IMPROVED PARAMETER ESTIMATION AND UNCERTAINTY PROPAGATION IN BAYESIAN RADIOCARBON-DATED EVENT COUNT [REC] MODELS

A PREPRINT

W. Christopher Carleton	David A. Campbell
Extreme Events Research Group	School of Mathematics and Statistics
Max Planck Institutes for	Carleton University
Chemical Ecology,	Ottawa, ON, Canada
The Science of Human History, and	davecampbell@math.carleton.ca
Biogeochemistry	
Jena, Germany	
wcarleton@ice.mpg.de	

December 16, 2021

ABSTRACT

Data about the past contain chronological uncertainty that needs to be accounted for in statistical models. Recently a method called Radiocarbon-dated Event Count (REC) modelling has been explored as a way to improve the handling of chronological uncertainty in the context of statistical regression. REC modelling has so far employed a Bayesian hierarchical framework for parameter estimation to account for chronological uncertainty in count series of radiocarbon-dates. This approach, however, suffers from a couple of limitations. It is computationally inefficient, which limits the amount of chronological uncertainty that can be accounted for, and the hierarchical framework can produce biased, but highly precise parameter estimates. Here we report the results of an investigation in which we compared hierarchical REC models to an alternative with simulated data and a new R package called "chronup". Our results indicate that the hierarchical framework can produce correct high-precision estimates given enough data, but it is susceptible to sampling bias and has an inflated Type I error rate. In contrast, the alternative better handles small samples and fully propagates

uncertainty into parameter estimates. In light of these results, we think the alternative method is more generally suitable for Palaeo Science applications.

Keywords Chronological Uncertainty; Uncertainty Propagation; Palaeo Data Science; Bayesian Regression; Radiocarbon-dated Event Counts

1 Introduction

Scientists studying the past have to contend with significant uncertainty about time. Whereas a physicist conducting an experiment in the present, for instance, now has access to clocks with uncertainties measured in attoseconds (1×10^{-18} of a second) (Pazourek et al., 2015), a scientist studying the past (a “palaeo scientist”) trying to determine the age of something is typically only able to narrow the range of possibilities to within a few centuries at best (Aitken, 2014). In other words, if an attosecond was equivalent to a millimeter, a palaeo science age estimate with 100-year errors on the same scale would be off by more than 300 million light-years. For experiments and measurements at all but the most extreme temporal scales, then, many scientists may largely be able to ignore chronological uncertainty. Palaeo scientists, on the other hand, must make it central to their research, and it affects almost every aspect of the work.

Take, for example, research into the relationship between climate changes and long-term human population dynamics. The former could be represented by a palaeoclimate proxy series, like a speleothem oxygen isotope record, while the latter could be estimated by counting some relevant archaeological indicator, like the number of human skeletons that entered the archaeological record over a given interval of time. To produce the population proxy, individual skeletons would first have to be dated and then those dates could be counted by binning them into a series of temporal intervals. The aggregation of individual skeleton dates into a count series is analytically downstream of the chronometrics and, so, uncertainties in the skeleton ages give rise to uncertainty in the counts. Clearly, uncertainties associated with individual dates should be reflected in the count series. Those counting uncertainties, in turn, should be considered when the whole series is compared to a climate indicator, which has its own chronological uncertainties that should be accounted for as well.

The need to account for chronological uncertainty recently inspired an approach called “Radiocarbon-dated Event Count” (REC) modelling (Carleton, 2020; Stewart et al., 2021). REC modelling refers to count-based regression models in which the dependent count variable is based on radiocarbon dated observations and the uncertainty in those dates is accounted for in the model’s parameter estimates. The approach was proposed because it has become increasingly common in various palaeo-sciences to use the abundance of radiocarbon-dated samples in a given place/time as a proxy for a past process of interest (Carleton and Groucutt, 2020). In archaeology, for instance, through-time variation in the frequency of radiocarbon samples found in archaeological deposits is often thought to reflect past changes in human population size (e.g., Kim et al., 2021; Riris and de Souza, 2021; Timpson et al., 2021), and among palaeoenvironmental scientists it has been used as a proxy for sea-level change and fire regime dynamics (e.g., Bleicher, 2013; Mooney et al., 2011; Thorndycraft and Benito, 2006). In each case, the primary variable is through-time variation in the abundance of

radiocarbon samples in sedimentary archives or deposits—essentially, radiocarbon-dated sample count series where each sample represents the age of a given event. As explained, though, uncertainty in the age estimates for the individual radiocarbon samples produces uncertainties in the count series. As a result, multiple probable count series exist for any given set of radiocarbon sample dates, and REC models are intended to account for this uncertainty in the context of regression analyses. The approach is part of a relatively recent surge in efforts aimed at improving the handling of chronological uncertainty in statistical Palaeo Science research (e.g., Bronk Ramsey, 2017; Crema et al., 2017; Crema and Bevan, 2020; Crema and Kobayashi, 2020; Dolman et al., 2021; Haslett and Parnell, 2008; Manning and Timpson, 2014; Mithen and Wicks, 2018; Parnell et al., 2015; Price et al., 2021; Timpson et al., 2021).

