Random Effects Old and New

James S. Hodges[1*], **Murray K. Clayton**[2]

[1]Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota USA 55414
[2]Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin USA 53706
[*]*email:* hodge003@umn.edu

February 2, 2011

SUMMARY

The term “random effect” is now used much more broadly than it was, say, 50 years ago. At that time, a random effect was an effect having (in analysis-of-variance jargon) levels that were draws from a population, and the draws were not of interest in themselves but only as samples from the larger population (e.g., Scheffé 1959, p. 238). By contrast, new-style random effects have levels that are not draws from any population, or that are the entire population, or that may be a sample but a new draw from the random effect could not conceivably be drawn; and the levels themselves are usually of interest. All such new-style random effects can be understood as formal devices to facilitate smoothing or shrinkage, interpreting those terms broadly. The distinction between old- and new-style random effects is not a mere nicety but has practical consequences for inference and prediction, simulation experiments for evaluating statistical methods, and interpretation of analytical artifacts. Therefore, this distinction should be developed beyond the catalog of examples and consequences given here and incorporated into statistical theory and practice.

Key Words: fixed effect, random effect, shrinkage, simulation, smoothing, spatial correlation

1 Introduction

The term "random effect" has come to be used more broadly than it was, say, 50 years ago. Many things that are now called random effects differ qualitatively from random effects as defined in, for example, Scheffé's classic 1959 text on analysis of variance. This distinction between old- and new-style random effects has important consequences, conceptual and practical, but few statisticians seem to recognize it and we do not have language to describe it. The present paper's purpose is to clarify this distinction, propose appropriate language for it, and describe its practical consequences. At least one prominent statistician has declared that the term "random effect" is now ill-defined to the point of being meaningless, so that he does not use it and would abolish it (Gelman 2005a, Section 6; Gelman 2005b). We have some sympathy with this view, but it would abolish not only the term "random effect" but also the distinction between old- and new-style random effects, which, we argue, would be a mistake.

Our argument, briefly, is this: The traditional, old-style definition of a random effect (Section 2) is quite explicit and specific. However, as analysis of richly-parameterized models becomes more unified in, for example, the mixed linear model framework (e.g., Ruppert, Wand, and Carroll 2003), more analyses use the *form* of old-style random-effects, but many, perhaps most, such random effects do not meet the traditional definition. Section 3 describes how new-style random effects fail to meet the old-style definition and gives three examples that together capture the sense of most if not all new-style random effects. Section 4 argues that these deviations from the old-style definition are not mere niceties but imply distinct ways to do inference and prediction, to do simulations for evaluating statistical methods, and to interpret analytical artifacts that turn up in applications of statistical methods involving random effects. In statistical theory and practice, therefore, we should identify whether a given random effect is the old or new kind and proceed accordingly. It may, therefore, ultimately be desirable to follow Gelman's advice partly, by abolishing the term "random effect" in favor of two new terms, one referring to each of old- and new-style random effects.

2 Old-style random effects

The traditional definition of a random effect is given, for example, in Scheffé (1959, p. 238): the levels of a random effect are draws from a population, and the draws are not of interest in themselves but only as samples from the larger population, which *is* of interest. (We have modified Scheffé's definition a bit; we will return to this below.) "Level" is ANOVA jargon that we use to highlight the difference between new- and old-style random effects. In this traditional definition, the population can be real and finite or hypothetical and infinite. Our example has a real finite population; an example with a hypothetical infinite population (e.g., samples of dental materials) would add nothing for the present purpose.

Example 1: Measuring epidermal nerve density. Dr. William Kennedy's neurology lab at the University of Minnesota uses novel methods to count and measure nerve fibers in skin and mucosa, as an objective way to measure the status and progress of neuropathies, especially among diabetics. A recent study (Panoutsopoulou et al 2009) compared two skin sampling methods using 25 "normal" (not diabetic) subjects, from each of whom skin was sampled by each of two methods, biopsy and blister, on the calf and on the foot. The analysis included three old-style random effects: subjects, the method-by-subject interaction ("method" being blister or biopsy), and the location-by-subject interaction ("location" being foot or calf). A fourth random effect, the method-by-location-by-subject interaction, is customarily treated as the residual error term because it cannot be identified without replicate measurements, and is deemed a variance component but not a random effect.

Using the notation of Scheffé (1959, Section 8.2), let factor A be methods, with levels $i = 1$, 2; factor B be location, with levels $j = 1$, 2; and factor C be subjects, with levels $k = 1, \ldots, 25$. Then the model representing this experimental situation is

$$\begin{aligned} y_{ijk} = \mu \quad &+ \quad \alpha_i^A + \alpha_j^B + \alpha_{ij}^{AB} \\ &+ \quad a_k^C + a_{ik}^{AC} + a_{jk}^{BC} + \epsilon_{ijk} \end{aligned} \tag{1}$$

where the αs are fixed effects, the as are random effects, and ϵ_{ijk} is the error term. In Scheffé's parameterization, each fixed effect satisfies sum-to-zero constraints, so each α sums to zero across each of its subscripts. Each interaction of subjects with a fixed effect also sums to zero across its subscript referring to the fixed effect, so for example $\sum_i a_{ik}^{AC} = 0$ for each subject k. Each random effect is drawn from a distribution with mean 0. For the subject effect $\text{var}(a_k^C) = \sigma_C^2$, and for the subject-by-method interaction, $\text{var}(a_{1k}^{AC}) = \sigma_{AC}^2$ for each k, with $a_{2k}^{AC} = -a_{1k}^{AC}$ to satisfy the sum-to-zero constraint. The subject-by-location interaction's variance is defined analogously. (We have simplified Scheffé's specification for the random effects.)

These random effects arise from sampling subjects and describe how the average nerve density varies between subjects (subject main effect); how the difference in nerve density, blister minus biopsy, varies between subjects (subject-by-method interaction); and how the difference in nerve density, foot minus calf, varies between subjects (subject-by-location interaction). Like many, perhaps most, applications of old-style random effects, these subjects were not selected by a formal sampling mechanism. This is in contrast with Scheffé's definition of a random effect, in which the draws "can reasonably be regarded as a random sample from some population" (1959, p. 238). Nonetheless, these are old-style random effects: the levels (subjects) are a sample, albeit not a random sample; the subjects are not interesting in themselves, but only as members of the population from which they were drawn (non-diabetic adults); and the object was to measure average differences in that population between methods and locations.

3 New-style random effects

Before we start, it is useful to distinguish three senses in which the notion of probability is used in practice. One familiar sense of "probability" involves draws using a random mechanism, either one we create and control (e.g., sampling using random numbers) or one we imagine is out there in the world (e.g., random errors in measuring a length). This conforms more or less to the frequentist notion of probability, the fraction of times an event occurs in a hypothetical infinite sequence of draws. A second familiar sense of probability describes uncertainty that does not arise from operation of a random mechanism, but rather is a person's uncertainty about an unknown quantity or about the association between two quantities. This is the subjective Bayesian notion of probability; de Finetti's slogan "Probability does not exist" summarized his argument that random mechanisms do not exist, and that probability properly has *only* this second sense.

The third and least familiar sense of probability is that a probability distribution can be used as a descriptive device. (For an argument that a descriptive notion of exchangeability is more primitive than probability, see Draper et al 1993. For an argument that this descriptive sense of probability can be used to justify a form of inference, see Freedman & Lane 1983.) For example, if we measured the heights of all US-born 52-year-old males employed by the University of Minnesota, those heights could be well described as following a normal distribution with some mean μ and variance σ^2. Describing the heights this way does not imply that anyone's height is a random draw from N(μ, σ^2): each man's height is fixed (for now, at least). Rather, this is an aggregate statement describing the heights of a group of men. If we used a random-number generator to select one man from this group and report his height, the height we report could be represented as a draw from N(μ, σ^2) in the first sense of probability (random mechanism). However, the selected individual's height is fixed; we create the randomness by our method of selecting him.

