Random Effects Old and New

Jim Hodges
Division of Biostatistics
University of Minnesota

Updated version of

Hodges (2013), *Richly Parameterized Linear Models*, Ch 13

Hodges & Clayton (2011) paper:
http://www.biostat.umn.edu/~hodges/RPLMBook/Manuscripts/Hodges-ClaytonRandomEffectsOldAndNew.pdf

Many things that are now called "random effects" would not have been recognized as REs 50 years ago.

This distinction, old- vs. new-style random effects, has important consequences, conceptual and practical.

Outline

Old style: Definition, example.

Some background: Three senses of "probability"

New style: They implement smoothing/shrinkage, and they're part of the model's mean, not its error variance

Consequences: Old & new REs require distinct ways to
- do inference and prediction
- interpret analytical artifacts
- do simulations for evaluating statistical methods

Old-style random effects

Scheffé (1959, p. 238):

-- the *levels* are *draws* from a *population*,

-- the draws are not of interest in themselves but only as samples from the larger population.

Example of old-style random effects

New objective methods to count and measure nerve fibers in skin and mucosa (Kennedy Lab).

Recent study (Panoutsopoulou et al 2013):

- -- 25 "normal" (non-diabetic) subjects
- -- Skin sampled by biopsy and blister (method)
- -- Sampled from the calf and on the foot (locations).

Three old-style random effects:

- -- subject main effect
- -- method-by-subject interaction
- -- location-by-subject interaction

The analysis also has a residual (error term) = method-by-location-by-subject interaction

This design has three old-style random effects:

-- Subject main effect
-- Method-by-subject interaction
-- Location-by-subject interaction

These random effects describe how:

-- Average nerve density varies between subjects
-- Blister minus biopsy varies between subjects
-- Foot minus calf varies between subjects

These *are* old-style random effects

- The levels of the RE (subjects) are a sample (though not a formal sample).
- The levels are not interesting in themselves but only as representatives of non-diabetic adults
- The object was to measure differences, in that population, between methods and locations.
- <u>These random effects are part of error variation</u>: They capture nuisance correlation within subject

A new measurement on a new person would involve a new draw of all the variance components.

A new measurement on one of these 25 people would involve a new draw on only the error term.

Some background: Three senses of "probability"

(1) Draws from a random mechanism, either one we create and control, or one we imagine is out in the world

$\approx$ frequentist notion of probability

(2) A person's uncertainty about an unknown quantity

= the subjective Bayesian notion of probability

(3) <u>A descriptive device</u>.

Example of #3:

The heights of US-born 52-year-old males employed by U of Minnesota

can be described as looking like

n iid draws from $N(\mu,\sigma^2)$

This doesn't imply anyone's height is a draw from a random mechanism or is even uncertain.

Definition (sort of): New-style random effects

A new-style random effect differs from old-style REs in at least one of these ways:

(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.

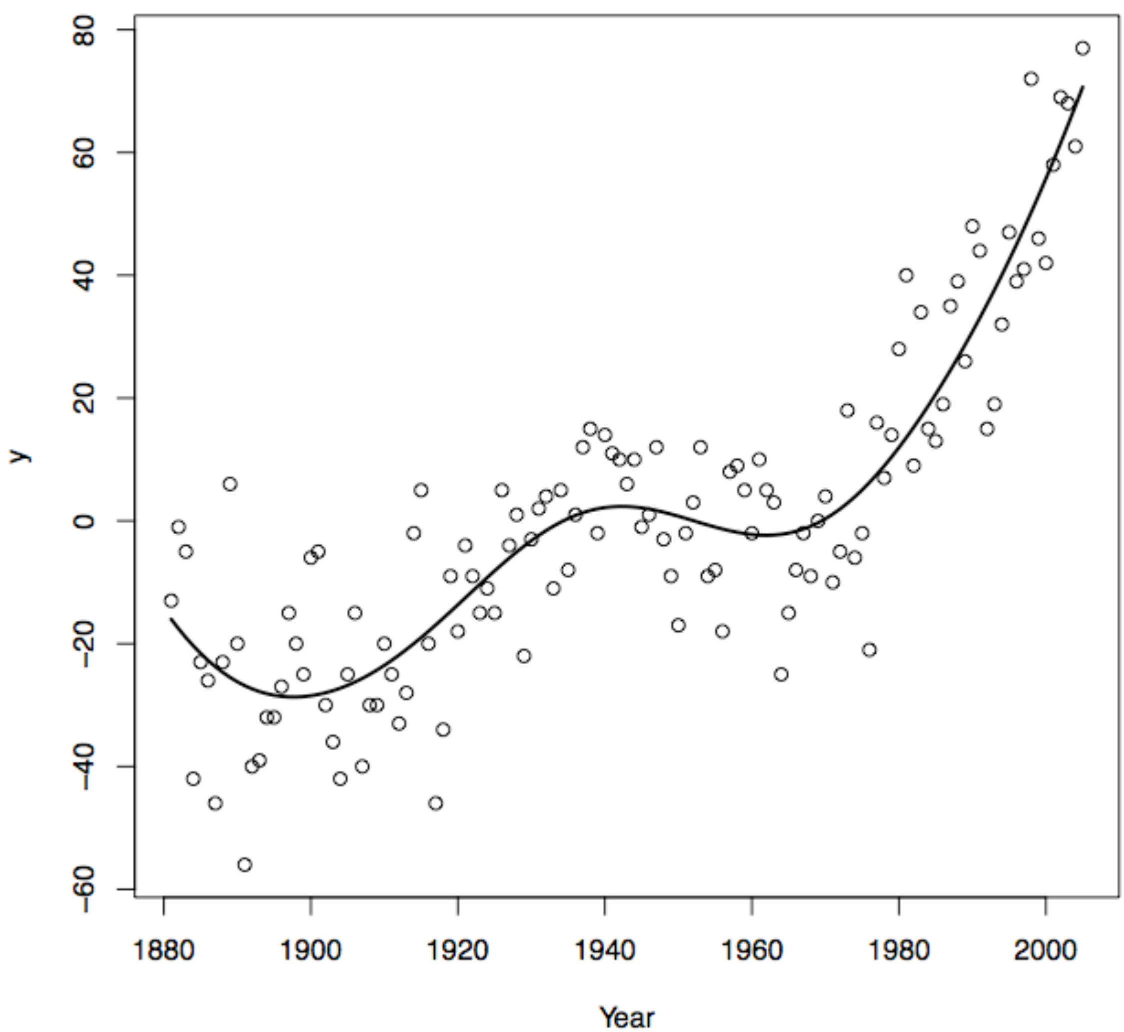
(2) The levels of the effect come from a meaningful population but they are the whole population and these particular levels are of interest.