So far, REC models have used a Bayesian hierarchical framework for addressing uncertainty (Carleton, 2020; Stewart et al., 2021). Hierarchical models, as the name suggests, involve parameters that can be arranged hierarchically and they are typically used to model data that are nested in hierarchical or multi-level arrangements, like classrooms of students within schools within districts (Gelman et al., 2013; Gelman and Hill, 2007). In the context of a REC model, the hierarchy is more abstract. Chronological uncertainty in event times means that any given set of radiocarbon-dated events could produce many probable count sequences, or realizations, and the same can be said of potential covariates that also have chronological uncertainty. Differences among probable realizations of dependent and independent variables, then, are partly a consequence of chronological uncertainty. REC models attempt to account for that by treating probable pairs of dependent and independent variables like random draws from a parent population. Each dependent–independent variable pair is placed in its own regression model, and the model parameters are then also thought of as random samples from a parent population. Consider a single parameter, like a regression coefficient for instance. With a hierarchical REC model, the coefficient estimated for a single dependent–independent pairing is a sample from a distribution of possible regression coefficients, and the parent population is defined by a probability distribution that has its own parameters. Those parent population parameters—or, “hyperparameters”—reflect the variation in dependent–independent variable pairings. Since the variation among variable pairs is caused in part by chronological uncertainty, the hyper-parameter estimates are thought to account for that uncertainty. Those hyperparameters, then, are ultimately the main target for statistical estimation and inference in hierarchical REC models. Simulation experiments indicate that this approach is useful and can be used to estimate parameters of interest when the data contain chronological uncertainty (Carleton, 2020).

The hierarchical arrangement imposes two potential limitations, however. The first is largely computational. Hierarchical REC models involve large numbers of parameters that have to be estimated, leading to long computation times. The total number of model parameters that need to be estimated is the number of parameters defined by a given model times the number of sampled dependent–independent variable pairs, plus the model’s hyper-parameters. Take, for example, a model defined by only 2 main parameters, like a linear regression with a mean and standard deviation for a lone regression coefficient. If we sampled only 100 probable dependent–independent variable pairs (to create 100 probable regressions), then, a hierarchical REC model would involve estimating 100 means, one for each sampled variable pair, plus the hyper-parameters at the top of the hierarchy. More parameters in a model’s definition implies

more computation time to estimate values for all of the dependent–independent variable pairs, which ultimately limits the number of those pairs that can feasibly be sampled. That constraint, in turn, limits the amount of chronological uncertainty that can be accounted for.

The second potential limitation is more theoretical and related to the way information is combined in hierarchical models. In hierarchical models, remember, the hyperparameter(s) represent a parent population from which samples have been drawn (Gelman and Hill, 2007). Imagine for instance a parameter of interest, B , that represents a population mean. We gather data from which we can estimate B by taking samples labelled s_j , where j just refers to a particular sample. For each sample, we can generate an estimate of B called β_j , and with a large enough sample we expect that estimate to approach the population value. In a hierarchical model, though, we would take multiple j samples and estimate the β_j ’s simultaneously, combining information to produce an estimate of B . The likelihood that B is a given value, then, depends on the whole set of s_j samples. Since we assume that the β_j values are drawn from the same parent population (characterised by B), extreme values will be down-weighted and all β_j ’s estimated in the hierarchical framework will tend to be pulled together (Gelman et al., 2013; Gelman and Hill, 2007). The pulling together of β_j estimates reduces the variance in the posterior density for B , which is generally considered an advantage of hierarchical models because it can yield high-precision parameter estimates. Essentially, information is being shared across data samples in order to derive a high-precision estimate for B ’s value and distribution. It works well as long as the separate j samples really are separate sources of information about the population in question. In that case, the samples can be thought of as each contributing additional information to the estimate of B , which justifies the increase in precision. But, when used to estimate REC model parameters, the samples are not separate sources of information about the population parameter(s) for the process that produced the events. There is only one set of events (radiocarbon samples) and using a hierarchical approach introduces a bias that might lead to a highly precise estimate of the sample parameter β_j but not necessarily a good estimate of the target population parameter. Consequently, sampling biases could lead to bad inferences.

With these potential limitations in mind, we investigated an alternative framework for handling uncertainty in REC models. The new approach involves a flatter parameter arrangement, and integrates the sampling of probable dependent–independent variable pairs into a Monte-Carlo simulation used to estimate parameter values. Here we present the results of a simulation study in which we compare the hierarchical approach to the alternative one using a newly developed R package called "chronup" (<https://github.com/wccarleton/chronup>).

2 Materials and Methods

2.1 The Negative-Binomial REC Model

We investigated a single type of REC model in order to compare approaches to managing chronological uncertainty. REC model’s assume the target process is a counting process—i.e., the number of objects (or more abstractly, “events”) dated to a given time is a random variable that can be modelled as a function of one or more other variables. A

suitable model for the counting process, then, could be one of the standard counting distributions, like the Poisson or Negative-Binomial. The Negative-Binomial has so far been preferred for REC models (Carleton, 2020; Stewart et al., 2021). It is a two-parameter distribution used for over-dispersed counting processes and autocorrelation (Hilbe, 2011; Zhu, 2011). The prototype REC model employs the Negative-Binomial (NB) in order to account for “temporal spread”, a phenomena whereby any samples drawn from a set of date distributions will tend to be spread out along the time domain with many sampled sequences occupying a longer span of time than the one covered by the true event dates (Bronk Ramsey, 2017; Carleton, 2020; Carleton and Groucutt, 2020).

