A Random Effect in the Analysis Need Not Imply A Random Draw in Data Generation

The term "random effect" has come to be used more broadly than it was, say, 50 years ago.

Analyses that are now described as including random effects apply to situations qualitatively different from those originally analyzed using random effects.

As usage of “random effect” has broadened, fewer people seem to recognize a related distinction that has important consequences, both conceptual and practical.

This lecture is about that distinction.

Briefly, I'm going to make this argument:

As analysis of richly-parameterized models become unified in the MLM framework (as in RWC), more analyses take the form of a random-effects analysis.

An analysis including an effect with the form of a random effect does not necessarily imply that, out in the world, any random mechanism produced that effect.

The form of the analysis should not be confused with the process in the world that produced the data.

I make this argument using examples, most of which you’ve seen before.

Older, narrower notion of a random effect:

- -- A random effect represents draws from a population, either a real finite population or a hypothetical infinite one, and

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

Example 1: Draws from a real finite population.

William Kennedy's neurology lab uses new methods to count nerve fibers in skin and GI mucosa, an objective way to measure e.g., diabetic neuropathies.

Recent dataset:

25 "normal" subjects; from each they took

-- two types of skin samples: biopsies and blisters;
-- each taken at two locations: calf and foot.

The analysis involves three random effects:

-- subjects,
-- method-by-subject int (method = blister/biopsy);
-- location-by-subject int (location = foot/calf).

These random effects arise from sampling subjects.

The analysis also includes a residual error = method-by-location-by-subject interaction.

Example 2: Draws from a hypothetical infinite population.

Dwight Anderson's lab in the dental school has done path-breaking studies working out, in molecule-by-molecule detail, the structure of the phi-29 virus.

One of their measurement processes has three steps:

-- producing, decomposing a batch of viral shells,
-- separating the molecules by weight on a gel,
-- burning blobs from each gel to give a measured weight for each blob.

To get 98 measurements for one molecule, they used

-- 9 batches, 11 gels, 7 oxidizer runs,
-- 98 measurements grouped irregularly into batches, gels, and oxidizer runs.

I treated batches, gels, and oxidizer runs as three random effects, with a residual error term for each of the 98 measurements.

In these examples, it makes sense to talk about these persons or measurements as at least plausibly random draws from a population, from which more draws could be made.

In either case, making new measurements would require drawing new samples from the random effects.

Also, these specific subjects or measurements clearly have no intrinsic interest, but are only interesting as representatives of a larger population of humans or viral measurements.

To do a simulation study relevant to methods used for either problem, it would be necessary to generate a simulated dataset by

-- first making a draw of each random effect
-- then drawing a residual error.

Now for examples not included in the old usage

Example 3: Global mean surface temperature, smoothed with a penalized spline.

Historically, penalized splines arose from the following kind of reasoning (as in RWC).

Each year t had a true global mean surface temperature (GMST). We represent these true GMSTs as a function f(t), which is fixed, but which could not be observed directly.

Instead, we have measurements y_t with error:

y_t = f(t) + error_t,   t = 1, . . . , T.

The fixed but unknown function f is presumed to be smooth.

The error is treated as arising by a random draw; assume error_t are independent of each other.

To estimate f at the observed t

- -- put a spline basis in the columns of matrix H,
- -- specify vector f = Hc for some coefficients c.
- -- in choosing c, penalize roughness in c;

For a particular penalty this implies the following optimization problem :

min{in c}((y - Hc)'(y - Hc))   where c'Dc <= K.

which turns out to be equivalent to:

min{in c}((y - Hc)'(y - Hc) + alpha* c'Dc)

That optimization problem can in turn be re-cast in the form of a random effect analysis:

$$y = Xb + Zu + e \quad e \sim N(0, v\ I\_T),\ u \sim N(0,S),$$

with v a variance and S a covariance matrix.

-- the design matrix X and its coefficients b capture the non-penalized columns of H,

-- the design matrix Z and its coefficients u capture the penalized columns of H,

-- S is the inverse of the non-zero part of D.

The penalized spline method of analysis now has the form of a random effect analysis,

We can implement this method using tools developed for random effect analyses.

This analysis has the form of a random effects analysis, BUT

f(t) is NOT a draw from a random mechanism. Rather, f(t) is fixed but unknown.

Bayesian terminology is delicate here,

-- f is fixed and unknown;

-- Your uncertainty about f is described using a probability distribution; it describes Your uncertainty about f, not something inherent in f.

Although we may call f a random variable in doing Bayesian theory and computations,

This does not imply that f was, in any sense, drawn from some random process.

This is the key conceptual distinction:

The analysis method formally treats f as a random effect,

BUT

The process that generated the data does not involve drawing f from a distribution.

This differs from the older notion of "random effect":

Here, it makes no sense to draw f or u again.

We are interested in this specific u and f, not in other conceivable realizations.

Possible counterargument:

- When f was generated out there in the world, it did involve draws from a random mechanism

(Note: This argument would work better if the analysis had been a dynamic linear model.)

In some cases, it may make sense to think that in producing the now fixed but unknown f, a random draw was made on one occasion.

BUT having once made that draw,

- the specific value of that single draw is of intrinsic interest, and

- it makes no sense to think of further draws.

Also, penalized splines are used in situations where it does not make sense to think of f as having been drawn even once:

RWC examples:

- -- LIDAR: f describes log ratio of light received from two laser sources, as a function of distance traveled before reflection back to source.

- -- Janka hardness: a structural property of wood which is a function of density.

- --NOx in engine exhaust, as a function of compression ratio.

These functions are implied by physical/chemical laws; the spline merely approximates them.

Also, I have used S spline functions to approximate complicated deterministic mathematical functions.