(3) A sample has been drawn, but the samples are all from a single level of (draw from) the model's random effect, and that level is of interest.

(1) The levels of the effect are not draws from a pop'n; there is no pop'n. The mathematical form of a random effect is used for convenience only.

Example: Mixed linear model representation of penalized splines

Object: Draw a smooth curve through the data.



A penalized spline is just a linear model with constrained coefficients.

The GMST data was fit using

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \Sigma_{j=1,30}\, u_j([x_i - \kappa_j]_+)^2 + \text{error}$$

Select $(\boldsymbol{\beta},\mathbf{u})$ to minimize

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \ni \mathbf{u}'\mathrm{D}\mathbf{u} \leq \mathrm{K}$$

This is equivalent to minimizing

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \alpha\, \mathbf{u}'\mathrm{D}\mathbf{u}$$

where α is a function of K

Minimizing

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \alpha\, \mathbf{u}'\mathrm{D}\mathbf{u}$$

is formally identical to estimating $(\boldsymbol{\beta},\mathbf{u})$ in the mixed linear model

$$\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \mathrm{Z}\mathbf{u} + \boldsymbol{\varepsilon},\ \varepsilon_i \sim N(0,\sigma^2),\ u_j \sim N(0, \boldsymbol{\tau}^2)$$

when σ^2 and $\boldsymbol{\tau}^2$ are taken as given; $\alpha = \sigma^2 / \boldsymbol{\tau}^2$

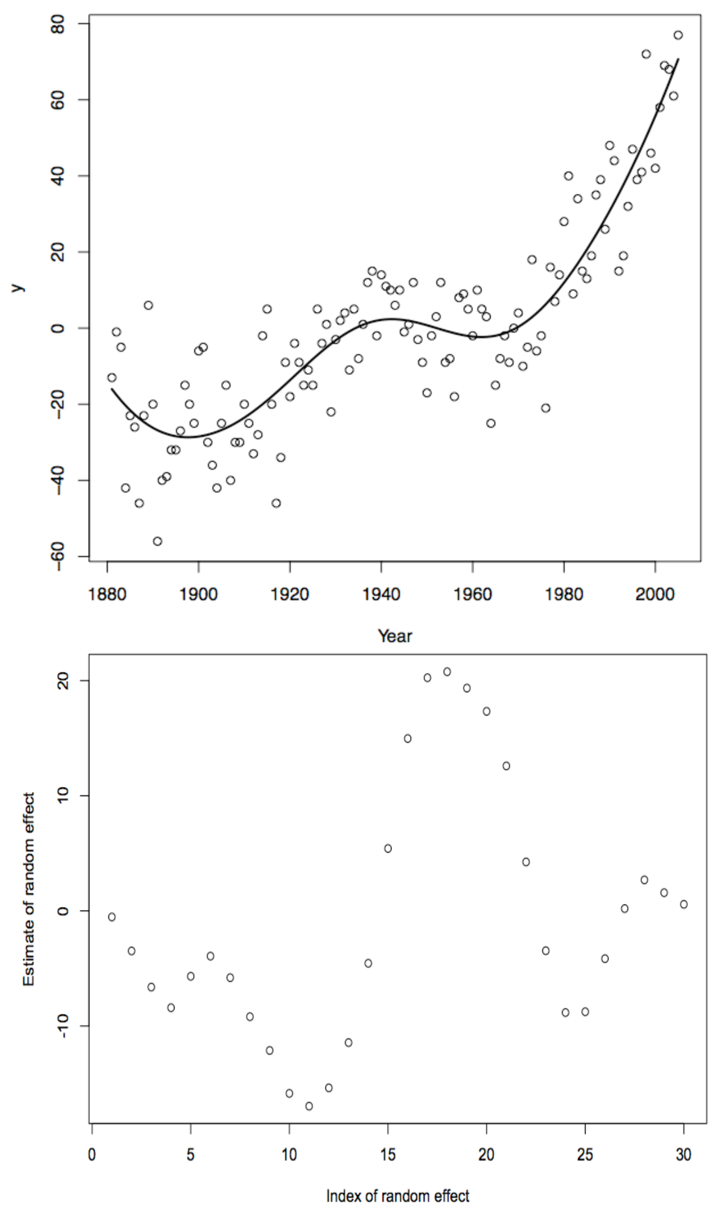
The penalized spline now has the mathematical *form* of a model with a random effect.