Let a_i refer to the time of event i —the date of a fossil, seed, piece of charcoal, etc. The event-time for i is uncertain and we can represent the uncertainty with a probability distribution. Often, this distribution will be a radiocarbon date density, which after calibration will typically be highly irregular and multi-modal. For our purposes, though, the distribution could represent any sort of chronological uncertainty and have any functional form, so we will simply refer to it as $\tau(\theta)$ where τ is just a reference to the time-domain and θ refers to whatever set of parameters are appropriate for the distribution. Thus, the probability, $P()$, that event i occurred at time a is

$$P(a_i) \sim \tau_i(\theta). \quad (2.1)$$

Counting the number of i can be expressed by summing an indicator function denoted $I_{(\cdot)}$ that returns 1 if the event occurred (i.e., the condition in the bracketed subscript is met) at time t and 0 otherwise. The count at time t can then be denoted y_t and the total number of events (dated objects) can be denoted by N . Then,

$$y_t|a_i, \dots, a_N = \sum_{i=1}^N I_{(a_i=t)}. \quad (2.2)$$

Given a set of realizations for the individual event-times, $a_i, \dots, a_N$, we can model the counting process with the NB distribution as follows,

$$y_t|a_i, \dots, a_N, r, p \sim NB(r, p). \quad (2.3)$$

In this equation, $NB(\cdot)$ refers to the NB distribution and r and p refer to its standard parameters. The first, r , is typically interpreted as the failures encountered in a series of Bernoulli trials (coin tosses) before the series is stopped, and the second, p , is the probability of success in each trial. The mean of the distribution, μ , is determined by a combination of the two:

$$\mu = r \frac{1-p}{p}. \quad (2.4)$$

Since the aim is to produce a regression model, we need to insert a regression function into the definition for this mean. For that, we can define r_t using a standard linear regression function with covariates, X_t , and one or more regression coefficients, β_t . Then, to account for temporal spread, the second NB parameter, p , can be allowed to vary with time. So it becomes p_t and with the inclusion of the regression function equation 2.3 becomes

$$y_t | a_i, \dots, a_N, r_t, p_t \sim NB(r_t, p_t), \quad (2.5)$$

where r_t is defined as follows,

$$r_t = e^{\beta X_t}. \quad (2.6)$$

Equations 2.5 and 2.6 together define the core prototype NB-REC model published previously (Carleton, 2020).

2.1.1 Hierarchical framework

The NB-REC models used so far have relied on a Bayesian hierarchical framework for including chronological uncertainty, as explained. Given the uncertainty in individual dates, there will be many probable y_t counts for any given time and, so, many probable count sequences (or vectors, denoted $\mathbf{y}$). For simplicity, we can conceive of the individual probable count sequences ($\mathbf{y}$) as each a single random draw from a joint-density describing chronological uncertainty for all the relevant events. The joint-density is defined by the set of chronological uncertainty distributions (each a $\tau(\theta)$ from earlier) corresponding to the events in a given sample. That joint-density represents the probability, for example, that one event is dated to 1000 BC, another to 1200 BC, a third to 1250 BC, etc. As before, we will denote the age of event i with a_i and there will be N events in a given database. Referring to the joint-density of all N corresponding chronological uncertainty distributions as $\phi(\cdot)$, the density describing the set of probable event times for all events would be

$$P(a_1, a_2, \dots, a_i) \sim \phi(\boldsymbol{\theta}) | i \in \{1, 2, \dots, N\}, \quad (2.7)$$

where $\boldsymbol{\theta}$ is a vector of parameters for each of the individual chronological distributions.

In the hierarchical framework, a set of J random draws from $\phi(\boldsymbol{\theta})$ are used to construct J corresponding count sequences using the indicator function and summation in equation 2.2. The randomly sampled times can be denoted by a vector $\mathbf{a}_j = \{a_{j,i}, \dots, a_{j,N}\}$ and the corresponding count sequences by $\mathbf{y}_j$. Along similar lines, we can imagine drawing J sets of covariate observations, X_j , from a given joint-density of the covariate measure(s) and associated chronological uncertainties. Together, these random samples comprise J probable dependent-independent variables pairs: a $\mathbf{y}_j$ for each X_j . To represent these sampled pairings in a REC model regression, we can re-write the core model from equations 2.5 and 2.6 as follows,

$$y_{j,t} | \mathbf{a}_j, r_{j,t}, p_{j,t} \sim NB(r_{j,t}, p_{j,t}) \quad (2.8)$$

$$r_{j,t} = e^{\beta_j X_{j,t}}. \quad (2.9)$$

This set of J models sit in the bottom level of the NB-REC model hierarchy. The set includes J regression coefficients (i.e., the β_j 's) and $J \cdot T$ NB 'p' parameters.

The hierarchical framework assumes that each of the J regression coefficients are drawn from a parent population represented by hyper-parameters that sit at the top of a hierarchy. Assuming that the β_j regression coefficients are drawn from a normal (Gaussian) hyper-parameter distribution, then we can represent the hyperparameter formally with

$$\beta_j \sim N(\mu, \sigma), \quad (2.10)$$

which can also be thought of as a prior placed on the β_j 's. Then, in the context of most REC models, the key targets for parameter estimation and inference become μ and σ . The first parameter, μ , represents the mean of the sample of lower-level β_j 's and the second, σ , represents the sample's variability. Since variability among the β_j 's is a consequence of chronological uncertainty—each coefficient is estimated for a sampled $\mathbf{a}_j$ set of event dates—the normal distribution with mean μ and standard deviation σ reflects that uncertainty.

