Title: A Radiocarbon Test for Demographic Events in Written and Oral History

Authors: Kevan Edinborough* (1), Marko Porčić (2), Andrew Martindale (3), T. J. Brown (3), Kisha Supernant (4), and Kenneth M. Ames (5).

Institution

(1) Institute of Archaeology, University College London, 31-34 Gordon Square, London, WC1H 0PY, United Kingdom.

(2) Department of Archaeology, Faculty of Philosophy, Čika Ljubina 18-20, 11000 Belgrade, Serbia.

(3) Department of Anthropology, University of British Columbia, Vancouver Campus, 6303 NW Marine Drive, Vancouver, British Columbia, Canada, V6T 1Z1.

(4) Department of Anthropology, 13-15 HM Tory Building, University of Alberta, Edmonton, Alberta, Canada, T6G 2H4.

(5) Portland State University, ANTH, P.O. Box 751, Portland, OR 97207.

Corresponding author

*Corresponding author: k.edinborough@ucl.ac.uk

Abstract

We extend an established simulation-based method to test for significant short duration (1-2 centuries) demographic events known from one documented historical and one oral historical context. The first case-study extrapolates population data from the Western historical tradition using historically derived demographic data from the catastrophic European Black Death bubonic plague (*Yersinia pestis*). We find a corresponding statistically significant drop in absolute population using an extended version of a previously published simulation method. Case-study two uses this refined simulation method to test for a settlement gap identified in oral historical records of descendant Tsimshian First Nation communities from the Prince Rupert Harbour (PRH) region of Pacific Northwest region of British Columbia, Canada. Using a regional database of n=523 radiocarbon dates, we find a significant drop in relative population using the extended simulation-based method consistent with Tsimshian oral records. We conclude that our technical refinement extends

the utility of radiocarbon simulation methods, and can provide a rigorous test of demographic predictions derived from a range of historical sources.

Keywords

Historical Record, Oral History, Archaeology, Simulation, Radiocarbon.

Significance Statement

Indigenous oral traditions remain a very controversial source of historical knowledge in Western scientific, humanistic and legal traditions. Likewise, demographic models using radiocarbon-based simulation methods are controversial. We rigorously test the historicity of indigenous Tsimshian oral records (*adawx*) using an extended simulation based method. Our methodology is able to detect short duration (1-2 centuries) demographic events. First we successfully test the methodology against a simulated radiocarbon data set for the catastrophic European Black Death/bubonic plague (*Yersinia pestis*). Second we test the Tsimshian *adawx* accounts of an occupational hiatus in their territorial heartland ca. 1500–1000 years ago. We are unable to disconfirm the oral accounts. This represents the first formal test of indigenous oral traditions using modern radiocarbon modelling techniques.

Introduction

We extend an established simulation-based method to test for significant regional scale demographic events known from documentary historical and oral historical sources. Simulation-based models based on real archaeological data-sets are proving increasingly useful for identifying population related changes in archaeological contexts (1–3). Such approaches offer a far more rigorous statistical assessment of a given demographic question than was previously possible (4).

Well-deployed simulation based demographic approaches have two main strengths. Firstly, data simulation can potentially account for the ubiquitous archaeological problem of finite, small sample sizes that diminish over time (5–7). Secondly, because simulation based approaches can avoid qualitative assessments of patterns within Summed Probability Distributions (SPDs), they can mitigate the thorny issue of confirmation bias. This problem is one long recognized by psychologists, wherein the influence of a favored hypothesis

inadvertently biases the choice of data and model selected by a researcher (8,9). The converse issue is one of rejection bias, where researchers reject an unfavorable model out of hand, without adequately considering or even replicating it (10).

Here we attempt to explicitly avoid these biases and encourage more researchers to follow our lead when using SPDs as a proxy for demographic signatures. We extend the methodological reach of a widely cited simulation-based demographic method (1,4). Then we test this method against the historically well documented population decline in 14th century Europe that was caused by the bubonic plague (*Yersinia pestis)* or 'Black Death' (11). We find support for this particular simulation-based approach using the established (known) data of this historical context following previously contested concerns about this approach raised by an earlier study (4,10). Using a newly collated radiocarbon dataset containing 523 results, we then apply a more conservative version of the same simulation method to test for a shorter duration demographic-settlement gap, known from the oral-historical record in Tsimshian territory in the Prince Rupert Harbour (PRH) Region of Northern British Columbia, Canada (12). The results of this test suggest that significant drops in relative population identified using the simulation-based method is also consistent with Tsimshian oral records. This paper presents one of the first cases of the rigorous testing of an indigenous oral record against demographic data derived from a statistically robust model. The absence of such tests is a common criticism of the use of Indigenous oral records in archaeology (13, 14). As we find support for demographic events extrapolated from both oral and historical records, we conclude that these simulation-based demographic models are consistent with other lines of evidence, which suggests that our results have considerable explanatory power. To encourage more researchers to use this approach, we include the associated freeware R code and data and a summarized explanation of the methods in Supplementary Information.