Ruppert, Wand, and Carroll (p. 138):

"[T]he mixed model formulation of penalized splines [is] a *convenient fiction* to estimate smoothing parameters. [It] is a reasonable (though not compelling) Bayesian prior for a smooth curve, and [maximizing the restricted likelihood] give[s] estimates of the smoothing parameter that generally behave well".

The analysis has the *form* of a random effects analysis, but $X\boldsymbol{\beta} + Z\mathbf{u}$ is not a draw from a random mechanism.

The fitted u_j don't look like iid $N(0, \tau^2)$:



Like $\boldsymbol{\beta}$, **u** is just part of the model's mean structure; **u**'s distribution just constrains the estimates of the u_j.

We <u>*choose*</u> this model to serve a particular purpose.

It is senseless to imagine that this model generated the data, or that more draws could be made from it.

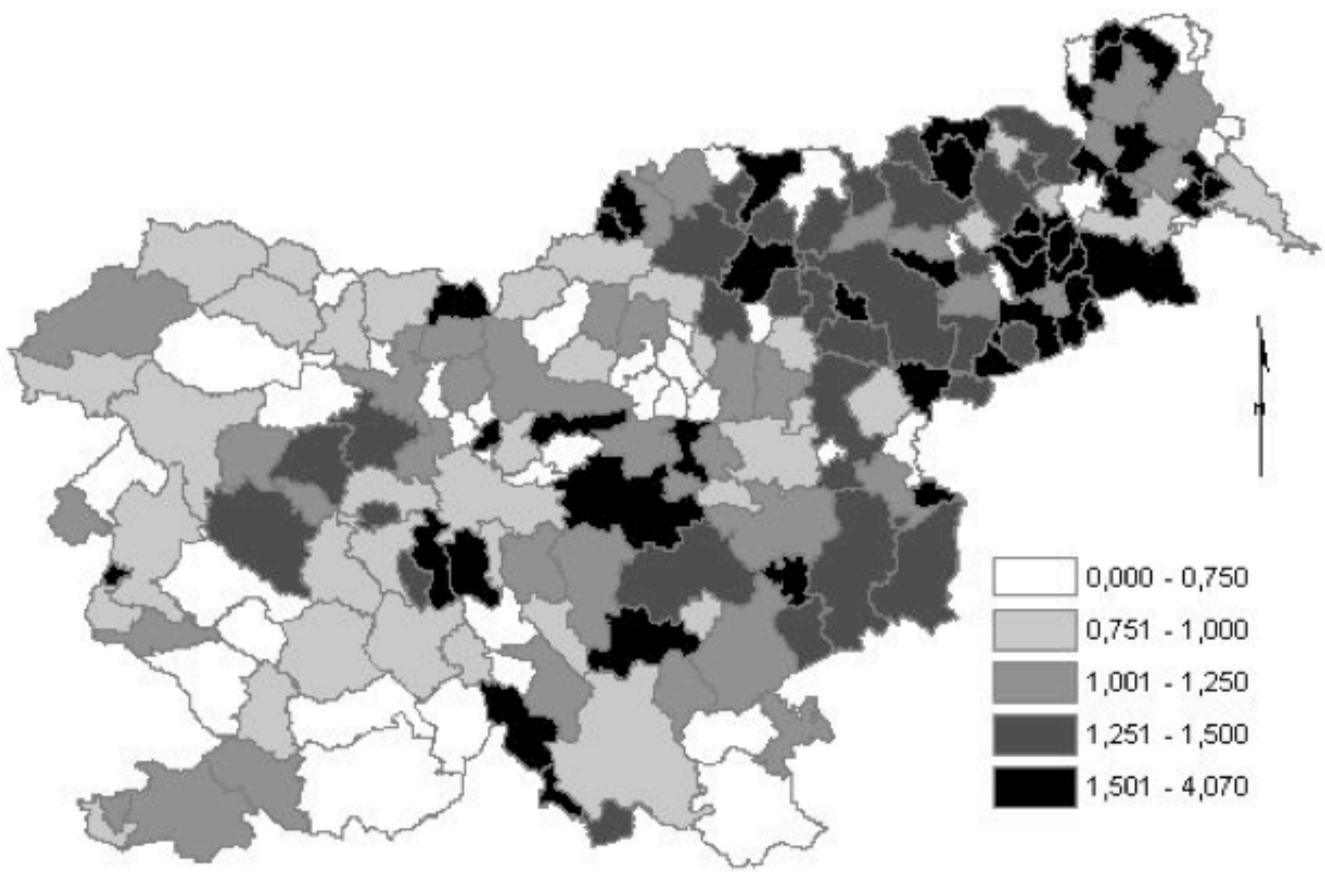
The RE form merely provides a flexible family of smooth mean functions and some discipline on the fit.