It should be emphasized that the chronological spread parameter, $p_{j,t}$, is not embedded in a hierarchy. It improves the fit of the model, reduces the bias in the estimation of the β_j 's parent distribution, and accounts for the known effect of chronological spread on count models. But, it is of no special interest, at least not in the archaeological and palaeoenvironmental science applications that inspired the exploration of REC models. Consequently, there is no hyperparameter to estimate in relation to $p_{j,t}$. That said, it is necessary to estimate values for the parameter and given the hierarchical paradigm it actually accounts for most of the parameters that have to be estimated—again, there will be JT of them to estimate for any given model compared to only J regression coefficients for each covariate plus an intercept if it is included.

As with many other Bayesian models, NB-REC model parameters are estimated with a Markov-Chain Monte-Carlo (MCMC) simulation. For the sake of clarity, imagine a model involving a single regression coefficient for only one covariate with no intercept. Recall that J probable event count sequences, y_j , are involved along with J probable covariate sequences, X_j . Thus, there would be J dependent–independent variable pairs to which J regression models would be fit and, therefore, J regression coefficients each denoted β_j . Again, these β_j ’s are thought to come from a parent population with a distribution and we can refer to that parent distribution as B with mean μ and standard deviation σ . An MCMC simulation for estimating the model’s parameters involves randomly trying potential values and keeping or discarding those guesses based on how likely they are given the data and the probability model (in this case an NB count regression).

The MCMC simulation for estimating the posterior distributions of B ’s parameters can be broken down into a series of steps. To keep the description uncluttered, we will concentrate on the algorithm with respect to just one variable, μ . The simulation begins by randomly drawing a single value from a prior distribution for μ (i.e., the guess). Next, that random draw is used as a guess for all the β_j ’s at once and its likelihood is calculated for all of the J dependent–independent variable pairs. A new guess for μ is made, and its likelihood is calculated and compared to the previous one. If the new one is higher, the corresponding new guess for μ is saved with a pre-determined probability in a ‘chain’—a sequence of guesses for the parameter value that were kept. Then, the whole process is repeated a large number of times and, after a settling-in period, the stored values in the chain constitute samples from μ ’s posterior distribution. The key takeaway here is that the J sampled dependent–independent variable pairs are static throughout the simulation process. They are treated like separate sets of observations that come from a parent population of potential pairs, and the simulation is trying to find the distribution of likely values for all the J regression coefficients (and other model parameters) simultaneously.

2.1.2 Alternative framework

An alternative sampling framework could be used to side-step the potential limitations of the hierarchical approach. It would involve viewing the probable count sequences (and covariates) as random variables instead of samples from a parent population. Referring back to equation 2.1, the joint-density representing chronological uncertainty for a given set of N events could be written as $\phi(\theta)$. Recall that a single sample from this joint-density would be a vector containing N dates, one probable date for each event in the set. Under the hierarchical sampling framework, each sample would be used to construct one of the J probable count series, $\mathbf{y}_j$. Each of the probable count series can potentially be produced by many different samples from $\phi(\theta)$, but for the purposes of the regression model the many-to-one correspondence does not matter—which specific events contribute to the count at a given time does not affect the model estimates. We can again refer to the count at a given time as y_t . Its distribution is proportional to the slice of $\phi(\theta)$ at the relevant time, t . Thus, by extension,

$$y_t | a_i, \dots, a_N, r_t, p_t \sim NB(r_t, p_t), \quad (2.11)$$

$$r_t = e^{\beta X_t}. \quad (2.12)$$

Compared to the hierarchical model (eq. 2.8–2.10), the model above is simpler. It has no j subscripts and there is only one β to estimate rather than J of them. Similarly, there are only T rather than JT spread parameters (p) of the NB distribution to estimate.

The MCMC simulation for estimating parameters from this non-hierarchical model proceeds in a slightly different way than the simulation for the hierarchical one. For simplicity, again imagine a single model with one regression coefficient and no intercept. We can then outline the algorithm with respect to a single model parameter, β , bearing in mind that the process is the same for the others.

As before, the MCMC simulation can be described by a series of steps. First, the algorithm randomly draws a value for β . Then, it samples a probable count sequence, $\mathbf{y}$, by drawing from $\phi(\theta)$ and counting the number of events that fall into a series of pre-defined intervals (i.e., binning the sampled event dates). Likewise, a probable covariate sequence is sampled with respect to the same set of intervals, taking into consideration measurement and chronological uncertainty. Next, the likelihood of the guess for β is calculated given the probable dependent–independent variable pair and the probability model specified in equations 2.11 and 2.12. This sampling process is repeated, and a new likelihood for a new β is calculated given a new probable dependent–independent variable pair. Lastly, the new likelihood is compared to the previous one and the new guess for β is retained with a pre-determined probability if its likelihood is higher than that of the previous β guess. The whole process is repeated a large number of times to generate MCMC chains containing samples from the posterior distribution for β (and the other parameters, of course). The key take-away from this brief description is that only one β distribution is being estimated instead of J of them. And, rather than estimating parameters for a fixed sample of probable dependent–independent variable sequences, those sequences are treated explicitly like instances of a random variable and sampled along with the model parameter values.