As the methods used here are advancing apace, the research lineage of our particular simulation approach is important to note. The following methodological progression is underpinned by the fundamental belief that population dynamics can be recovered from the archaeological record, given a sufficient observed sample of dated human activity. Our position is that whilst this sample itself may be a skewed approximation of true population

levels and dynamics, our results will reflect the underlying population signal if they meet the strenuous criteria set by sufficiently rigorous methodological protocols.

Uncalibrated radiocarbon dates have long been used as evidence around the world for inferring general human settlement patterns (15–17), and this paper builds out of this approach. These tentative First Order approaches always come with stated cautions and caveats. Recently, given the increasing availability of computer power, the promise of the approach has encouraged a controversial demographic turn in archaeology (18–21). Initially, uncalibrated radiocarbon data were simply collated from subsets of a defined geographical region of archaeological interest, and then summed over one to produce a temporally coarse-grained histogram, a time-series of the relative intensity of uncalibrated radiocarbon data (22). After a comprehensive radiocarbon analysis of a well-excavated prehistoric region of southern Scandinavia by Edinborough (23–25), Shennan and Edinborough (26) summed and calibrated discreet bins of archaeological radiocarbon results from across northern and central Europe, to produce a broad scale calibrated population model spanning selected parts of the Neolithic transition there. Radiocarbon dates were binned in this way into archaeologically determined units, or phases, to avoid inadvertent sampling biases caused by oversampling of specific sites or periods. Collard et al. (19) developed the method further, summing the calibrated archaeological radiocarbon date bins, or phases, producing a new demographic boom and bust model for the Neolithic transition of Great Britain. The potentially confounding effects of exponential human growth rates (4,27,28), and archaeological site formation processes producing a general exponential taphonomic loss over time of archaeological data (6,7), necessitated a further refinement of this method. The most sophisticated method which accounts for research bias, taphonomic loss, and the long-term population trends was developed by Shennan et al. (1). The research bias is reduced by the specific binning procedure, which gives equal weight to sites/site phases with differential numbers of dates. To account for the effects of taphonomy and long-term population growth on the empirical curve, an exponential model is fitted to the empirical curve by regression. The resulting exponential model is used as a null model against which the empirical SPD is statistically evaluated.

Method

In order to assess the statistical significance of the deviation of the empirical curve from the null model, a large number of simulated radiocarbon datasets is generated by randomly sampling calendar dates from the specified time interval according to the probabilities given by the null model (see R Code in Supplementary Information). The number of dates for each simulated dataset is equal to the number of bins in the empirical dataset. The sampled calendar dates are "back calibrated" by simulating a radiocarbon date which might have produced the particular calendar date. The "back calibrated" dates are then re-calibrated and summed. This procedure is repeated several thousand times in order to create a distribution of simulated values for each moment in time.

In order to assess the statistical significance of the empirical SPD pattern, the empirical curve is compared to the 95% percentile intervals calculated from the simulated data for each year. For time intervals where the empirical summed calibrated probability distribution is above or below the simulated 95% confidence intervals (CI), there is a statistically significant growth or decline, respectively, of population relative to the null model. Given that in 5% of cases the curve will be outside of the 95% CI limits even if the underlying population dynamics was identical to the null model, false positive results are identified through a global significance statistic. This is calculated by first transforming both empirical and simulated probability density values into Z scores in relation to the simulated distribution for each time unit. Z scores outside the 95% CI are then summed both for the empirical and simulated curves. The empirical sum of Z scores is compared to the distribution of summed Z scores from simulated datasets. The global significance value is the relative frequency of simulated Z score sums, which are equal to or greater than the empirical value. Recent progress in this particular research lineage now allows formal comparison of entire regional radiocarbon assemblages using different datasets, for instance from different areas of Jomon culture in Japan, that also produces global significance tests, so inter-regional demographic models can be critically assessed and productively compared (29).