... After adopting this convenient fiction, Ruppert et al behave like conscientious statisticians, checking for heteroskedastic errors, non-linearity, etc.

(2) The levels of the effect come from a meaningful pop'n but they're the whole pop'n and are of interest.

Example: Stomach cancer in Slovenia, 1995-2001

Standardized incidence ratio, by municipality



Disease maps are commonly smoothed using models with random effects like this one (Besag et al 1991)

O_i = stomach cancers in municipality i

$O_i \sim$ Poisson with

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

E_i = expected # of cancers.

SEc_i = SES, centered

Heterogeneity: $\mathbf{H} = (H_1, ..., H_{194})' \sim \text{iid } N(0,\tau^2)$.

Spatial clustering: $\mathbf{S} = (S_1, ..., S_{194})' \sim$ ICAR

$$p(\mathbf{S} \mid \sigma^2) \propto \exp(- \mathbf{S}'\mathbf{Q}\mathbf{S} / 2\sigma^2)$$

Q describes spatial neighbor pairs

Unlike the spline:

- There's a meaningful population BUT
- The municipalities are the *whole* population and they're of interest.

Like the spline:

Perhaps some random process produced the O_i, but

S + **H** was not produced by a draw from ICAR + iid, and

Even though new counts could be made for 2002-2008, they wouldn't be an iid draw from the same mechanism.

It is hard to see how **S + H** can usefully be described as a draw from a random mechanism.

The intuition motivating a spatial model --
near municipalities are more similar than far --
is *descriptive*, not mechanical.

Less problematic: **S**'s ICAR model is a descriptive device (3rd sense of probability).

We could say the 192 Slovenian S_i, if observed, would *look like* a draw from a particular ICAR model.

Like the heights of 52-year old men, this doesn't mean **S** *was* drawn from a random mechanism; it's a convenient way to describe a group of constants.

Some would say it's natural to think of **S**'s ICAR distribution as a *statement of subjective probability* (2nd sense of probability).

If we view **S**'s distribution as

• <u>description</u>, we can use that description in a statistical method.

• <u>subjective probability</u>, we can use it in a Bayesian computation.

Either way:

-- The random effect is a device *we choose* for a particular purpose.
-- It is senseless to imagine that this model *generated* the data.

(3) A sample has been drawn, but the samples are all from <u>one</u> level of (draw from) the model's random effect, and <u>that level is of interest</u>.

In a specific region, we're interested in the fraction of iron at a certain depth, $\mu + \mathbf{W}(s)$ at location s

μ and $\mathbf{W}(s)$ are fixed but unknown

We observe $y(s_i)$ and model it as

$$y(s_i) = \mu + W(s_i) + \text{error}(s_i)$$

with iid $\text{error}(s_i)$

$\mathbf{W}(s)$ is commonly modeled as a Gaussian process:

$$(W(s_1), ..., W(s_n)) \sim \text{multivariate normal}$$

This seems to be <u>unlike</u> the previous examples:

-- The s_i are a sample of possible locations.
-- These s_i are not so interesting; we want to know about the region, which we might call a population.
-- We could draw new s_i or make new measurements at the original s_i.

This sounds like an old-style random effect ...

But it's not.

A draw from this Gaussian process model is a *function* on the whole 2-D region of interest.

As with the spline, there is no population: The random effect's "population" is the hypothetical infinite population of possible draws from the GP.

As in the stomach-cancer example:

• Exactly one draw has been made from this random effect and no more can or will be made.

• The whole point of gathering data is to learn about this one draw. If we measure y_i at new s_i, we just learn more about this one draw.

• We may *describe* W(s) by saying $\{W(s_i)\}$ *looks like* a draw from Normal with covariance Σ; or that this distribution represents *our uncertainty* about W(s), and we can deploy either of these in a statistical method.