In practice, sampling of dependent–independent variable sequences could happen before the MCMC simulation proceeds. This separation may be necessary in order to create a general computational procedure/toolkit because the uncertainties defining $\phi(\theta)$ for the dependent count sequences and covariates can be very complex (e.g., Bayesian chronological models for radiocarbon dates). The complexity would likely require custom MCMC algorithms to be written if sampling $\phi(\theta)$ was intended to occur within the MCMC for the regression model. Instead, we wrote an algorithm that takes two matrices as inputs, one containing probable event-count sequences ($\mathbf{y}$) and the other containing probable covariate sequences (X). This approach means that radiocarbon-date calibration and other chronological modelling complexity can be considered apart from the regression parameter estimation. It also means that the MCMC simply walks over two matrices instead of directly sampling probable dependent–independent variable pairs at each

iteration of the simulation. Still, the overall estimation framework is essentially the one we just outlined with the same implications for parameter estimation, inference, and interpretation.

2.2 Simulation

We ran a straightforward simulation study to compare the two estimation frameworks. First, we simulated some data from a simple process. This process had a Poisson distribution with a mean defined by a regression function dependent only on time (i.e., a linear increasing function as the sole covariate). Thus, the count y at time t was a random variable with the following conditional distribution,

$$y_t | \lambda_t \sim \text{Pois}(\lambda_t), \quad (2.13)$$

$$\lambda_t = e^{\beta X_t}, \quad (2.14)$$

where β is the sole regression coefficient and X is a discrete sample from a linear function. Equations 2.13 and 2.14, then, describe the true, underlying process responsible for a sequence of counts. This process resulted in a simple exponential increase in event counts (see Figure 1), which was sampled to produced a simulated set of true event dates and a corresponding count sequence with β fixed to 0.004.

Figure 1: Simulated process and sample data.

Next, chronological uncertainty was added to the simulated data. The sampled event dates were passed to a function in the IntCal package in R (Blaauw, 2021). This function produced a set of plausible radiocarbon determinations

(uncalibrated radiocarbon date densities) that were then calibrated with another function from the same package. The resulting joint-density of event dates was sampled to produce a set of probable event-count sequences. A similar process was used to add chronological uncertainty to the sole covariate used in the analysis. The process involved adding a random deviation produced with a Normal distribution to each value in the covariate sequence. This had the effect of causing temporal bins to swap positions relative to their true arrangement. For every probable event count sequence sampled, then, there was also a probable covariate sequence—a probable dependent–independent variable pair.

Lastly, NB-REC model parameters were estimated and compared under the two propagation frameworks in two simulation experiments. The key parameter of interest was the sole regression coefficient, β , which was used as the target benchmark for evaluating the two approaches. There are, of course, several parameters that would affect the level of chronological uncertainty in our simulation, including chiefly individual radiocarbon sample measurement errors and features of the relevant radiocarbon calibration curve. In the present study, though, our main concern was determining how each model handles chronological uncertainty and what the differences in their estimates might be holding external sources of that uncertainty constant. So, we focused here on only two simulation parameters, namely the number of probable dependent-independent variable pairs involved and the size of the event sample (i.e., number of radiocarbon dates involved). The former is relevant because it indicates how much chronological uncertainty in a given sample of dated events is being accounted for. The latter is relevant because, as we explained in the introduction, hierarchical models in particular might produce highly-precise but biased estimates of model parameters when samples are biased, which means event sample size may be important for understanding how each framework handles chronological uncertainty.

We ran two experiments to compare the frameworks. In the first, we looked at the impact of varying the number of dependent–independent variable pairs sampled for the hierarchical model. We denoted that sample size by J and drew samples of four sizes for the hierarchical model: $J = 5$, $J = 10$, $J = 50$, and $J = 100$. For this experiment, we included $N = 1000$ randomly sampled events. The posterior density estimates for β given the different sizes for J were then compared to one another and to the posterior estimate from the non-hierarchical model. Then, in the second experiment, we investigated the impact of increasing the number of events included from $N = 1000$ to $N = 5000$. This sample size, remember, refers to the number of events used to produce the event-count time-series, not the number of variable pairs sampled from the joint-density representing chronological uncertainty. As before, we compared the posterior densities for the key rate parameter across the experimental conditions, including different J sample sizes for the hierarchical model and all of those results were compared to the non-hierarchical estimate.

All analyses were conducted in R (R Core Team, 2021) and re-run several times. For the hierarchical framework, an R package called "Nimble" (de Valpine et al., 2021, 2017) was used for parameter estimation and the "chronup" package was used for the alternative framework. Each hierarchical MCMC was run for at least 1 million iterations and each "chronup" MCMC was run for 200,000 iterations. Convergence of the resulting MCMC chains in both cases was checked visually and with the help of standard Geweke (Geweke, 1992) diagnostics available in the "coda" package (Plummer et al., 2006). The code used for this study is available on GitHub (github.com/wccarleton/chronupintro) as is

the pre-release version of the "chronup" package (github.com/wccarleton/chronup). Several other R packages were instrumental including "bigmemory" (Kane et al., 2013), "parallel" (R Core Team, 2021), and "ggplot2" (Wickham, 2016). The state of the GitHub repositories for the scripts used in the present analysis and the code behind the "chronup" package will be preserved as archived snapshots with Zenodo (zenodo.org/) and given DOIs (DOIs TBD after acceptance). In addition, all of the analyses can be easily replicated by following the detailed instructions in the Supplementary Information provided alongside this article.

