it can't, because we don't have them all, and I will soon argue that putting a spatial random effect in the analysis model can't fix that. It would seem to follow that asking whether restricted spatial regression (RSR) recovers the parameter $\boldsymbol{\beta}$ from a data-generating model that includes a spatial random effect $\boldsymbol{\eta}$ is not a pertinent question, because data-generating mechanisms don't have spatial random effects, only analysis models do. A more pertinent question would be how the fits of different analysis models behave

when the data are produced by different data-generating mechanisms (models), specifically different scenarios of unavailable or unknown covariates, which are missing covariariates from the point of view of the analysis model.

With this in mind, here are some scenarios I find it useful to consider, along with what theory or simulations show about them. The first two scenarios are unrealistic, but the third is fully realistic and the unrealistic ones help us understand it.

First, though, I need some set-up. I assume the analysis model has one explicit explanatory variable in a column $\mathbf{X}$, and a spatially-correlated random effect, and an iid normal error term, so the theory is tractable and exact. Second, I'll talk about a conventional analysis with two steps, one in which the covariance parameters are estimated by maximizing the restricted likelihood (RL) and a second in which the mean-structure parameter β is estimated by plugging the maximum-RL estimates into the covariance model. It is crucial to note that confounding has very different effects in these two stages, as I'll describe below. (Note also that a Bayesian analysis does not change anything important here.)

Here are the scenarios.

1. Besides the covariate of interest $\mathbf{X}$, the data-generating mechanism has just one covariate $\mathbf{H}_X$ that we don't know about (so from the point of view of the analysis model, it's a missing covariate). Suppose further that $\mathbf{H}_X$ has correlation exactly 1 with $\mathbf{X}$. The restricted likelihood (RL) attributes to $\mathbf{X}$ *all* of $\mathbf{y}$'s variation arising from $\mathbf{H}_X$, so the RL-maximizing variance of the random effect is very close to zero. (For ICAR random effects, this is in my book, Section 15.2.3. I'm also implicitly relying on an arm-waving extention to other spatial random effects.) When β is estimated, even though the analysis model includes the spatial random effect, the estimate of β is biased because all of $\mathbf{H}_X$'s contribution to $\mathbf{y}$ is mistakenly attributed to $\mathbf{X}$.

2. Now suppose the data-generating mechanism has several covariates that we don't know about (missing covariates, from the point of view of the analysis model). Let them be the columns in a matrix $\mathbf{H}_{notX}$ and suppose further that $\mathbf{X}'\mathbf{H}_{notX} = \mathbf{0}$, i.e., $\mathbf{H}_{notX}$ is orthogonal to $\mathbf{X}$. In this scenario, the data $\mathbf{y}$ have variation with spatial structure that is orthogonal to $\mathbf{X}$, so the RL-maximizing variance of the random effect is positive. Thus, in estimating β, some variation in $\mathbf{X}$'s direction is attributed to the fitted random effect because that random effect is not shrunk to zero in the direction of $\mathbf{X}$. In other words, the analysis model's random effect introduces a spurious confounder, corresponding to nothing in the data-generation model, that steals some of the variation in $\mathbf{y}$ that properly belongs to $\mathbf{X}$, and it does so because of variation in $\mathbf{y}$ that's orthogonal to $\mathbf{X}$. I'm pretty sure this is what's happening in Section 5.2's simulation.

3. Finally, suppose the data-generating mechanism combines the first two scenarios: the

data-generating mechanism includes one covariate perfectly correlated with $\mathbf{X}$, in a column $\mathbf{H}_X$, and other covariates orthogonal to $\mathbf{X}$, which are columns in a matrix $\mathbf{H}_{notX}$. In the RL, all variation in $\mathbf{X}$'s direction is attributed to $\mathbf{X}$, but because of the variation in $\mathbf{y}$ in the space spanned by $\mathbf{H}_{notX}$, the RL-maximizing estimate of the random effect's variance is positive. Thus, when estimating the fixed effect β, some of the variation in $\mathbf{y}$ in the direction of $\mathbf{X}$ is attributed to the random effect. The *amount* of variation in $\mathbf{y}$ in $\mathbf{X}$'s direction that the analysis model attributes to the random effect is determined entirely by the amount of variation in $\mathbf{y}$ in the space spanned by $\mathbf{H}_{notX}$: if $\mathbf{y}$'s variation in this subspace is large, the random effect variance is has a large estimate and the random effect is strongly confounded with $\mathbf{X}$; if $\mathbf{y}$'s variation in this subspace is small, the random-effect variance is has a small estimate and the random effect is only weakly confounded with $\mathbf{X}$. In other words, the bias in estimating β is large or small and positive or negative depending on how the data-generating mechanism produces variation in $\mathbf{y}$ in the subspace *orthogonal to* $\mathbf{X}$.