Comments on new-style random effects, 1

New-style random effects can all be understood as formal devices that implement smoothing or shrinkage.

This is obvious for penalized splines.

This is less clear for the other examples, perhaps because we habitually think of spatial models in terms of covariance matrices.

Comments on new-style random effects, 2

New-style random effects are part of the model's mean, not part of its variance structure.

The distribution of a new-style random effect does not embody or represent the mechanism that produced the data.

> $X\boldsymbol{\beta} + Z\mathbf{u}$, $\mathbf{S}$, or W(s) is just a group of fixed, unknown constants.

The random effect's distribution is something *we choose*.

In a given situation, some choices are better than others.

But "better" means they give better estimates of the unknowns, not that they better represent the mechanism that produced the data.

Comments on new-style random effects, 3

Maybe the true $X\boldsymbol{\beta} + Z\mathbf{u}$, $\mathbf{S}$, or W(s) arose as a draw from some random mechanism.

But it makes no sense to imagine further draws, and the draw(s) we have is (are) of interest.

As hypothesized *mechanisms* for producing the data, these models are silly.

- If we make this mistake, we are mistaking the shovel for the process that produced the soil.

It is more accurate to think of these models as descriptive – superficially – but useful.

Comments on new-style random effects, 4

Re the Slovenian example, I've heard

- "$S_i + H_i$ is only in the model as an error term, like the error term in a linear regression."

What "error" does $S_i + H_i$ capture?

- Local variation in the mean of the data-generating process that's not captured by the predictor SEc.
- That is, local bias or lack of fit in the fixed effects.
- Berkson distinguished this kind of error from "classical" error, e.g., a linear regression's error term.

These new-style REs are part of the model's mean.

Comments on new-style random effects, 5

Some spatial analyses do involve old-style random effects.

Example:

- Ozone in the Boston area, daily data for m years.
- Days may be an old-style random effect.

But not necessarily ...

- Suppose we're interested in a specific week.

- We may describe that week's spatial ozone gradient using an RE, but it's part of the model's mean.

- There's a meaningful sense in which this fixed, unknown feature of Boston was drawn from a probability distribution, but it's not relevant to our question.

Practical Consequences

(1) In simulation experiments for evaluating statistical methods, data should be simulated differently for old-style and new-style random effects.

(2) Some analytic artifacts have different interpretations for old-style and new-style random effects, and different remedies. (Example: Spatial confounding.)

(3) Appropriate inference and prediction may depend on whether a random effect is old-style or new-style. (Example: Confidence intervals for penalized splines)

Practical Consequences: Simulation experiments to evaluate statistical methods

Principle:

- A simulation experiment is intended to answer specific questions.
- The experiment's design must enable it to answer those questions.

For old-style REs, the question is: Measure behavior of estimates & intervals for FEs and variance components.

We'd answer this (in the nerve-density example) by simulating a subject's 4 measurements this way

- Make a draw of each of the three REs
- Make four draws from residual error
- Each fake observation is a sum of true FEs, drawn REs, and error.

This mimics the way the real data were generated: by sampling subjects.

Simulation experiments for new-style random effects

Draws never <u>need</u> to be made from a new-style RE.
It's usually <u>incorrect and self-defeating</u> to do so.

Common questions asked about methods using new-style REs:

- What is the fit's bias or MSE at specific predictor values?
- What is the coverage of a particular type of interval?

For such questions, it's <u>*wrong*</u> to draw from a new-style RE

• <u>First principles</u>: A new-style REs is just a convenient fiction; taking it literally is a conceptual error.

• <u>Pragmatic</u>: Simulating from a new-style RE doesn't give data with relevant features and is thus self-defeating.

Argument from first principles

A new-style RE is just a convenient fiction; taking it literally is a conceptual error

In evaluating a penalized spline procedure (e.g., a basis), the question is how well it captures turns, peaks, valleys.

Therefore, data should be simulated by adding residual error to specific true f(x) having turns, peaks, valleys.

Simulating data by drawing

$$f(x) \sim X\boldsymbol{\beta} + Z\mathbf{u}, \qquad u_j \sim \text{iid } N(0,\tau^2)$$

- leaves out precisely the most relevant features.
- has a different true f(x) for each draw.

Pragmatic argument

<u>*Draws*</u> from models with new-style REs are inconsistent with our intuition, which arises from <u>*fitting*</u> such models.

Applied to a spline with a truncated quadratic basis:

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{u} + \boldsymbol{\varepsilon}, \quad \varepsilon_i \sim N(0,\sigma^2), \quad u_j \sim N(0, \boldsymbol{\tau}^2)$$

u is the changes in the quadratic coefficient at the knots

<u>*Fitting*</u> this spline, larger $\boldsymbol{\tau}^2$ => wigglier, rougher fit.

<u>*Draws*</u> from $X\boldsymbol{\beta} + Z\mathbf{u}$, $u_j \sim N(0,\boldsymbol{\tau}^2)$ don't do this.