3 Results

The results indicate that the two approaches to handling uncertainty produced results that were similar in some ways and quite different in others. Both approaches produced estimates of the posterior distribution for β that included the target parameter value (0.004) within the 99% and/or 95% credible intervals or were very close to it. As can be seen in Figures 2 and 3, the means of each estimate were also roughly in agreement irrespective of the size of the dependent-independent variable sample (indicated by the variable J in the plot labels), the event count sample size (N), or the estimation framework.

Figure 2: Posterior densities (scaled to 1) for β across experimental conditions given the smaller event sample ($N = 1000$). The number at the top of each plot indicates the number of probable dependent-independent variable pairs used to estimate the posterior density with the hierarchical framework. The top four plots show the densities estimated with the hierarchical framework and the bottom blot shows the density estimated with the alternative framework. The vertical dashed lines indicate the target parameter value.

Nevertheless, important differences with respect to bias and variance were also evident. The posterior distributions produced by the hierarchical approach were much narrower in most cases than the ones produced with the alternative approach. For samples with more than 5 dependent-independent variable pairs ($J > 5$), the difference in variances between the two approaches were quite dramatic when sample sizes were smaller (see Figure 2). Often when running the hierarchical model using variable pair samples larger than 10 (i.e., $J > 10$) the posterior distributions for β would exclude the true value at the usual levels of confidence (95% and 99%). When event sample size (N) was larger, though, the posteriors for the hierarchical models included the true value while retaining their high precision (see Figure 3). The alternative approach from the "chronup" package, in contrast, produced much wider posteriors irrespective of event sample size and those posteriors nearly always included the true value for β within the 95% and 99% credible regions.

Figure 3: Posterior densities (scaled to 1) for β across experimental conditions given a larger event sample ($N = 5000$). The number at the top of each plot indicates the number of probable dependent-independent variable pairs used to estimate the posterior density with the hierarchical framework. The top four plots show the densities estimated with the hierarchical framework and the bottom blot shows the density estimated with the alternative framework. The vertical dashed lines indicate the target parameter value.

4 Discussion and Conclusion

The two estimation frameworks handle chronological uncertainty in very different ways. Chronological uncertainty in both cases is initially summarized by a random sample of probable count sequences drawn from the joint-density of all the calibrated radiocarbon dates for a given set of events. But the way that information affects parameter estimates differs between the two approaches. The hierarchical model combines information about event counts from a limited sample of those probable sequences to produce a high-precision estimate of the target parameter. In contrast, the

alternative framework, implemented in the R package "chronup", uses a much larger sample of probable sequences and does not combine the chronological information in a way that constrains parameter estimates. As a result of the differences, the hierarchical estimates generally have a lower variance than the non-hierarchical estimates. But, the hierarchical ones can be biased and the increased precision can cause the relevant posterior density to exclude the target population parameter value at the usual levels of confidence, especially when small event samples are involved.

The difference is an example of the bias–variance trade-off. The trade-off is sometimes discussed in slightly different terms in different fields (e.g., theoretical statistics compared to machine learning), but for the present discussion we will think in terms of model performance with respect to the bias and variance of its out-of-sample predictions—i.e., its estimates of the value(s) of unobserved data given its estimated parameter(s) based on a training data sample (Hastie et al., 2017). For those unsure, bias in this context refers to systematic offsets from the true value a model is trying to estimate (i.e., the population parameter), while variance refers to the average size of the difference between the mean predicted value and any individual prediction. It would be optimal to create a model that had both low prediction bias and low prediction variance, but it is typically thought to be impossible to reduce both at the same time. They trade off.

Prediction in Palaeo Science is important and understanding why will help to make sense of how the trade-off manifests in the two estimation frameworks. Even though the aim in Palaeo Science research is often to explain a pattern in a fixed data-set (e.g., a database of radiocarbon determinations), using a statistical model to explain those patterns entails making a "prediction" about what sort of patterns would be observed if we repeated the data gathering process, even if new data are never actually gathered. Imagining this hypothetical do-over is how we generalize from the single observed data-set to the process that produced it. This is the type of reasoning we must use, for instance, to infer something about a general process involving through-time fluctuations in radiocarbon-dated event counts on one hand and climate changes on the other. If a given model fit our observations very well, but would likely fail to adequately predict new data, we would have to conclude that the model described patterns that cannot be generalized. Such a model would not accurately reflect the population/process that produced the training data or any subsequently gathered data (hypothetical or real).