As it is, the Shennan et al. method (henceforth the UCL method) tests for the departure of the empirical SPD curve from the null model SPD curve by simulating SPD curves from the

null model and constructing confidence intervals for each point in time. However, this method cannot tell whether a difference in the values of the SPD between the two points on the empirical curve is significant relative to the null model when it comes to differences in the shape of the curve or parts of the curve. For example, if the true population scenario looked like the left panel in Figure 1, the UCL method would pick up the general deviation from the null model. The uniform null model is used here for simplicity, because there are no taphonomic effects given that these are simulated data. If the uniform model is applied to a set of 350 randomly simulated radiocarbon dates from the hypothetical population model, it would not be able to tell us whether the changes in the part of the curve which is already outside the confidence intervals are significant (right panel of Figure 1); the original method does not detect the small trough in the high population zone (vertical difference between A and B in Figure 1). Likewise, the method would not be able to detect the subtleties of the situation shown in Figure 2, where 350 dates are sampled from the underlying hypothetical model and summed. The SPD curve is consistent with the null model when it comes to the range of variation for each calendar year; however, we know that the shape of the true underlying model is different from the uniform model. In spite of this, we would not be able to detect a significant drop in the curve between points A and B in Figure 1 or 2 using only the original UCL method.

Insert Figure 1. If the true population scenario looked like the left panel, the UCL method would pick up the general deviation from the null model. If applied to a set of 350 randomly simulated radiocarbon dates from the hypothetical population model, it would not be able to tell us whether the changes in the part of the curve which is already outside the confidence intervals are significant (right panel).

Insert Figure 2. A set of 350 randomly simulated dates are sampled from the underlying hypothetical model and summed. The SPD curve is consistent with the null model when it comes to the range of variation for each calendar year; however, we know that the shape of the true underlying model is different from the uniform model.

A simple extension of the original UCL method is proposed to resolve this. The main idea of this refinement is that the significance of the relative changes in the SPD curve can be tested by statistically comparing the difference between two points on the empirical SPD curve to the distribution of differences between the points with the same coordinates in calibrated time on the SPD coming from the null model. The statistical test is based on drawing a large number of samples from the probability distribution of calendar dates given by the null model, back-calibrating them, re-calibrating them and summing them, and calculating the vertical difference between points A and B on the simulated SPD curve for each sample draw. This will produce the distribution of vertical differences between two fixed points in calibrated time under the null model. Then we just compare the empirical difference to the distribution of differences under the null model.

Case Study 1: A Historical Recorded Demographic Drop Tested by Simulation

The UCL method has continued to receive some criticism (10,30), so to test its efficacy we use a known historical dataset (29), containing the start, duration and end of the European Black Death, to determine if we can accurately approximate the historically recorded population crash estimated at c. 30% mortality rate of the (census) population. To do so, we simulate a random sample of 1000 radiocarbon dates according to the probabilities given by Contreras and Meadows' (10) historical population dynamics curve and then apply our refined version of the UCL model (see above) to this hypothetical set of data. The sampled radiocarbon dataset is provided with a randomised standard error of dates between 30 and

40 radiocarbon years. A stationary (uniform) population model is used as a null model against which the SPD is evaluated, as no taphonomy is involved since the dates are sampled randomly/directly from a known historical curve. The results (Figure 3) show that the empirical curve dips under the lower 95% CI limit between 1300 and 1400 AD exactly when the Black Death de-population episode occurs. The calculated global significance is <0.001 based on 10,000 simulated iterations. The vertical difference between points A and B is also significant (p = 0.0169), although in this case the original UCL method is sufficient to demonstrate the deviation as the empirical curve does go beyond and above the 95% CI limits at the expected time. We provide the results of both the UCL and extended method test applied to the same data but on multiple simulated samples in Supplementary Information. These results clearly show that even when the original UCL method cannot demonstrate the significance of a change in the curve, the extended method can. This indicates that both the original UCL method and our extension can test for short-duration demographic events in history.

Insert Figure 3. 1000 randomly sampled radiocarbon dates from the period between 1000 and 1700 AD, with the standard error of dates between 30 and 40 radiocarbon years.