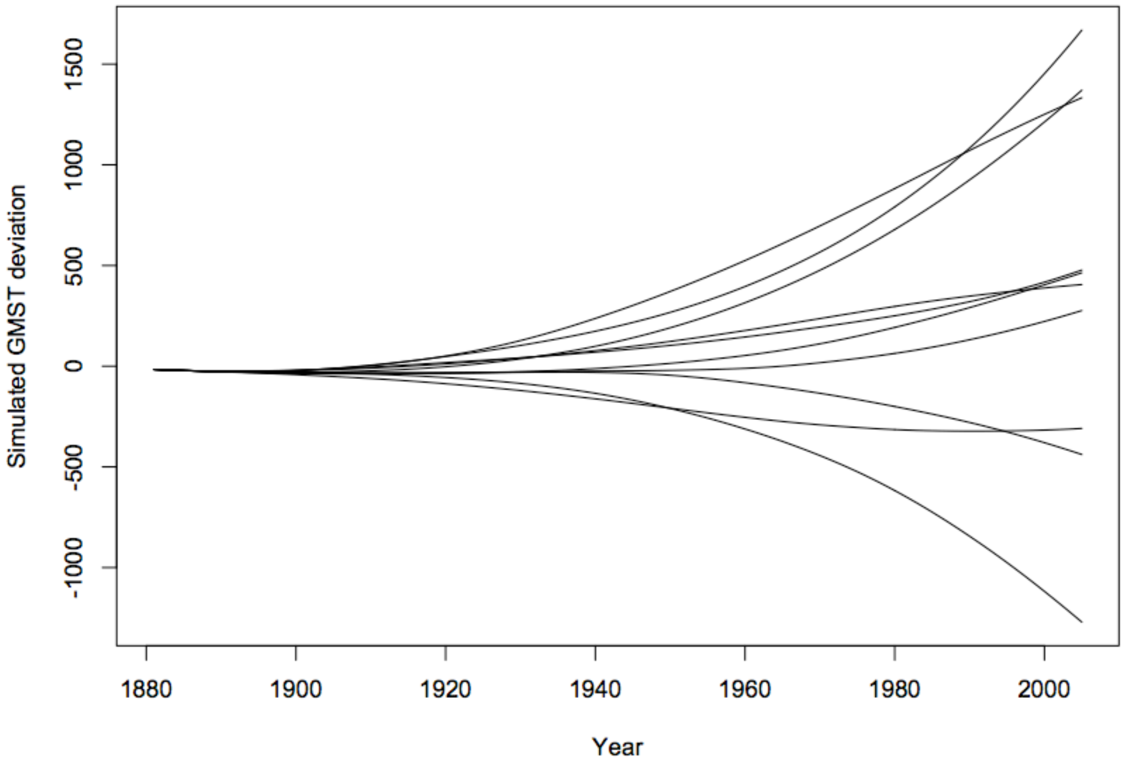
Proof of pragmatic argument:

Fit a spline (truncated-quadratic basis, 30 knots) to the GMST data.

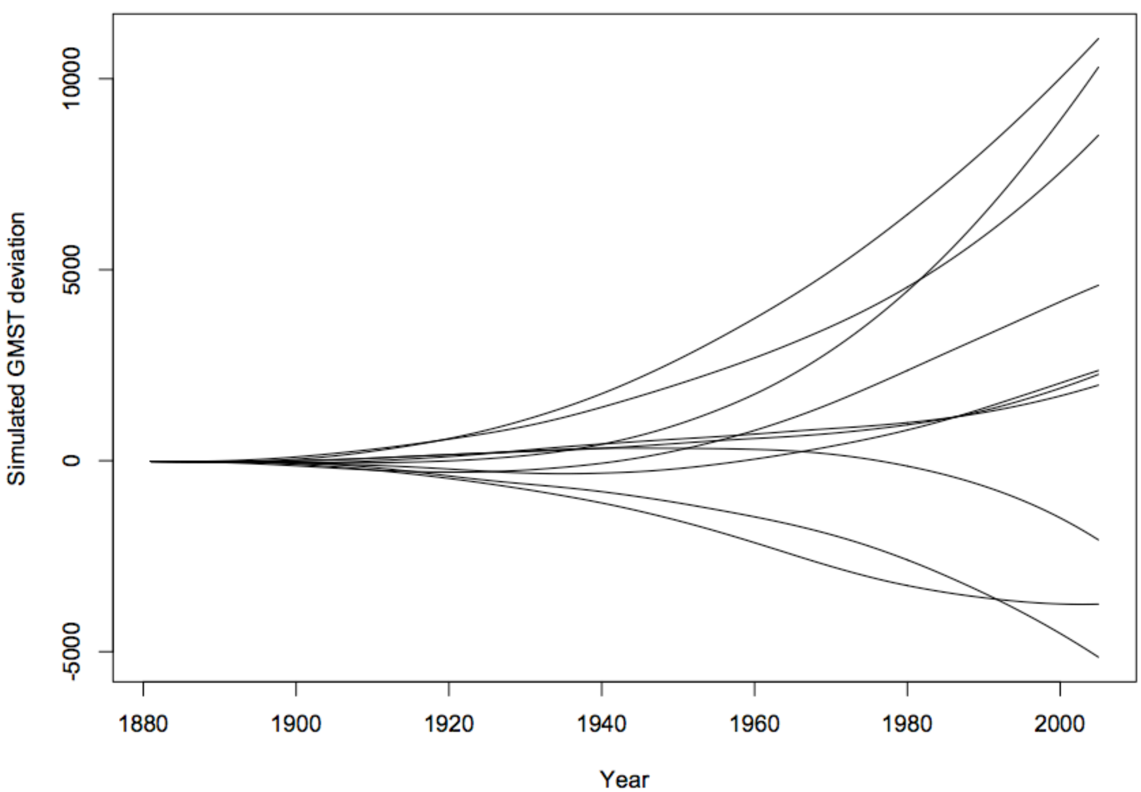
Draw 10 curves from $X\boldsymbol{\beta} + Z\mathbf{u}$ using the estimated $\boldsymbol{\beta}$ and u_i's estimated variance $\boldsymbol{\tau}^2 = 947$.