The hierarchical framework for REC models is an instance of decreased variance at the expense of enlarging the effect of bias. Sharing of information across probable event count sequences leads to lower variance posterior distributions for a REC model's regression coefficients. But, as the posterior densities we plotted demonstrate, those low-variance, high-precision estimates can be biased. The model is fit very tightly to the training data, producing a high-precision estimate of the sample regression coefficient. It may, therefore, fail to predict out-of-sample observations more frequently than typically expected because of the difference between a given training sample's properties and those of the population (or process). Essentially, the hierarchical framework will be more sensitive to sampling fluctuations. Of course, given an adequate sample, estimated parameters will be close to the population parameters. In our experiment with a much larger sample size, for example, the hierarchical model does capture the true population parameter in its posterior estimates despite the low variance. Nevertheless, by conditioning on J sequences of events as though they were observed exactly rather than being random variables, the uncertainty in the data is misrepresented and variance

reduced at the expense of unknown bias. The conditioning also means the model's credible intervals (CIs) no longer refer to the target population parameter, decoupling the meaning of CIs from their intended definition (i.e., the interval within which an unobserved parameter falls with a given probability, usually 95% or 99%). This, in turn, increases the chances that the population parameter falls outside the 95% or 99% credible region of the model's posterior density even with larger samples. Sample size (really, adequacy) is affecting the bias side of the trade-off and the number of probable variable pairs is affecting the variance side. The resulting inflation in prediction error can still occur, though, because differences between the sample estimate and the population parameter value will never be zero, except in the limit of infinite sample sizes. Ultimately, then, the increase in precision from using the hierarchical framework comes at the cost of an increase in Type I errors (false positives).

In contrast, the alternative framework for REC model parameter estimation favours variance over bias. The MCMC implemented in "chronup" selects a new probable count sequence (and corresponding covariate sequence) each iteration. So, every time it estimates the likelihood of a value for a model parameter, it is comparing the previous value's likelihood to the new value given a new dependent-independent variable pair. If the new value is kept, it gets added to the MCMC chain as a sample from the relevant posterior distribution. Thus, a new value may be kept even if the randomly selected variable pair under consideration differs greatly from the others. Consequently, there is no "pulling together" of parameter estimates as occurs in hierarchical models, and the result is a posterior density estimate that is wider than the one estimated with the hierarchical approach. Importantly, it can also be expected to widen with increasing variance in the joint-density representing chronological uncertainty in the relevant event times. More variability among the probable event count sequences (and corresponding covariate sequences) leads to greater variability in the model's posterior density estimates. This means that uncertainty from any source, including chronological errors, leads ultimately to higher-variance parameter estimates. Put another way, the alternative framework involves "uncertainty propagation", which means ensuring uncertainties associated with primary observations are passed up the chain of inference to the model's parameter estimates. In sum, the alternative framework is more likely to produce estimates that include the true population parameter value(s) even with smaller/less-adequate samples and increased observation uncertainty will increase parameter uncertainty. We may lose precision when compared to the hierarchical estimates, but the benefit is a lower Type I error rate with fewer assumptions about sample adequacy.

With these differences and the bias–variance trade-off in mind, we recommend the alternative estimation framework for REC modelling in most Palaeo Science applications. There are really two key reasons behind this recommendation. First, Palaeo Science samples are typically small and biased in ways that are difficult to identify and control for. Thus, it may be hard to avoid biased parameter estimates for REC models that use a hierarchical estimation framework. And, second, false positive findings are already more likely to make it into the scientific literature for a number of reasons. Therefore, since prominent scientists have argued that Palaeo Science research has relevance to current issues facing the world (e.g., Andermann et al., 2020; Caseldine and Turney, 2010; Mitchell, 2008; Shennan and Sear, 2021), a conservative approach to inferential errors is preferable. It is better to be uncertain than certainly wrong if research is going to have real-world consequences.

We see two key avenues for future research into REC models. One involves a fuller exploration of key parameters likely to affect REC model analyses. These include looking at the effect of calibration curve features on the model estimates, identifying grounded justifications for selecting one particular temporal sub-interval instead of another in a real-world setting where the temporal span of the true process is unknown, and exploring the effect of combining radiocarbon-dating and other types of chronometric uncertainties (e.g., Uranium-Thorium or Optical Stimulated Luminescence dating) into the same analysis.

The other avenue for future research involves exploring and developing other kinds of regression models that can be added to the "chronup" toolkit. As of the time of writing, the package supports only two kinds of count model (the Poisson and Negative-Binomial for REC models). Neither model currently supports autocorrelation terms, which means that dependent variable autocorrelation cannot be accounted for, potentially leading to biased parameter estimates in some real-world cases. So, autocorrelation count models should be explored and added to the available suite of tools. Expansion of the tool kit will make it more attractive and widely applicable, hopefully leading to uncertainty propagation becoming the norm in Palaeo Science research.

5 Acknowledgements and Funding

To be added after acceptance.

References

- Aitken, M.
2014. *Science-Based Dating in Archaeology*. Routledge.
- Andermann, T., S. Faurby, S. T. Turvey, A. Antonelli, and D. Silvestro
2020. The past and future human impact on mammalian diversity. *Science Advances*, 6:eabb2313.
- Blaauw, M.
2021. *IntCal: Radiocarbon Calibration Curves*. R package version 0.2.2.
- Bleicher, N.
2013. Summed radiocarbon probability density functions cannot prove solar forcing of Central European lake-level changes. *Holocene*, 23(5):755–765.
- Bronk Ramsey, C.
2017. Methods for Summarizing Radiocarbon Datasets. *Radiocarbon*, 59(6):1809–1833.
- Carleton, W. C.
2020. Evaluating bayesian radiocarbon-dated event count (rec) models for the study of long-term human and environmental processes. *Journal of Quaternary Science*, 36:110–123.

Carleton, W. C. and H. S. Groucutt

2020. Sum things are not what they seem: Problems with point-wise interpretations and quantitative analyses of proxies based on aggregated radiocarbon dates. *The Holocene*.