We can now identify three varieties of new-style random effects, which we first list and then describe in detail with examples. We use the ANOVA term "levels" to emphasize how these random effects fail to meet Scheffé's definition and how foreign the old-style language seems when used to describe new-style random effects.

- The levels of the effect are not draws from a population because there is no population. The mathematical form of a random effect is used for convenience only.
- The levels of the effect come from a meaningful population but they are the whole population and these particular levels are of interest.
- A sample has been drawn, and the sample is modeled as levels of a random effect, but a new draw from that random effect could not conceivably be made, even if it made sense to imagine the random effect was drawn in the first place.

Rendered in ANOVA language, these new-style random effects are hard to recognize, but they will be recognized easily when described in contemporary language, below.

3.1 The levels of the effect are not draws from a population because there is no population. The mathematical form of a random effect is used for convenience only.

The mixed-model representation of penalized splines (Ruppert et al 2003) is an example of random effects with levels that are not draws from any conceivable population. In the simplest case of a one-dimensional penalized spline with a truncated-line basis, the levels of the random effect are the changes in the fit's slope at the knots, and the random effect's distribution is simply a device for penalizing the changes. The senselessness of imagining further draws of such random effects is clearest for the examples in Ruppert et al (2003) in which penalized splines are used to estimate smooth functions in the physical sciences. As Ruppert et al (2003) put it, "we used the mixed model formulation of penalized splines as a *convenient fiction* to estimate smoothing parameters. The mixed model is a reasonable (though not compelling) Bayesian prior for a smooth curve, and ... [maximized restricted-likelihood] estimates of variance components give estimates of the smoothing parameter that generally behave well" (p. 138; emphasis added). To make this concrete, consider an example that briefly recapitulates the mixed linear model formulation of penalized splines.

Example 2. Global mean surface temperature (GMST). Consider smoothing the annual series of global mean surface temperature deviations from 1881 to 2005, plotted in Figure 1's top panel and available at http://www.biostat.umn.edu/~hodges/GMST.txt. (This version of the series was downloaded in 2006; an up-to-date series and related series are available at NASA 2010.) Call the GMST measurements y_i, $i = 1, \ldots, 125$ (units are 0.01C) and let $\mathbf{y}' = (y_1, \ldots, y_{125})$. In the analyses presented here, the years are centered and scaled so they have finite-sample variance 1.

The object is to draw a smooth line through the data that captures its general shape. Penalized splines do this by, in effect, fitting a linear model with many right-hand-side variables and constraining their coefficients. For the right-hand-side variables, we use a truncated quadratic basis, so the fitted spline for year i has the form

$$\beta_0 + x_i\beta_1 + x_i^2\beta_2 + \sum_{j=1}^{30} u_j([x_i - \kappa_j]_+)^2 \qquad (2)$$

where x_i is year i after centering and scaling, $[z]_+ = 0$ if $z < 0$ and z otherwise, and the κ_j are 30 knots, the years 1880 + {4, 8, ..., 120}, centered and scaled in the same manner as the x_i. The spline is fit by choosing values for $(\boldsymbol{\beta}, \mathbf{u})$, where $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2)$ and $\mathbf{u}' = (u_1, \ldots, u_{30})$. We can write (2) in the form $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$, where the i^{th} row of $\mathbf{X}$ is $(1, x_i, x_i^2)$ and the i^{th} row of $\mathbf{Z}$ is $(([x_i - \kappa_1]_+)^2 \ldots, ([x_i - \kappa_{30}]_+)^2)$.

If we simply fit a conventional linear model $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$, it would give a wiggly fit, wildly over-fitting the data. Penalized splines avoid this by constraining $\mathbf{u}$. This is commonly done by selecting a positive semi-definite matrix $\mathbf{D}$ and then fitting the spline — selecting $(\boldsymbol{\beta}, \mathbf{u})$ — by solving this optimization problem:

$$\text{minimize} \;\; (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})) \;\; \text{subject to} \;\; \mathbf{u}'\mathbf{D}\mathbf{u} \leq K \qquad (3)$$

for some constant K. This optimization problem turns out to be equivalent to another optimization problem:

$$\text{minimize} \;\; (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})) + \alpha\mathbf{u}'\mathbf{D}\mathbf{u} \qquad (4)$$

where α is a function of K and can be chosen by any of several methods, e.g., cross-validation.

The optimization problem (4) can in turn be re-cast in the *form* of a random effects analysis. For simplicity, let $\mathbf{D}$ be the identity matrix, the default choice in Ruppert et al (2003). In this analysis, the vector of GMSTs $\mathbf{y}$ is modeled as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), u_j \overset{iid}{\sim} N(0, \tau^2), \tag{5}$$

and $(\boldsymbol{\beta}, \mathbf{u})$ is selected by solving the generalized least squares problem that arises from (5) when the variances σ^2 and τ^2 are fixed at particular values. The optimization problem (3) is thus formally identical to computing the usual estimate of $\boldsymbol{\beta}$ and the best linear unbiased predictors (BLUPs) of $\mathbf{u}$ for model (5) given σ^2 and τ^2. The penalized spline analysis now has the form of a random-effects analysis, and as Ruppert et al (2003) noted, we can use this convenient fiction as a source of tools for (implicitly) selecting the penalty K or α in (3) and (4) respectively, by using restricted likelihood to estimate the variances in (5).

Although the analysis of (5) has the *form* of a random effects analysis, nowhere is there any suggestion that $\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}$ is a draw from a random mechanism. Rather, $\mathbf{u}$, like $\boldsymbol{\beta}$, is simply unknown before the optimization problem is solved. The representation (5) provides a flexible collection of smooth curves and some discipline to avoid overfitting, but $\mathbf{u}$ is not a draw from anything. Even if we assume that a random process produced year i's true GMST and measurement y_i, it would be senseless to imagine making more draws from $\mathbf{u}$'s distribution, which is merely an analytical convenience, a device to facilitate smoothing. Indeed, the GMST series clearly does not fit a model of iid errors around a smooth curve (see, e.g., Adams et al 2000), but maximizing the restricted likelihood arising from (5) is just a way to pick α in (4) and thus an unobjectionable way to draw a smooth line through the data capturing its general shape, which was the object of this exercise. (After adopting this convenient fiction, however, Ruppert et al 2003 behave like conscientious statisticians, checking for heteroscedastic errors, non-linearity, etc.)

In many cases there cannot be *any* sense in which $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ arose from a random process, even once. Many examples in Ruppert et al (2003) come from the physical sciences, e.g, the LIDAR, Janka hardness, and NOx examples, in each of which the smooth function estimated by $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ is implied by physical or chemical laws. Even more plainly, we have used splines to approximate complicated deterministic functions, to simplify computation or presentations. In these cases, $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ may be considered a random variable in the Bayesian use of that term (the second sense of probability), but it did not come into being as a random draw.

Bayesian terminology is delicate here, making it is easy to fall into conceptual errors. In a Bayesian approach to a penalized spline analysis, Your uncertainty (as de Finetti put it) about $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ is described using a probability distribution, the second sense of "probability", but that distribution describes You, not $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$.

It is easy to see that that $\mathbf{u}$ is not an iid draw from $N(0, \tau^2)$ by fitting (5) to the GMST data and examining the BLUPs for $\mathbf{u}$. Figure 1's top panel shows the fit obtained using (5) and $\tau^2 = 947$ and $\sigma^2 = 145$, which maximize the restricted likelihood. The fixed effect estimates are $\hat{\boldsymbol{\beta}}^{'} = (85.35, 176.40, 68.40)$; Figure 1's bottom panel plots the BLUPs for $\mathbf{u}$. Runs of positive and negative signs as in this figure, while necessary to fit a curve like the one fit to the GMST data, are spectacularly improbable if $\mathbf{u}$ is truly iid normal with mean zero. This grates on our statistical sensibilities. However, iid is not the wrong model for $\mathbf{u}$; rather, when we use (5) as a device to fit a smooth curve, there is no reason to expect the fitted $\mathbf{u}$ to look like an iid sample.