Scenario 3 is realistic: If the data-generating mechanism has any unknown or unobserved covariates (missing covariates, from the point of view of the analysis model) that are correlated with $\mathbf{X}$, then you can WLOG re-parameterize the data-generating model so it has one covariate that's perfectly correlated with $\mathbf{X}$ with other covariates orthogonal to $\mathbf{X}$, and then Scenario 3 applies. I would conclude, therefore, that putting a spatial random effect in the analysis model to adjust for covariates in the data-generation model that are unavailable or unknown, and thus missing from the analysis model, will adjust $\mathbf{X}$'s coefficient haphazardly; in other words, adding a spatially-correlated random effect to the analysis model cannot fix this fundamental problem of observational studies. You are absolutely right that if the data-generating model includes unknown or unavailable covariates that are confounded with $\mathbf{X}$, an RSR fit will give a biased estimate of $\mathbf{X}$'s coefficient β. But an SGLMM fit doesn't fix that; it just gives an estimate with a different and haphazardly determined bias. We cannot know what that bias is, any more than we can with an RSR fit, because we either don't have or don't know the needed (missing) covariates that are part of the data-generation model.

Should we nonetheless prefer the estimate produced by the SGLMM? I would say "no". To attribute to $\mathbf{X}$ all the variation in $\mathbf{y}$ that lies in the space spanned by $\mathbf{X}$, as in RSR, is not a strong assumption; it's just a recognition that it is impossible to adjust for unknown or unavailable covariates, and we mislead ourselves by trying to do it.

2 The spurious (or is it?) association between x and $\boldsymbol{\eta}$

One consequence of the foregoing is a different interpretation of your very cool results in Section 3 about the association between the explicit covariate $\mathbf{x}$ and the random effect

$\boldsymbol{\eta}$. "When the spatial random effect $\boldsymbol{\eta}$ is spatially smooth and has a large effective range of spatial autocorrelation" (p. 243), this means $\boldsymbol{\eta}$'s covariance matrix is dominated by a very few eigenvectors with large eigenvalues. The association between $\mathbf{x}$ and $\boldsymbol{\eta}$, when their range parameters are large, arises from that dominance because as their respective ranges get large, the j^{th} dominant eigenvectors in $\mathbf{x}$ and $\boldsymbol{\eta}$ are highly correlated with each other. Indeed, because $\mathbf{x}$ and $\boldsymbol{\eta}$ are observed at the same locations, their respective eigenvectors are identical if $\mathbf{x}$ and $\boldsymbol{\eta}$ have the same range and smoothness parameter.

The intriguing thing here is that the independence of $\mathbf{x}$ and $\boldsymbol{\eta}$ implies that for any n-vector $\mathbf{c}$, $\mathbf{c}'\mathbf{x}$ and $\mathbf{c}'\boldsymbol{\eta}$ have correlation zero, but because when their range parameters are large their covariance matrices are dominated by a few eigenvectors, any *realizations* of $\mathbf{x}$ and $\boldsymbol{\eta}$ necessarily have a correlation that's large in absolute value. As the ranges become large, for both $\mathbf{x}$ and $\boldsymbol{\eta}$ the probability mass is concentrated around their respective first eigenvectors, so $\mathbf{x}$ and $\boldsymbol{\eta}$ are positively correlated with probability about 1/2 and negatively correlated with probability about 1/2. Thus your results have to be in terms of $\mathrm{E}(R^2)$, not the expected correlation, which is necessarily zero.

Following my argument above, I don't see this artifactual association between $\mathbf{x}$ and $\boldsymbol{\eta}$ as a flaw of RSR. Rather, I see it as an inherent drawback of using a spatial random effect in the analysis model as an expedient for absorbing lack of fit to the fixed effects. When $\boldsymbol{\eta}$'s range is large, $\boldsymbol{\eta}$ is *inherently* indiscriminately highly correlated with covariates $\mathbf{X}$ that vary smoothly in space, even if the data-generation model has no unknown or unavailable covariates corresponding to the dominant spatially-smooth eigenvectors of $\boldsymbol{\eta}$'s covariance matrix. This problem is negligible for fixed effects $\mathbf{X}$ that have fine-scale or no spatial structure because the corresponding eigenvectors of $\boldsymbol{\eta}$'s covariance matrix have small eigenvalues, so unless $\boldsymbol{\eta}$'s variance is estimated to be huge, these components of $\boldsymbol{\eta}$ will be shrunk near zero and thus cannot be confounded, to any degree that matters, with such an $\mathbf{X}$.

In this view, RSR is not defective but simply recognizes (I apologize for repeating myself) that it is impossible to adjust for unknown or unavailable covariates by putting a spatial random effect in the analysis model, and we're better off if we don't try.