Caseldine, C. J. and C. Turney

2010. The bigger picture: Towards integrating palaeoclimate and environmental data with a history of societal change. *Journal of Quaternary Science*, 25:88–93.

Crema, E., A. Bevan, and S. Shennan

2017. Spatio-temporal approaches to archaeological radiocarbon dates. *Journal of Archaeological Science*, 87:1–9.

Crema, E. R. and A. Bevan

2020. Inference from large sets of radiocarbon dates: Software and methods. *Radiocarbon*, 00:1–17.

Crema, E. R. and K. Kobayashi

2020. A multi-proxy inference of jōmon population dynamics using bayesian phase models, residential data, and summed probability distribution of 14c dates. *Journal of Archaeological Science*, 117:105136.

de Valpine, P., C. Paciorek, D. Turek, N. Michaud, C. Anderson-Bergman, F. Obermeyer, C. Wehrhahn Cortes, A. Rodríguez, D. Temple Lang, and S. Paganin

2021. *NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling*. R package version 0.11.1.

de Valpine, P., D. Turek, C. Paciorek, C. Anderson-Bergman, D. Temple Lang, and R. Bodik

2017. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26:403–413.

Dolman, A. M., T. Kunz, J. Groeneveld, and T. Laepple

2021. A spectral approach to estimating the timescale-dependent uncertainty of paleoclimate records - part 2: Application and interpretation. *Climate of the Past*, 17:825–841.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin

2013. *Bayesian Data Analysis*, 3rd edition. Boca Raton: CRC Press.

Gelman, A. and J. Hill

2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.

Geweke, J.

1992. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In *Bayesian Statistics*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., Pp. 169–193. Oxford: Clarendon Press.

Haslett, J. and A. Parnell

2008. A simple monotone process with application to radiocarbon-dated depth chronologies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(4):399–418.

Hastie, T., R. Tibshirani, and J. Friedman

2017. *The Elements of Statistical Learning*. Springer New York.

Hilbe, J. M.

2011. *Negative Binomial Regression*. Cambridge University Press.

Kane, M. J., J. Emerson, and S. Weston

2013. Scalable strategies for computing with massive data. *Journal of Statistical Software*, 55(14):1–19.

Kim, H., G.-A. Lee, and E. R. Crema

2021. Bayesian analyses question the role of climate in Chulmun demography. *Scientific Reports*, 11(1):23797.

Manning, K. and A. Timpson

2014. The demographic response to holocene climate change in the sahara. *Quaternary Science Reviews*, 101:28–35.

Mitchell, P.

2008. Practising archaeology at a time of climatic catastrophe. *Antiquity*, 82:1093–1103.

Mithen, S. and K. Wicks

2018. The interpretation of mesolithic structures in britain: New evidence from criet dubh, isle of mull, alternative approaches to chronological analysis for inferring occupation tempos settlement patterns. *Proceedings of the Prehistoric Society*, 84:77–110.

Mooney, S. D., S. P. Harrison, P. J. Bartlein, A. L. Daniau, J. Stevenson, K. C. Brownlie, S. Buckman, M. Cupper, J. Luly, M. Black, E. Colhoun, D. D’Costa, J. Dodson, S. Haberle, G. S. Hope, P. Kershaw, C. Kenyon, M. McKenzie, and N. Williams

2011. Late Quaternary fire regimes of Australasia. *Quaternary Science Reviews*, 30(1-2):28–46.

Parnell, A. C., J. Sweeney, M. Salter-townshend, J. R. M. Allen, B. Huntley, and J. Haslett

2015. Bayesian inference for palaeoclimate with time uncertainty and stochastic volatility. *Applied Statistics*, 64(1):1–24.

Pazourek, R., S. Nagele, and J. Burgdörfer

2015. Attosecond chronoscopy of photoemission. *Reviews of Modern Physics*, 87:765.

Plummer, M., N. Best, K. Cowles, and K. Vines

2006. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

Price, M. H., J. M. Capriles, J. A. Hoggarth, R. K. Bocinsky, C. E. Ebert, and J. H. Jones

2021. End-to-end bayesian analysis for summarizing sets of radiocarbon dates. *Journal of Archaeological Science*, 135.

R Core Team

2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Riris, P. and J. G. de Souza

2021. Formal Tests for Resistance-Resilience in Archaeological Time Series. *Frontiers in Ecology and Evolution*, 9:906.

Shennan, S. and R. Sear

2021. Archaeology, demography and life history theory together can help us explain past and present population patterns. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376:20190711.

Stewart, M., W. C. Carleton, and H. S. Groucutt

2021. Climate change, not human population growth, correlates with late quaternary megafauna declines in north america. *Nature Communications*, 12:965.

Thorndycraft, V. R. and G. Benito

2006. The Holocene fluvial chronology of Spain: Evidence from a newly compiled radiocarbon database. *Quaternary Science Reviews*, 25(3-4):223–234.

Timpson, A., R. Barberena, M. G. Thomas, C. Méndez, and K. Manning

2021. Directly modelling population dynamics in the South American Arid Diagonal using 14 C dates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1816):20190723–20190723.

Wickham, H.

2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Zhu, F.

2011. A negative binomial integer-valued GARCH model. *Journal of Time Series Analysis*, 32(1):54–67.