Draw 10 more curves, with only one change: $\boldsymbol{\tau}^2 = 94{,}700$

Results are on the next page. Can you tell which is which?



Simulated GMST deviation
1500
1000
500
0
-500
-1000
1880
1900
1920
1940
1960
1980
2000
Year



Simulated GMST deviation
10000
5000
0
-5000
1880
1900
1920
1940
1960
1980
2000
Year

The draws with bigger variance only have a larger vertical scale.

These curves lack interesting features: to produce that feature, you need a spectacularly improbable **u**.

Instead, to answer any kind of real question, you must specify interesting true f(x) and simulate datasets *by adding residual error only*.

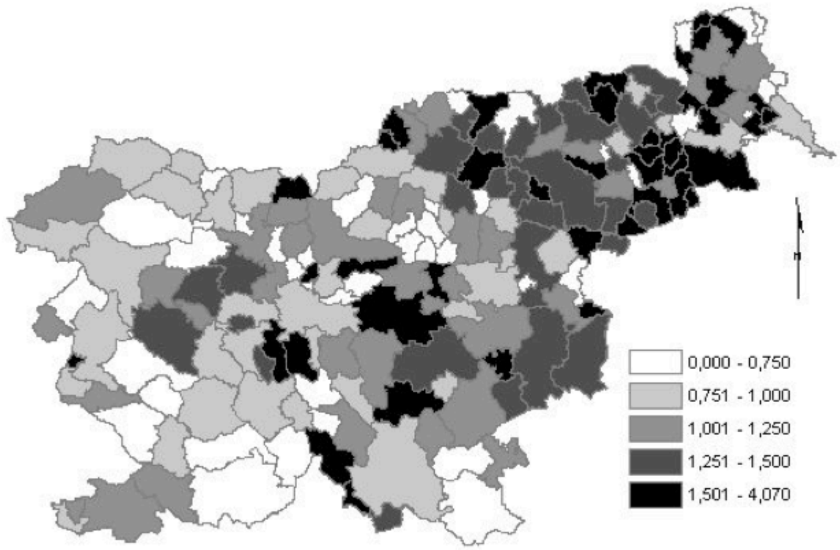
Sometimes it seems harmless to draw a true curve from a new-style random effect and then repeatedly draw residual errors.

But this gives only bland curves lacking the interesting features you'd want in a simulation experiment.

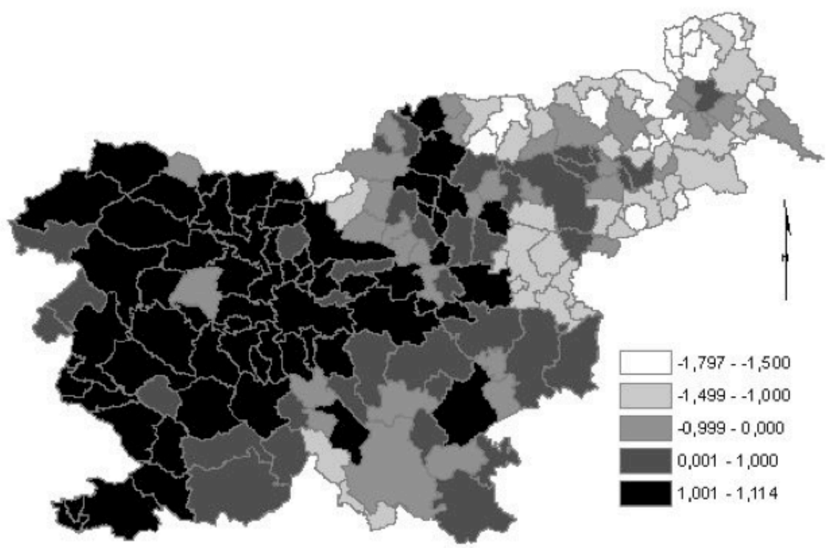
Consequences: Interpreting analytical artifacts

Dr. Vesna Zadnik collected the Slovenian stomach-cancer data to test whether it was associated with SES

SIR of stomach cancer, 1995-2001



Socioeconomic status



<u>First, a non-spatial analysis</u>: $O_i \sim \text{Poisson}(\mu_i)$, where

$$\log\{\mu_i\} = \log\{E_i\} + \alpha + \beta\, SEc_i$$

$\beta \mid O_i \sim$ median -0.14, 95% interval -0.17 to -0.10.