3.2 The levels of the effect come from a meaningful population but they are the whole population and these particular levels are of interest.

A common example here is smoothing disease maps, in which a geographical entity is partitioned into smaller regions and some color-coded measure of a disease, e.g., incidence, is shown on the map for each region. If the disease is rare or many regions have small populations, a map showing each region's raw disease measure will be noisy and of little use. A smoothed map replaces each region's raw disease measure with a measure obtained by "borrowing strength" from neighboring regions. (We use quotes because of the value judgment implicit in this expression, but it is so well-established as to be *de rigueur.*) Smoothing is often implemented by means of a random effect: the regions are the random effect's levels, the collection of regions is the entire population, and individual regions are of interest. Such analyses often include regressors as fixed effects.

Example 3. Stomach cancer in Slovenia. Dr. Vesna Zadnik, a Slovenian epidemiologist, collected counts of stomach cancers in the 194 municipalities that partition Slovenia, for the seven years 1995 to 2001 inclusive. She was studying the possible association of stomach cancer with socioeconomic status, as measured by a composite score calculated from 1999 data by Slovenia's Institute of Macroeconomic Analysis and Development. (Zadnik & Reich 2006 gives her findings). As part of this analysis, she used the model of Besag, York, & Mollié (1991), in which the counts of stomach cancers O_i are conditionally independent Poisson random variables with mean

$$\log\{E(O_i)\} = \log(E_i) + \beta SEc_i + S_i + H_i. \tag{6}$$

E_i is the expected number of cancers, the rate of stomach cancers per person in all of Slovenia multiplied by the population of municipality i (indirect standardization). SEc_i is the centered measure of socioeconomic status. The intercept is the sum of two random effects, $\mathbf{S} = (S_1, \ldots, S_{194})'$ capturing spatial clustering and $\mathbf{H} = (H_1, \ldots, H_{194})'$ capturing heterogeneity. The H_i were modeled as independent draws from a normal distribution with mean zero and precision (reciprocal of variance) τ_h. The S_i were modeled using an L_2-norm improper conditionally autoregressive (ICAR) model, also called a Gaussian Markov random field, representing the intuition that neighboring municipalities tend to be more similar than municipalities that are far apart. This model for $\mathbf{S}$ can be represented as an improper n-variate normal distribution specified by its precision matrix,

$$p(\mathbf{S}|\tau_s) \propto \tau_s^{(n-G)/2} \exp(-0.5\tau_s \mathbf{S}'\mathbf{Q}\mathbf{S}), \tag{7}$$

where G is the number of islands (disconnected groups of municipalities) in the spatial map (Hodges, Carlin, & Fan 2003). The unknown τ_s controls the smoothness of $\mathbf{S}$, with larger τ_s forcing neighboring S_i to be more similar to each other. $\mathbf{Q}$ encodes the neighbor pairs, with diagonal elements q_{ii} = number of municipality i's neighbors, and $q_{ij} = -1$ if municipalities i and j are neighbors and 0 otherwise. For the Slovenian data, $G = 1$ and municipalities i and j were deemed neighbors if they shared a boundary.

In this example, the spatially-correlated random effect $\mathbf{S}$ does not meet Scheffe's definition of a random effect because the levels (municipalities) are the entire population and are themselves of interest. This differs from penalized splines in that the present example has a meaningful population, but as with splines, it is awkward to imagine $\mathbf{S}$ as a draw from a random mechanism. As for the GMST data, it is not hard to imagine that some process with a stochastic element produced $S_i + H_i$ and obviously counts of stomach cancers could be made for the following 7-year period. However, it would do serious violence to the subject matter to imagine or model these new counts as arising from an iid draw from the same mechanism that produced the counts for 1995-2001, or that it would even be possible to make a second draw from the same mechanism that produced the counts for 1995-2001. Those of a modeling inclination might model the *change* from 1995-2001 to the next 7-year period using a stochastic model and argue that the change from each 7-year period to the next could be considered an iid draw from some process. However, that has no relevance to Dr. Zadnik's problem, in which she had only the 7-year period 1995-2001. Thus it is hard to see how the random effect $\mathbf{S}$ can be usefully described as arising from a random mechanism, the first sense of probability, or how we could usefully imagine making a second draw of this random effect.

It seems less problematic to consider the random-effect model for $\mathbf{S}$ as an example of the third sense of probability, a descriptive device. Just as we could say that the heights of the 52-year-old US-born males at the University of Minnesota look like draws from a particular normal distribution, we could also say that the 194 Slovenian S_i, were we to observe them, would look like a draw from an ICAR with particular neighbor pairings. As with the heights of 52-year-old men, this would not imply that the S_i were actually drawn from any random mechanism; it would only be a convenient way to describe the ensemble of fixed but unknown S_i.

Using a probability distribution to describe $\mathbf{S}$ is decidedly less concrete than using a probability distribution to describe the heights of a group of men, which are at least in principle observable while $\mathbf{S}$ is necessarily unobservable. But the intuition that motivates use of a spatial model — that municipalities near each other tend to have more similar S_i than municipalities far from each other — is ultimately descriptive. Those of a subjective Bayesian turn of mind might say at this point that it is equally natural to think of $\mathbf{S}$'s random-effect distribution as a probability statement in the second sense, describing Your uncertainty about how the S_i tend to be similar to each other. We have no objection to this, but many people still find it awkward or impossible to think of probability statements as describing personal belief. Those of a conciliatory turn of mind might suggest that for the present problem, at least, there is no difference between the descriptive and subjective-Bayesian interpretations of this random effect. In neither of these senses of probability, however, could we imagine making another draw from $\mathbf{S}$'s distribution. If we view $\mathbf{S}$'s distribution as descriptive, we might choose to deploy that description as part of a statistical method, the operating characteristics of which we can study using simulations, for example. If we view $\mathbf{S}$'s distribution in the subjective-Bayesian manner, we would use it like any other probability statement in the

Bayesian calculus.

We have belabored this point because in the subculture of spatial analysis, at least some people tend to ignore the nature of a random effect and proceed directly to mathematical specifications and calculations, but as we argue below, the nature of the random effect has practical implications.

3.3 A sample has been drawn, and the sample is modeled as levels of a random effect, but a new draw from that random effect could not conceivably be made, even if it made sense to imagine the random effect was drawn in the first place.

Consider the classic problem of geostatistics, mineral exploration. Suppose that in the region being explored, we are interested in the fraction of iron in rock at a certain depth, called W. Each location $\mathbf{s}$ in the region has a particular true W, $W(\mathbf{s})$, which is fixed but unknown. $W(\mathbf{s})$ is not observed directly, but measured with error at a sample of locations $\{\mathbf{s}_i\}$, giving measurements $y(\mathbf{s}_i)$ which could be modeled as $y(\mathbf{s}_i) = W(\mathbf{s}_i) + error(\mathbf{s}_i)$, where the $error(\mathbf{s}_i)$ are independent of each other.

We could estimate $W(\mathbf{s})$ from the $y(\mathbf{s}_i)$ using a two-dimensional penalized spline, executing the analysis by giving $W(\mathbf{s})$ the *form* of a random effect. If we did, $W(\mathbf{s})$ would be another instance of the new-style random effect in Section 3.1. However, we want to make a different point that strikes closer to the heart of spatial statistics.

The present example, mineral exploration, differs from the global-mean surface temperature and Slovenian stomach-cancer examples in a few ways. The measurement locations $\mathbf{s}_i$ are a sample of possible locations (in practice rarely selected by a formal sampling mechanism, but no matter) and they are not of interest in themselves but for what they tell us about the region as a whole, which may be understood as the population from which the $\mathbf{s}_i$ were drawn. We could go back and draw a new sample of $\mathbf{s}_i$, and we could also make a new measurement at our original measurement locations $\mathbf{s}_i$, or at least very close to them.