Case Study 2: An Indigenous Oral Historical Record Tested by the Extended UCL Method

We next apply the extended UCL method to an archaeological context to test the hypothesis that oral records provide evidence for an occupation gap that may be recoverable in the radiocarbon dates. Marsden (32) proposed that Tsimshian oral records, called *adawx*, record a regional conflict known as the 'War with the Tlingit' that resulted in the wholesale abandonment of the coastal territories of the Tsimshian located along the northern coast of British Columbia, Canada (Figure 4) sometime between 1500 and 1000 years ago (12). As a test of our revised method, we evaluate the potential for a demographic gap around this time from radiocarbon dates derived from coastal Tsimshian archaeological sites in the Prince Rupert Harbour, a main population center of the Tsimshian (33–35). All radiocarbon dates for Prince Rupert Harbour were audited and calibrated using the latest calibration curve, otherwise they would be inaccurate, imprecise, and incomparable (36–38).
Firstly, the calibrated radiocarbon results are examined for visually obvious gaps in Prince

Rupert Harbour settlement history that may correspond to the oral historical record. A battery of models using OxCal radiocarbon calibration software (see Supplemental Information) are used to construct two groups (phases) of dates around the most obvious candidate gap following a well-established research protocol derived from two recent exceptional archaeological cases, sequenced using ideally dated and stratified radiocarbon material from Fiji and Tonga in Polynesia (39,40).

Insert Figure 4. The Prince Rupert Harbour area, showing archaeological sites with terrestrial and marine based radiocarbon samples.

Only one OxCal model gave a sufficiently good agreement index that allowed the data to be sequenced into two phases. This model provides an interval between these two groups of dates (the gap) to be calculated in calendar years; in this case c. 42-259 years happening between 1240-1060 cal BP (median 1166) and 1070-945 (median 994) cal BP, (see Supplementary Information). To avoid any confirmation bias of our own, we treat this OxCal result cautiously as a working hypothesis, and then test it with our new extended UCL method.

Radiocarbon dates are also summed (1,4,29) to see if this gap could be detected by a conservative simulation test. This summing and simulation method uses bespoke computer code written in open-sourced R statistical software (see Supplementary Information). We applied the UCL method and its extension as described above with the difference that we used the Surovell et al. (7) exponential curve equation which models the effects of taphonomy instead of fitting the exponential model to the empirical SCPD curve. We deviate from the original formulation of the UCL method where the null model is constructed by fitting the exponential curve to the empirical summed probability curve, with an aim to account both for *assumed* effects of taphonomy and a secular population growth trend. We make no assumptions about a secular population growth trend, and use the null model curve constructed independently of our data which only accounts for the assumed effects of taphonomy (7). In this case, we consider a potential secular population growth trend to be a separate demographic phenomenon to be discovered, if it is there.

Insert Fig. 5. Prince Rupert Harbour area (above plot), with an illustrative kernel density heat map showing both distribution and relative intensity of marine based radiocarbon results . Below plot: Prince Rupert Harbour marine based radiocarbon data summed with

extended UCL method with 100 year data bins. Points 'A' and 'B' in blue, show a significant drop outside the 95% confidence intervals.

The solid red line in Figure 5 shows a general trend of the real data by fitting a rolling 200-year average to the real data (the black line). The interval between ~ 2800 – 1500 cal BP remains outside of the expected confidence range, which suggests ~ 1300 years of a large yet fluctuating relative population, prior to a significant demographic drop in the region starting somewhere between 1800 and 1100 cal BP. Although we do see a c.200 year gap at c. 1000 BP (95% confidence interval delineated by the two solid grey lines) with a significant general downward trend in population, interpretative caution is required. Assuming a single Marine Reservoir Effect (MRE) value for the entire marine radiocarbon result dataset may be problematic if it insufficiently accounts for all local ΔR variation in the dataset. Although we are confident that this value is accurate for the last c.5000 years following the most recent conservative calculation of a local ΔR at the site of Kitandach (273±38) (38), interpretations of marine based radiocarbon results remain variable and potentially problematic in this region, even when using the most rigorously calculated MRE values with the latest radiocarbon methods (38). Furthermore, lack of calibration data present in the smoother Marine 13 curve (37) compared with its terrestrial counterpart obscures smaller features and smooths SPD results, despite the larger radiocarbon sample size (N=336) used in this case.

Insert Figure 6. Prince Rupert Harbour area (above plot), with an illustrative kernel density heat map showing both distribution and relative intensity of terrestrial based radiocarbon results. Below plot: Prince Rupert Harbour terrestrial data summed with extended UCL method, using 100 year radiocarbon data bins. Points 'A' and 'B', and 'C' and 'D' in blue, show a significant drop outside the 95% confidence intervals.