Immediate practical consequence of distinguishing between a random effect in the analysis method and a random draw in the data-generation process

In generating data for a simulation experiment involving p-splines

-- It makes no sense to simulate data by drawing f from Xb + Zu, u ~ N(0,S).

-- Data should be generated by repeatedly drawing vectors of errors e and adding them to specific fixed true fs,

-- The true fs should be chosen depending on what we want to learn from the experiment.

Simulating a dataset by drawing f ~ Xb + Zu, u ~ N(0,S) and adding an error e . . .

obliterates precisely the relevant features of the data being analyzed.

Example 4: Generic geostatistical data.

Classic spatial problem: in a given region, estimate X, say, fraction of iron in rock at 1000 m depth.

Each location s has a true value of X, X(s) which is fixed but unknown.

We measure X(s) with error at specific locations s_i, y(s_i) = X(s_i) + error(s_i), error(s_i) independent

We could estimate X(s) using a 2-D spline, in which (per RWC) X(s) has the form of a random-effect.

The previous argument applies: giving X(s) the form of a random effect in the analysis does not imply any practical sense in which X(s) is a draw from a random mechanism.

However, I'd like to do this more closely to traditional spatial analysis. To do so, I must first take a step back.

A probability distribution can be used as a descriptive device.

-- Measure the heights of all US-born 50-year-old males employed by the U of MN
-- Those heights could be described as following a Gaussian density with a particular mu and sigma2.

This does not imply my height is a random draw from N(mu, sigma2).

Rather, this is an aggregate statement about the heights of a group of men.

If we randomly selected one man and measured his height, the height we reported could meaningfully be represented as a draw from N(mu, sigma-squared)

The selected individual's height is fixed; we create the randomness by our method of selecting him.

Returning to the geostatistical example:

We may describe the true X(s) by saying that at locations {s_i}, {X(s_i)} follows N with mean mu and covariance W({s_i},theta), W being some parametric form.

As above, one might argue that the process that produced X(s) was indeed random, so that X(s) represents a draw from a random process.

If so, however, that fact is now irrelevant: X(s) is now fixed, it makes no sense to contemplate further draws, and the actual X(s) are of intrinsic interest.

We merely choose to describe aggregate features of X(s) using a probability distribution.

As with the heights of 50-year-old men, it makes no sense to behave as if the X are a random draw from that distribution.

We may now choose to use that description in trying to estimate true X(s_i) from measurements y(s_i).

Non-Bayesian: That description is simply a part of an analytic method.

Although the analytic method has the form of a random effect, X(s) is in no useful sense generated by a draw from a random distribution.

Bayesian: That probability distribution is a piece of information You choose to use, describing how the unknown X(s) tend to be related to each other.

That probability distribution does not imply that X(s_i) is not fixed, or that a new X(s_i) could be drawn.

We can immediately draw the same implication we drew for the penalized splines:

Simulation experiments should not draw X(s_i) from the distribution we use to describe X(s).

Instead, they should use fixed true X(s), preferably several sets of them chosen to serve the purposes of the simulation experiment.

Simulating a dataset by drawing an X(s) from the random effect distribution used to describe it obliterates precisely the features of X(s) that a simulation experiment would be used to test.

Geostatistical models often include a spatial covariance matrix with a "nugget" of measurement error on top of spatial correlation in underlying X(s).

This mixes together two very different things.

This may be harmless in analyzing a dataset, but it is a serious error if it motivates simulations in which a new X(s) is drawn for each simulated dataset.

There are cases of spatial analysis where it does make sense to consider true but unknown values as generated by draws from a probability distribution.

Hypothetical example:

-- In a given square mile of the Atlantic Ocean, at 1000 meters depth,
-- Every 30 minutes in a given week, we measure water temperature at fixed locations.

It is meaningful to describe the different times as representing draws from a probability distribution.

-- More such draws could be made
-- The draws themselves have no intrinsic interest.

Suppose, however, we are interested in the specific week in which we made these measurements.

The square mile of interest has, for the study's week, a smooth spatial temperature gradient; the half-hourly temperatures vary around it.

The smooth spatial gradient is a fixed feature of our square mile for the study's week, which we happen not to know.

Thus, if we did a simulation experiment comparing methods for estimating this gradient, it would defeat the experiment's purpose to generate each simulated dataset by drawing a new gradient.

There is a meaningful sense in which this fixed (for the study week) but unknown feature of our square mile was drawn from a probability distribution, but that sense is not relevant to the problem of studying this particular week

.

Confession:

The alternative derivation of the Slovenia result ("add the random effect and the fixed effect goes away") put the CAR structure in the error covariance.

This contradicts the point I'm making today:

- -- Municipality i's mean x_i beta + S_i is meaningfully described as looking like a CAR random effect,

- -- But it is senseless to treat S_i as a draw from a random distribution. The true S_i is fixed, but unknown.

- -- Putting the CAR structure in the error covariance implies the true S_i is a random draw that could be re-drawn.

Thus: In a simulation experiment relevant to Vesna Zadnik's problem,

If we simulate cancer counts n_i by

-- drawing S_i from a CAR distribution then
-- drawing n_i from a Poisson distribution with
mean exp(x_i b + S_i),

the estimate of b will be unbiased in an analysis with a CAR-distributed S_i.

But this, I've argued, makes no sense.

Instead, the process that produced the actual cancer counts begins with a fixed but unknown collection of x_i b + S_i and then for municipality i makes a Poisson draw with mean x_i b + S_i.

If we generate simulated data this way, the estimates produced by the same analysis will be biased.

This fact is understood, to some extent, for regular random models.

RWC discusses a related point on p. 139-140.

Their treatment raises some interesting issues – because one part if it is, in my view, erroneous – but that goes beyond the scope of the present paper, and I will stop here.