So far this sounds like an old-style random effect, but it is not. If we were to take new measurements at old $\mathbf{s}_i$, or take measurements at new $\mathbf{s}_i$, it would be senseless to imagine that these would involve a new draw of the random effect $W(\mathbf{s})$. Rather, $W(\mathbf{s})$ is fixed but unknown. We have observed it (with error) at some $\mathbf{s}_i$ but its value is already determined at *all* $\mathbf{s}$, observed and unobserved. It may be convenient to *describe* $W(\mathbf{s})$ in aggregate by saying that at any finite set of locations $\{\mathbf{s}_i\}$, $\{W(\mathbf{s}_i)\}$ looks like a draw from a normal distribution with mean $\mu(\mathbf{s}_i)$ and covariance $C(\mathbf{s}_i, \mathbf{s}_j, \theta)$ with parameter θ, and we may understand this probability distribution in either the second (subjective Bayesian) or third (descriptive) senses. We may imagine that $W(\mathbf{s})$ came into being by some stochastic process, but it is now fixed and no more draws will be made from that stochastic process. In these respects, $W(\mathbf{s})$ is identical to $\mathbf{S}$ in the Slovenian stomach-cancer example.

3.4 Comments on new-style random effects

(a) With some exceptions noted in Section 3.1, one might argue that for all our examples of new-style random effects, when the true $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$ or $\mathbf{S}$ or $W(\mathbf{s})$ came into being out there in the world, it *did* involve draws from a random mechanism. In such cases, it may make sense to imagine that producing the now fixed but unknown $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$, $\mathbf{S}$, or $W(\mathbf{s})$ involved a random draw *on one occasion*, after which $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$, $\mathbf{S}$, or $W(\mathbf{s})$ became fixed. Having once made such a draw, however, it makes no sense to imagine further draws, and the specific value of that single draw is of intrinsic interest: estimating it is, in many if not all cases, the entire purpose of the analysis. It is therefore a hazardous distraction to conceive of $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$, $\mathbf{S}$, or $W(\mathbf{s})$ arising as a draw from a random mechanism and better to think of it as simply fixed and unknown.

(b) Instead of being a draw from a random mechanism, a new-style random effect is more clearly seen as something we *choose* as an analytic tool. We do not mean to suggest that the things we have labelled $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$, $\mathbf{S}$, or $W(\mathbf{s})$ do not have real physical existence; $W(\mathbf{s})$ certainly does, in our example, and $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$ and $\mathbf{S}$ do in at least some cases. Accordingly, some choices of analytical tools — random effect distributions — will be better suited than others to estimating the actual (but unknown) $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$, $\mathbf{S}$, or $W(\mathbf{s})$. But “better suited” here means they will yield better estimates, not that they are better characterizations of any stochastic process that might have produced the actual $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$, $\mathbf{S}$, or $W(\mathbf{s})$. In other words, the probabilistic form of our analytical tool does not correspond to a probabilistic mechanism in the world that produced $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$, $\mathbf{S}$, or $W(\mathbf{s})$.

(c) New-style random effects can all be understood as merely formal devices for implementing smoothing or shrinkage. This is explicitly the case for penalized splines, in which the random effect is a “convenient fiction”. It is less explicit for the kinds of spatial models in Sections 3.2 and 3.3, perhaps because people in our discipline habitually think of these models in terms of covariance matrices and thus draws from random mechanisms.

(d) Regarding the Slovenian stomach-cancer example, we have heard it said that $S_i + H_i$ is only in the model as an error term, in the same way as we include a simple error term in an ordinary linear regression. The same could be said of many problems like the mineral exploration example. But what is “error” here? We would argue that error, in these instances, is nothing more than local lack of fit, which in the Slovenian example we use to capture spatial clustering and heterogeneity that are not captured by the predictor of interest, *SEc*. This still does not provide a rationale for imagining that new-style random effects are anything other than formal devices for implementing shrinkage or smoothers. In particular, it seems strained to imagine that $S_i + H_i$ capture a few omitted covariates entering the true model linearly, and would be unnecessary if only we had these missing covariates.

(e) There are, of course, cases of spatial analysis that are properly conceived as involving old-style random effects. For example, suppose we are studying ozone in the Boston area, and have a dataset with measurements taken each day for several years. It may be meaningful, depending on the questions being asked, to represent days as draws from a distribution with spatial correlation. More such draws could be made and for many questions the draws themselves have no intrinsic interest, the interest lying instead in the population of days from which they were drawn.

Suppose however we are interested in a specific week because, for example, a particular kind of temperature inversion occurred that week. In this case, Boston has, for this week, a smooth spatial ozone gradient around which daily measurements varied. It violates the substantive question to treat that smooth spatial ozone gradient as a draw from a random mechanism. Rather, it is a fixed feature of Boston for the study's week, in which we have a specific interest but which we happen not to know. We may choose to make an aggregate description of this fixed but unknown feature of Boston using a probability distribution as a descriptive tool. There is a meaningful sense in which this fixed but unknown feature of Boston was drawn from a probability distribution, but that sense is not relevant to this particular question.

4 Practical consequences of the differences between old and new random effects

The distinction between old- and new-style random effects has implications for how we should do inference and prediction, design simulation experiments for evaluating statistical methods, and interpret analytical artifacts that turn up in uses of statistical methods involving random effects.

4.1 Inference and Prediction

For the present purpose, we use "inference" to refer to analyses focused on the present set of observations and on the models we posit as having generated them. We use "prediction" to refer to analyses focused on other observations related to the present set but as yet unobserved or unknown to us.

4.1.1 Inference

In a Bayesian analysis, all inferences are based solely on the posterior distribution of unknowns in the model or models being entertained. For such Bayesian computations, it makes no difference whether a random effect is old- or new-style.

By contrast, the old-style notion of a random effect is implicit in non-Bayesian terminology and concepts. Consider, for example, best linear unbiased predictions (BLUPs). Here, the term "prediction" is used, by an old convention, to refer to estimates of random effects, while this convention reserves "estimate" for fixed effects. The notion of unbiasedness in BLUP refers to the expectation over random effects as well as error terms. Computing an expectation over a random effect — outside of a Bayesian analysis — seems to require that repeated sampling of the random effect be meaningful. While it is defensible to imagine that new draws can be made from old-style random effects, the idea of new draws makes no sense for new-style random effects, as we have argued. Thus, it makes little sense to consider BLUPs unbiased for new-style random effects. It is well known, for example, that penalized splines — which are BLUPs in the mixed-model formulation — flatten peaks and fill valleys (e.g., Ruppert et al 2003, p. 141), as do spatial random-effect models.

These tendencies are plainly biases if we conceive of a new-style random effect as simply a tool for estimating ensembles of fixed but unknown quantities. Such an understanding was present in early work on shrinkage estimation, which was called "biased estimation" among other things, although that term has largely died out. We do not mean to impugn BLUPs, which are immensely useful (per GK Robinson's classic 1991 paper), but it does not make sense to consider them unbiased for new-style random effects.

The old-style notion of random effects is so deeply embedded in non-Bayesian theory that cataloging its effects is beyond the scope of this paper; one example will have to suffice. The question is how to compute pointwise confidence intervals for a penalized spline fit when the spline is implemented as a mixed linear model as in Section 3.1. We would argue that if the random effect is new-style, the confidence interval should be computed conditioning on — treating as fixed, although unknown — the realized value of the random effect $\mathbf{u}$, while if the random effect is old-style, the confidence interval should not condition on $\mathbf{u}$'s realized value. To make the discussion concrete, we follow the development in Ruppert et al (2003, Section 6.4) using the GMST example from Section 3.1. We do so not to single out these authors for criticism; rather, the exceptional clarity and care of their writing makes this an ideal platform for discussing this issue. (We have changed their notation slightly.)