<u>Now do a spatial analysis</u>:

$$\log\{\mu_i\} = \log\{E_i\} + \alpha + \beta\, SEc_i + S_i + H_i$$

where **S** is an L_2-norm improper CAR

H is iid Normal mean 0, precision τ_h.

$\beta \mid O_i \sim$ median -0.02, 95% interval -0.10 to +0.06.

<u>What happened? And what should you do?</u>

The interpretation depends on whether S is an old-style or new-style random effect

(Hodges & Reich, *Amer. Stat'n* 2010)

<u>(A) S is a new-style random effect</u>

(i) Spatially-correlated errors can introduce or remove bias and are not necessarily conservative.

<u>(B) S is an old-style random effect</u>

(ii) The spatial effect is collinear with the fixed effect; there's no bias (Dave Nelson).

(iii) Adding the spatial effect creates "information loss"; there's no bias (Dave Nelson).

(iv) *Both* estimates of β are biased because error is correlated with SEc (Paciorek 2010).

Except for (iv), these take SEc as fixed & known: observed without error, not drawn from a distribution.

Practical consequences: Inference & Prediction

"Inference" = analyses focused on the present dataset and on models that supposedly generated it.

"Prediction" = analyses focused on data related to the present set but as yet unobserved or unknown.

Inference

Bayesian:

Old- or new-style, posteriors are computed the same way.

But priors aren't. Priors for old-style REs can draw on intuition and previous data; new-style REs can't.

Non-Bayesian:

Old-style REs are deeply embedded in the terminology.

"BLUP": "Unbiased" refers to the expectation over random effects as well as error terms.

A penalized spline fit is a BLUP; it flattens peaks and fills valleys. These are biases.

So the term "BLUP" shouldn't be used for new-style REs.

But the confusion has more serious consequences ...

Confidence intervals for fitted values in penalized splines.

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{u} + \boldsymbol{\varepsilon},\ \varepsilon_i \sim N(0,\sigma^2),\ u_j \sim N(0, \boldsymbol{\tau}^2)$$

Taking **u** as a new-style RE, a CI can use

var(f-hat(x) | **u**), a function of σ^2 and $\boldsymbol{\tau}^2$ but not **u**

But this CI's coverage is too low where f-hat(x) is biased, because it is centered at E[f-hat(x) | **u**], not f(x).

The obvious fix is to subtract the bias from the fitted spline and center the CI on that.

But some treat **u** as an old-style RE:

Ruppert, Wand, & Carroll (2003, p. 139):

"But, since E(**u**) = 0, the unconditional bias is E[f-hat(x) - f(x)] = 0. Thus, on average over the distribution of **u**, f-hat(x) is unbiased for f(x). To account for bias in the confidence intervals, the [conditional] variance var(f-hat(x) | **u**) should be replaced by the conditional mean-squared error E[{f-hat(x) - f(x)}^{2} | **u**] ... then averaged over the **u** distribution."

This gives a wider interval "because [it] accounts for both ... variance and squared bias".

But it's still centered in the wrong place:

Coverage is still too low in areas of high curvature and too high in areas of low curvature.

Prediction

For old-style REs (using the nerve-density example), there are two cases:

(1) Predicting biopsy and blister nerve density measurements from a new subject's calf and foot.

Each new measurement's variance is the sum of all three REs (subjects, method-by-subject, location-by-subject), and residual error.

(2) Prediction of (say) a new biopsy from the foot of an already-sampled subject.

Each new measurement's variance is simply the residual error variance.

(Prediction uncertainty also accounts for "prediction" of the three REs for that subject.)

Prediction for New-style REs

<u>All</u> predictions are like the old-style RE, predicting a new measurement on an already-sampled subject: the new measurement's variance is error variance only.

For some problems (GMST or Slovenia) predictions of new measurements may be impossible or ill-advised.

Apart from these cases, no new sampling of the RE is possible, so you must condition on the existing "draw" of the RE.

In the mineral-exploration example:

It is possible to measure again at an observed s_i or
to measure at a new s_0

In either case, the only interpretation that does not do violence to the subject matter is that the random effect W(s) has already been drawn and is simply unknown.

Otherwise, making a prediction would involve re-drawing the process that produced the ore seam.