To begin, Ruppert et al (2003) take a step beyond the model-free development of the penalized spline in Section 3.1 and assume the observations y_i follow the model

$$y_i = f(x_i) + \epsilon_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \epsilon_i \tag{8}$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ and $u_j \overset{iid}{\sim} N(0, \tau^2)$, reserving comment for now on what $\mathbf{u}$'s distribution means. The estimate of $f(x)$, $\hat{f}(x)$, is obtained by maximizing the resulting restricted likelihood in τ^2 and σ^2, computing the estimated BLUPs (EBLUPs) $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ at the estimates of τ^2 and σ^2, and computing $\hat{f}(x) = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}$. The question is how to construct pointwise confidence intervals to present with this estimate.

Ruppert et al (2003) note that "Variability estimates differ depending on whether randomness in $\mathbf{u}$ is taken into account. One argument ... is that randomness of $\mathbf{u}$ is a device used to model curvature, while $\boldsymbol{\epsilon}$ accounts for variability about the curve. According to this argument, variance calculations should be done with respect to the conditional distribution $\mathbf{y}|\mathbf{u}$ rather than the unconditional distribution of $\mathbf{y}$." Define $\mathbf{C} = [\mathbf{X}|\mathbf{Z}]$ and $\mathbf{C}_i = [\mathbf{X}_i|\mathbf{Z}_i]$, where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are the rows of $\mathbf{X}$ and $\mathbf{Z}$ corresponding to year i. Then

$$\mathrm{var}\{\hat{f}(x_i)|\mathbf{u}\} = \mathbf{C}_i \mathrm{cov}\left(\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} |\mathbf{u} \right) \mathbf{C}_i', \tag{9}$$

where $\mathrm{cov}([\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}]'|\mathbf{u})$ is a function of τ^2 and σ^2 but not a function of $\mathbf{u}$ despite the conditioning. (Odd but true; see Ruppert et al 2003, p. 139, equation 6.10.) Because $\hat{f}(x)|\mathbf{u}$ has a normal distribution with mean $E[\hat{f}(x)|\mathbf{u}]$ and variance (9), a 95% confidence interval can be constructed from this distribution in the usual way. This interval is centered at $E[\hat{f}(x)|\mathbf{u}]$, but we will call it the conditional confidence interval for $f(x)$. (As Ruppert et al note, (9) ignores variance arising from estimating τ^2 and σ^2, standard practice in conventional analyses of mixed linear models, but this is irrelevant for the present purpose.)

"If there is no appreciable bias", Ruppert et al continue, "then $E[\hat{f}(x)|\mathbf{u}] \approx f(x)$, and this interval can be interpreted as a confidence interval for $f(x)$". But the mixed-model framework allows an estimate of the bias conditional on $\mathbf{u}$, i.e., taking $\mathbf{u}$ as fixed though unknown, which for model (8) is easily shown to be

$$E[\hat{f}(x) - f(x)|\mathbf{u}] = -r\mathbf{C}(\mathbf{C}'\mathbf{C} + r\mathbf{I})^{-1}\begin{pmatrix}\mathbf{0}_3\\ \mathbf{u}\end{pmatrix} \tag{10}$$

where $\mathbf{0}_3$ is a 3-vector of zeroes and $r = \sigma^2/\tau^2$, with larger values implying a smoother fit. This bias is a function of x and non-zero in general. Thus, the conditional confidence interval is an interval for $E[\hat{f}(x)|\mathbf{u}]$, not for $f(x)$. Although Ruppert et al do not say it in so many words, the conditional confidence interval is not a proper confidence interval for $f(x)$ in general because its coverage is too low, and coverage is low because the interval has the wrong center.

Up to this point, Ruppert et al's development is consistent with $\mathbf{u}$'s nature as a new-style random effect, a convenient way to describe an ensemble of fixed but unknown constants. But now Ruppert et al make a choice that seems odd for a new-style random effect; we follow their development to its conclusion, then give a choice that seems more consistent with $\mathbf{u}$'s nature.

After Ruppert et al give (10), the bias of $\hat{f}(x)$ given $\mathbf{u}$, and note that it is non-zero in general, they observe "But, since $E(\mathbf{u}) = \mathbf{0}$, the unconditional bias is $E[\hat{f}(x) - f(x)] = 0$. Thus, on average over the distribution of $\mathbf{u}$, $\hat{f}(x)$ is unbiased for $f(x)$. To account for bias in the confidence intervals, the [conditional] variance $\text{var}\{\hat{f}(x)|\mathbf{u}\}$ should be replaced by the conditional mean-squared error $E[\{\hat{f}(x) - f(x)\}^2|\mathbf{u}]$ . . . and then averaged over the $\mathbf{u}$ distribution." They do so, obtaining

$$E[\{\hat{f}(x_i) - f(x_i)\}^2] = \mathbf{C}_i\text{cov}\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}} - \mathbf{u}\end{pmatrix}\mathbf{C}_i', \tag{11}$$

and suggest using a confidence interval centered at $\hat{f}(x)$ with upper and lower limits determined by (11). This new interval, which we call the unconditional confidence interval, is wider than the conditional interval arising from (9): $\mathbf{C}_i\text{cov}([\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}} - \mathbf{u}]')\mathbf{C}_i' > \mathbf{C}_i\text{cov}([\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}]'|\mathbf{u})\mathbf{C}_i'$, "because [(11)] accounts for both components of error (variance and squared bias) whereas [(9)] accounts only for variance and covers $E[\hat{f}(x)]$, not $f(x)$." Ruppert et al mention a result of Nychka (1988), that the unconditional pointwise confidence interval using (11) gives 95% coverage averaging over the values of the *predictor* (year, in the GMST example), but note that coverage is too low in areas of high curvature, where bias is greatest, and too high in areas of low curvature, where bias is negligible.

Now let us reconsider Ruppert et al's choice to account for bias in $\hat{f}(x)$ given $\mathbf{u}$ by inflating the variance of $\hat{f}(x)$ conditional on $\mathbf{u}$, specifically by adding squared bias and taking the expectation of this inflated variance with respect to $\mathbf{u}$'s distribution. If $\mathbf{u}$ is understood as a new-style random effect — where $\mathbf{u}$ is fixed but unknown and its distribution is just a device to induce smoothing — Ruppert et al's choice appears to be a *non sequitur*. The problem with the conditional interval for $f(x)$ based on (9) is not that it is too narrow, but that it is centered at the wrong value for each x. Widening the confidence interval does not solve the problem, as Ruppert et al note; the unconditional interval's coverage is too low for precisely those values of x where bias in $\hat{f}(x)$ is largest. Viewing $\mathbf{u}$ as a new-style random effect, the obvious way to fix this problem is not to inflate (9) but to center the confidence interval in the right place. The obvious (though not necessarily best) way to do this is by subtracting the estimated bias — (10) evaluated at $\hat{\mathbf{u}}$ — from the fit $\hat{f}(x)$ and constructing a confidence interval around this bias-corrected estimate using the conditional standard error derived from (9). (Sun & Loader 1994, while presenting methods for confidence intervals that were simultaneous, not pointwise, argued against this simple bias correction based on simulation results in their Section 5. Unfortunately, their Section 5 is so terse that we cannot determine how they simulated their data. If they simulated data by drawing a new value of the random effect for each simulated dataset – which, we argue below, cannot be justified for new-style random effects – then their case against the simple bias correction is in doubt.)

Figure 2 compares the various intervals using the GMST example. It shows the data (circles), the spline fit (the central solid line), and the three nominally 95% confidence intervals for a period with large curvature, 1925 to 1955. The simple conditional interval around the (biased) fit $\hat{f}$ is drawn using solid lines; the unconditional interval centered around $\hat{f}$ is the dashed lines; and the bias-corrected fit and the conditional interval around it are indicated by dotted lines. The bias-corrected interval is shifted up by a maximum of about 1.5 units compared to the simple unconditional interval, while the upper limit of the unconditional interval is shifted up by only about 0.3 units at most. This is consistent with the conditional interval's undercoverage in areas of greatest curvature, which the theoretical development leads us to expect. Because the bias-corrected interval addresses the bias problem directly, we would expect its pointwise coverage to be closer to 95% for all values of x and particularly in intervals of x with substantial curvature, where the unconditional interval is known to have low coverage. Hence, we should not need extraordinary expedients like averaging over x to get 95% coverage, i.e., using a global calibration for an interval that is interpreted pointwise.

We address one final observation to partisans of Bayesian analyses. The marginal posterior variance of $f(x)$ is the expectation of (9) with respect to the joint marginal posterior of $\boldsymbol{\beta}, \mathbf{u}, \tau^2$, and σ^2. In most cases this is larger than (9) because it accounts for uncertainty about $\boldsymbol{\beta}, \mathbf{u}, \tau^2$, and σ^2, though it is not necessarily larger than (9) or (11). Like the unconditional confidence interval, a Bayesian interval will tend to have high coverage in areas of low curvature and low coverage in areas of high curvature. This would be of no interest to someone holding the traditional Bayesian view that bias and interval coverage over repeated sampling are irrelevant, but it would concern those who like Bayesian intervals because they account for uncertainty about τ^2 and σ^2 (e.g., Ruppert et al 2003, p. 102).

4.1.2 Prediction

For old-style random effects, everyone distinguishes two cases, which we will describe in terms of Example 1, measuring epidermal nerve density. The first case is a prediction that involves drawing a new level of the random effect; in the epidermal nerve density example, this would be predicting biopsy and blister nerve density measurements from a new subject's calf and foot. Because a new subject is drawn, each new measurement's variance is the sum of all the components of variation: the three customarily called random effects (subjects, method-by-subject interaction, and location-by-subject interaction), and the one customarily called residual error. The second case is a prediction of a new measurement on a level of the random effect that has already been drawn. In the epidermal nerve density example, this would be a new skin sample from, say, the foot of an already-sampled subject. Because no new subjects are sampled, each new measurement's variance is simply the residual error variance. The proper measure of prediction uncertainty for this new measurement also accounts for uncertainty about the true nerve density of this specific subject's foot in the area being sampled, i.e., of a function of some fixed and random effects. But the only variance component that contributes to the new draw is residual error.

For *all* new-style random effects, predictions are like the latter case, if predictions are possible at all. For all such random effects, the device of a random effect is used to estimate an ensemble of quantities that are fixed but unknown, like the true foot nerve density in the subject from whom we are taking a new skin sample. For some problems, like the global-mean surface temperature data, predictions of new measurements may be impossible or ill-advised: it is not possible to make a new observation on the years 1881-2005, and using a spline fit to extrapolate past 2005 is a bad idea. However, when a new measurement can be made (e.g., some physical-science examples in Ruppert et al 2003), predictions about these new measurements are simply inferences about the fitted spline surface, as in Section 4.1.1, plus the variance component arising from residual error. In this respect, the new-style random effects in Sections 3.1 and 3.2 are the same.

Now consider the mineral-exploration example in Section 3.3. It is possible to take a second measure at an existing location $\mathbf{s}_i$ (or very close to it) or to take a measurement at a new location $\mathbf{s}_0$. In either case, the only interpretation that does not do violence to the subject matter is that the random effect $W(\mathbf{s}_i)$ or $W(\mathbf{s}_0)$ has already been drawn and is simply unknown; otherwise, making a prediction would involve re-drawing the process that produced the ore seam. Rather, the spatial process $W(\mathbf{s})$ simply facilitates spatial smoothing.

4.2 Simulation experiments to evaluate statistical methods

In a simulation experiment to evaluate methods involving old-style random effects, for example analyzing the epidermal nerve density data with model (1), simulated datasets must be drawn by first drawing the random effects for new subjects, a_k^C, then drawing the interaction random effects for each subject, a_{ik}^{AC} and a_{jk}^{BC}, then drawing a residual error ϵ_{ijk}, and finally adding these random effect draws to the fixed effects μ, α_i^A, α_j^B, and α_{ij}^{AB}. This reflects the way the real data are presumed to have been generated — by sampling subjects — and permits evaluation of the relevant aspects of the statistical methods, namely the behavior of estimates and intervals for fixed effects and variance components. This way of simulating data reflects the nature of the old-style random effects used to model the data.

The foregoing is not controversial. However, based on journal articles and communication with colleagues, many statisticians appear to believe that for simulation experiments evaluating a method involving *new-style* random effects, data should be simulated in the same manner, with repeated draws from the random effects. Views on this matter do not seem to correspond to any Bayes-frequentist divide; it seems to be the default position of most statisticians of both persuasions. We will argue, to the contrary, that in simulating data for experiments intended to evaluate methods involving new-style random effects, repeated draws should *not* be made from those new-style random effects; it may be inappropriate to use even one such draw.

We give two arguments, one from first principles and one pragmatic. The argument from first principles is that new-style random effects are convenient fictions used to implement smoothing and it is a mistake to take these fictions literally and draw from them. The question in simulation experiments intended to evaluate such methods is how well they capture features of the unknown function or surface to be estimated. The pragmatic argument is that in general, simulated draws from new-style random effect models do not behave the way our intuition, developed from *fitting* such models, would suggest. Therefore, simulating from a new-style random effect does not necessarily produce data with relevant features.

To make this concrete, consider penalized splines and the GMST example. In fitting a penalized spline, we assume there is a fixed but unknown $f(x)$, and the object is to estimate it. In the mixed-model representation $f(x) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$, with $u_j \overset{iid}{\sim} N(0, \tau^2)$, the random effect $\mathbf{u}$ is simply a device that penalizes overfitting. The argument from first principles is that in evaluating a penalized spline procedure (e.g., a basis or method of choosing knots), the question is how well the procedure captures certain features of $f(x)$ such as sharp turns or valleys. Therefore, data should be simulated by adding residual error to specific true $f(x)$ having such interesting features. To simulate a dataset by repeatedly drawing $f(x)$ from $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ with $u_j \overset{iid}{\sim} N(0, \tau^2)$ is to leave out precisely the most relevant features.

The pragmatic argument applied to penalized splines is about the effect of increasing the variance τ^2 of the random effect $\mathbf{u}$. In *fitting* a penalized spline, larger τ^2 implies a wigglier, rougher fit: larger τ^2 means changes at the knots are penalized less, so larger changes will be accepted in the fit. But *draws* from $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ do not behave this way, as is easily verified. Consider, for example, the penalized spline model fit to the GMST data used in Sections 3.1 and 4.1.1, with the truncated quadratic basis. Figure 3a shows 10 draws from this spline fit, using the estimated fixed effects and the REML estimates of the smoothing and error variances, $\tau^2 = 947$ and $\sigma^2 = 145$. None of these curves shows anything like the key feature of the spline fit to the GMST data, namely the turn down about 1940 and back up about 1970. In hundreds of draws from this model, we have seen no shapes deviating notably from the ones in Figure 3a. Indeed, we should not expect any: the u_j fitted to the GMST data (Figure 1) do not look like anything like draws from the iid model used for a penalty, with long runs of negative and positive u_j. The chance of drawing such a collection of u_j from an iid normal centered at zero is vanishingly small. (Note also that most of these simulated $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ span a far wider vertical range than the actual GMST data, and Figure 3a does not even include the error term.)

Figure 3b shows 10 draws from a model identical to Figure 3a's model except that τ^2 is 94,700, larger by a factor of 100. Contrary to intuition, Figure 3b's draws are no more wiggly than Figure 3a's; the two plots differ only in the vertical scale. Thus, if a simulation experiment is intended to compare different penalized splines according to how well they detect non-smooth features of curves, such an experiment can only be done by specifying particular non-smooth $f(x)$ and simulating datasets by adding residual error to them, not by drawing curves from $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ with large τ^2 and adding residual error.

The same arguments apply to the new-style random effects in Sections 3.2 and 3.3. In *fitting* models using either of these random effects, increasing the variance of the random effect makes the fitted curve or surface rougher. In *generating data from* these random effects, increasing the random effect's variance does not qualitatively change the shape of the generated curve or surface, it merely changes the scale. Thus, for the same arguments given above, in a simulation experiment intended to compare different spatial models for the mineral-exploration example, the experiment should generate data from a variety of fixed true $W(\mathbf{s})$, with shapes chosen to serve the purposes of the simulation experiment.

In some cases, it may be harmless to draw one true curve from a new-style random effect and then generate simulated datasets by holding fixed that one true curve and repeatedly drawing residual errors. But as we saw with the GMST data, we can only get a true curve with an interesting shape by using u_j that do not look like draws from the model we intend to fit. Generating true surfaces from new-style random effects will produce exclusively bland surfaces lacking features of the sort one would like in a simulation experiment.

4.3 Interpretation of analytical artifacts

We give one example where the interpretation of an analytical artifact depends on whether a random effect is an old- or new-style random effect. With the recent explosion in use of models with random effects, no doubt more will come to light.

Consider the Slovenian stomach-cancer data, Example 3. In analyzing these data, Dr. Zadnik first fit a non-spatial model in which municipality i's count of stomach cancers O_i was Poisson with mean $\log\{E(O_i)\} = \log(E_i) + \alpha + \beta SEc_i$, with flat priors on α and β. In this analysis, β had posterior median -0.14 and 95% posterior interval $(-0.17, -0.10)$, capturing the strong negative association between $SIR_i = O_i/E_i$ and SEc_i. (These data are shown in Reich et al 2006, Figure 1.) Dr. Zadnik then fit the spatial model given in Section 3.2. After adding the random effects for spatial clustering ($\mathbf{S}$) and heterogeneity ($\mathbf{H}$), the coefficient for SEc_i, β, had posterior median -0.02 and 95% posterior interval $(-0.10, 0.06)$. Compared to the non-spatial analysis, the 95% interval was wider and the spatial model fit better, with deviance information criterion (DIC; Spiegelhalter et al 2002) decreasing from 1153 to 1082 even though the effective number of parameters (p_D) increased sharply, from 2.0 to 62.3. These changes were expected. The surprise was that the negative association, which is quite plain in the maps and the non-spatial analysis, had gone away. What happened?

Hodges & Reich (2010) explore this question in detail for normal-errors models with a spatial random effect **S**. For the present purpose, the relevant finding is that such spatial-confounding effects (first reported, as far as we know, in Clayton et al 1993) have a different interpretation depending on whether the random effect **S** is interpreted as new- or old-style. If **S** is a new-style random effect — a mere formal device to capture local lack of fit and not to be taken literally as a draw from a random mechanism — then in general adding it to the model introduces a bias into the estimate of β. If **S** is intended to adjust the estimate of β to account for a missing covariate, the adjustment from adding **S** to the model is biased, perhaps highly. If, on the other hand, **S** is an old-style random effect — as in the example in comment (e) of Section 3.4 — and the fixed effect (SEc in this case) is treated as measured without error and not a draw from a probability distribution, then adding the spatial random effect **S** does not introduce a bias into the estimate of β, but simply inflates its posterior variance (or standard error).

Hodges & Reich (2010) argue that the proper choice of analysis differs depending on which interpretation is given to **S**. The details of their argument are beyond the scope of the present paper but do not affect the present point, which is that this curious analytical artifact cannot be interpreted without taking account of whether the random effect is old- or new-style.

5 Conclusion

We have attempted to define two distinct types of random effects, an older type consistent with the definition given in Scheffé (1959), and a newer type that does not fit Scheffé's definition and can generally be described as a formal device to implement smoothing or shrinkage. While this distinction is not novel, it is not well-developed or consistently observed even among authors who recognize it (including ourselves, we admit), so that sometimes the convenient fiction of a random mechanism is taken literally, as in Section 4.1.1. We have attempted to develop some language and a body of examples (Section 3) to show how the two types of random effects differ in interpretation and how their differences affect inference and prediction, design of simulation experiments, and interpretation of analytical artifacts (Section 4). We do not hope that we have completely cataloged new-style random effects or the implications of their difference from old-style random effects. Rather, this is a first step. In particular, as analyses involving new-style random effects become more common, new analytical artifacts will no doubt turn up and call for interpretations that old-style random effects cannot provide.

Acknowledgements

The authors have benefitted from the comments of Sudipto Banerjee of the University of Minnesota Division of Biostatistics, Melanie Wall of Columbia University's Department of Biostatistics, Yue Cui of Abbott Laboratories, Yi (Philip) He of Eisai Inc., and students in the first author's class on richly-parameterized models. These helpful people cannot be blamed for the results.

References

Adams JL, Hammitt JK, Hodges JS (2000). Periodicity in the global mean temperature series? In Morton SC, Rolph JE, eds., *Public Policy and Statistics: Case Studies from RAND*, New York:Springer-Verlag, 75-93.

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

Clayton DG, Bernardinelli L, Montomoli C (1993). Spatial correlation in ecological analysis. *Int. J. Epi.*, **22**:1193-1201.

Draper DC, Hodges JS, Mallows CL, Pregibon D (1993). Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, Series A*, **156**:9-37.

Freedman D, Lane D (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business and Economic Statistics*, **1**:292-298.

Gelman A (2005a). Analysis of variance — why it is more important than ever (with discussion). *Annals of Statistics*, **33**:1-53.

Gelman A (2005b). Why I don't use the term "fixed and random effects". Blog entry, January 25, 2005, URL
http://www.stat.columbia.edu/∼cook/movabletype/archives/2005/01/why_i_dont_use.html

Hodges JS, Carlin BP, Fan Q (2003). On the precision of the conditionally autoregressive prior in spatial models. *Biometrics*, **59**:317-322.

Hodges JS, Reich BJ (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, **64**:325-334.

NASA (2010). Current version of global-mean surface temperature deviation series.
http://data.giss.nasa.gov/gistemp/tabledata/GLB.Ts.txt

Panoutsopoulou IG, Wendelshafer-Crabb G, Hodges JS, Kennedy WR (2009). Skin blister and skin biopsy to quantify epidermal nerves: a comparative study. *Neurology*, **72**:1205-1210.

Reich BJ, Hodges JS, Zadnik V (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, **62**:1197-1206. Errata available from the corresponding author.

Robinson GK (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science*, **6**:15-51.

Ruppert D, Wand MP, Carroll RJ (2003). *Semiparametric Regression.* New York:Cambridge.

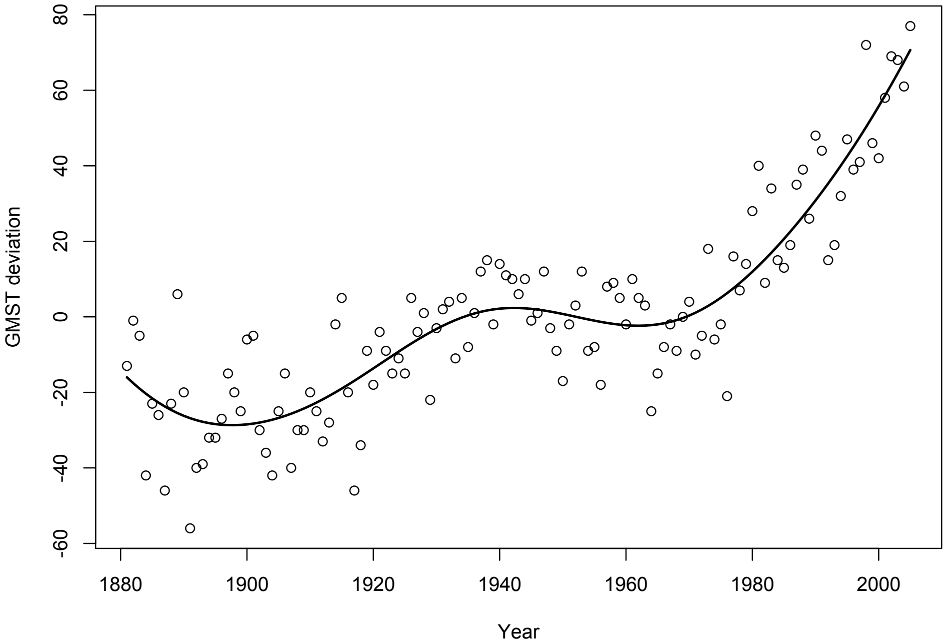
Scheffé H (1959). *The Analysis of Variance.* New York:Wiley.

Spiegelhalter DM, Best DG, Carlin BP, Van der Linde A (2002). Bayesian measures of model complexity and fit (with discussion). *J. Royal Statistical Society, Ser. B*, **64**:583-639.

Sun J, Loader CR (1994). Simultaneous confidence bands for linear regression and smoothing. *Annals of Statistics*, **22**:1328-1345.

Zadnik V, Reich BJ (2006). Analysis of the relationship between socioeconomic factors and stomach cancer incidence in Slovenia. *Neoplasma*, **53**:103-110.

Figure 1: Global mean surface temperature deviation data. Top: The data and spline fit. Bottom: Fitted values for the random effects u_i.



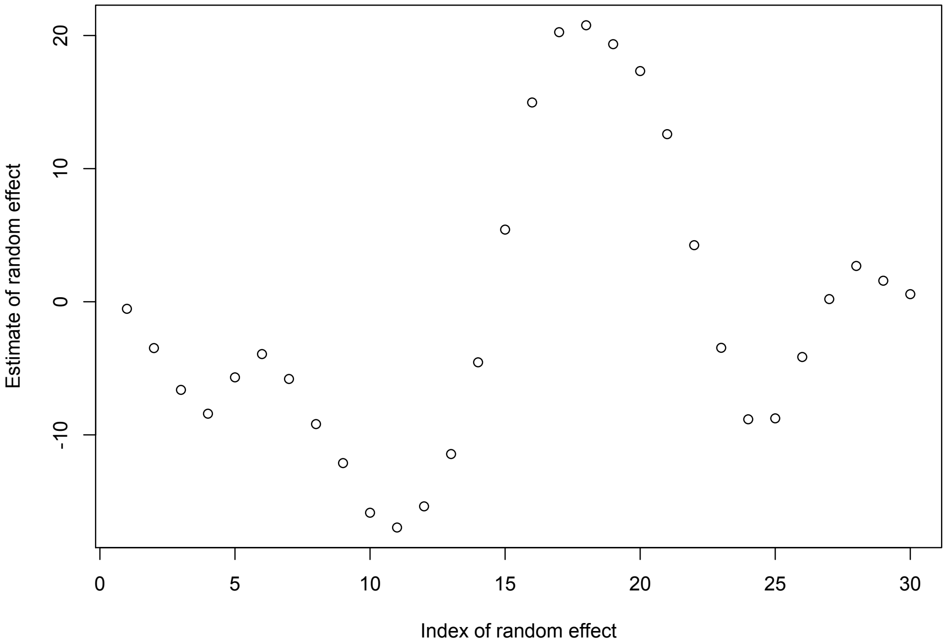


Figure 2: Detail of GMST series and fit for the years 1925-1955, with different confidence intervals. Solid lines: spline fit and simple conditional interval around this (biased) fit $\hat{f}$; dashed lines: unconditional interval centered around $\hat{f}$; dotted lines: bias-corrected fit and conditional interval around it.

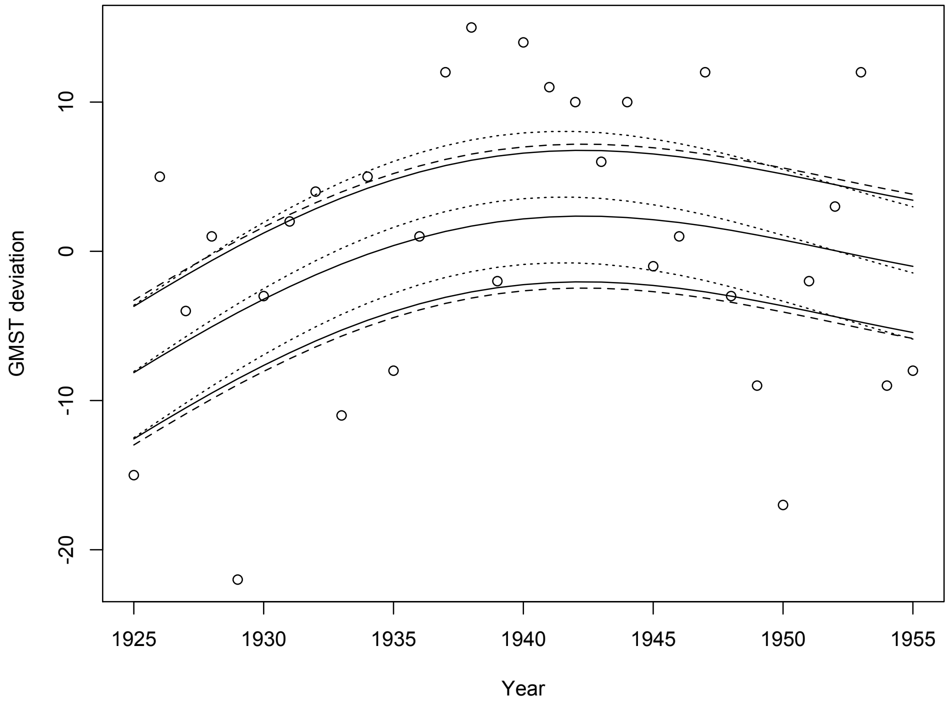
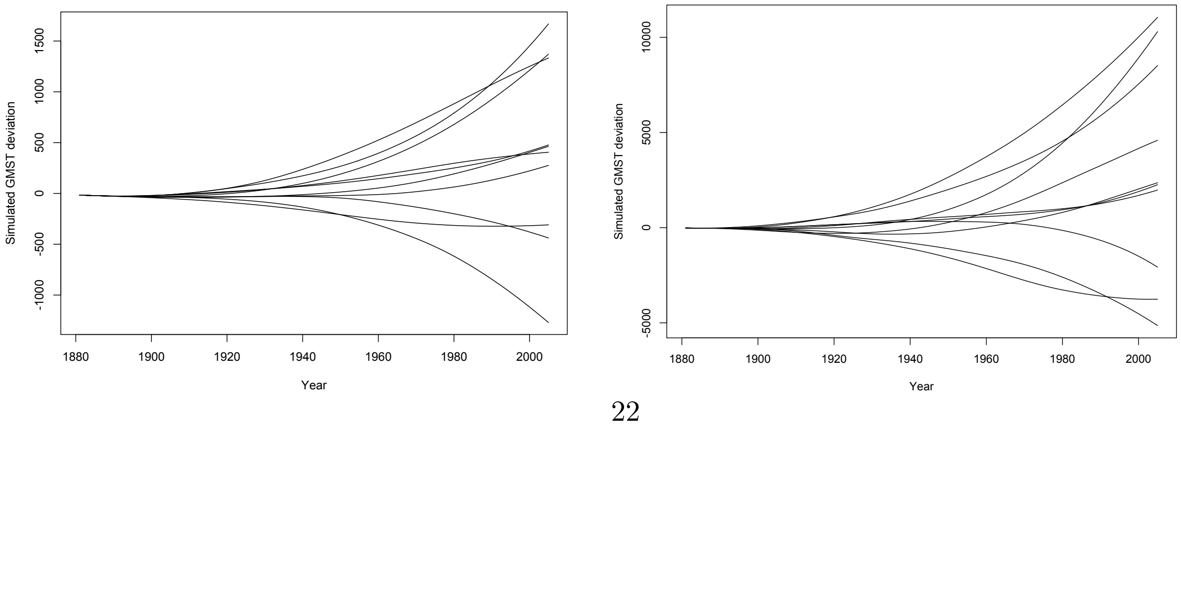


Figure 3: Draws from the convenient fiction used to fit a penalized spline to the GMST data. Panel (a): $\sigma_u^2 = 947$; Panel (b): $\sigma_u^2 = 94700$.


(a) $\sigma_u^2 = 947$

(b) $\sigma_u^2 = 94700$