Figure 6 results are generated again using 100 year data bins for the available PRH terrestrial radiocarbon data. The real radiocarbon data crosses (D), or is marginally close to the 95% CI (B) of the simulated data (the solid grey line) in two places. Thus we find two candidate occupation gap horizons (B and D) indicated by this method. The global p value is highly significant ($p = 0.0095$) indicating that deviations of the empirical curve from the null model are greater than chance. The extended method shows that both gaps may be significant as the differences between points A-B and C-D are statistically significant at the 0.05 level (with Bonferonni correction the threshold would be 0.025 – significance values associated with A-B and C-D differences are below this value, 0.0049 and 0.0091, respectively). The more recent gap, ~ 1200-1000 cal BP at D, is in broad agreement with the results of our OxCal radiocarbon model detailed above, so we suggest this is the best candidate gap for correspondence with other lines of material evidence for the hiatus described in the oral record (12). Additionally, the earlier gap is more likely to be a result of sampling bias (see Supplementary Information). Our preferred gap model of ~1200-1000 cal BP is consistent with our Marine sample SPD model, as only the later gap (at D) persists in both datasets.

Discussion

There is wide consensus that demographic patterns are potentially visible in radiocarbon data if the data are representative of historical trends. In archaeological contexts with smaller numbers of dates, the UCL method provides a means of assessing demographic trends via a comparison between the actual data and iterations of modeled data. We propose a refinement to this method that allows for a test of specific population trends of short duration, on the order of 100-200 years. Our test correctly identified such trends in modeled scenarios and against the known historical effects of the Black Death bubonic plague. Our results validate the UCL method using a conservative testing approach.

We also used this method to evaluate whether an event recorded in Tsimshian oral records was visible in radiocarbon data. While in this particular case there is considerable historical and archaeological evidence for this event, our test remains a conservative approach that provides both accurate and precise results for specific population level questions. With all modelling caveats in mind, we conclude that the event as recorded in the oral record – a settlement hiatus of the coastal Tsimshian region – occurred between 1200 and 1100 years ago. This represents the first time an Indigenous oral record has been subjected to such rigorous testing. Our result, that the Tsimshian oral record is correct (properly not disproved) in its accounting of events from over 1000 years ago, is a major milestone in the evaluation of the validity of Indigenous oral traditions.

Independent testing of hypotheses derived from the oral and historical records in this way avoids both confirmation and rejection biases. In our case, we tested events as recorded in documentary and oral records, but this approach would serve to test any explicit demographic hypothesis, regardless of the source. Our extension of the UCL simulation and summing method allows formal demographic questions to be more rigorously tested whilst accounting for small sample sizes and short duration events.

Acknowledgements

We wish to thank the following people and organizations: Lax Kw'alaams Indian Band, Metlakatla Indian Band, Susan Marsden, David Archer, Bryn Letham, Iain McKechnie, Ian Hutchinson, Eric Guiry, Steven Dennis, David Leask. Gordon Cook and Phillipa Ascough (University of Glasgow) and the Scottish Universities Environmental Research Council (SUERC) staff are thanked for the 14C sample preparations and measurements. Samples for dating were collected through the Social Sciences and Humanities Research Council of Canada Grant Number 410-2011-0814 awarded to Andrew Martindale as Principal Investigator (PI). Radiocarbon measurements were obtained through (US) National Science Foundation Grant Number 216847 awarded to Kenneth Ames as PI.

Prof. Stephen Shennan, Prof. Mark Thomas and Adrian Timpson (University College London), and Dr. Enrico Crema (Division of Archaeology, Cambridge), are thanked for their support and inspiration.

References

1. Shennan S, Downey SS, Timpson A, Edinborough K, Colledge S, Kerig T, et al. Regional population collapse followed initial agriculture booms in mid-Holocene Europe. Nat Commun [Internet]. 2013 [cited 2017 Jan 21];4. Available from: http://www.nature.com/ncomms/2013/131001/ncomms3486/full/ncomms3486.html

2. Edinborough K, Crema E, Shennan S, Kerig T. An ABC of lithic arrowheads: A case study from south-eastern France. In: Brink K, Hydén S and Jennbert K, and Larsson L, and Olausson D, (eds.) Neolithic Diversities (pp. 213-223): Lund. In: Neolithic Diversities [Internet]. Department of Archaeology and Ancient History, Lund University; 2015 [cited 2017 Jan 21]. p. 213–23. Available from: http://lup.lub.lu.se/record/7791229

3. Crema ER. Modelling temporal uncertainty in archaeological analysis. J Archaeol Method Theory. 2012;19(3):440–461.

4. Timpson A, Colledge S, Crema E, Edinborough K, Kerig T, Manning K, et al. Reconstructing regional population fluctuations in the European Neolithic using radiocarbon dates: a new case-study using an improved method. J Archaeol Sci. 2014;52:549–557.

5. Lenth RV. Some Practical Guidelines for Effective Sample Size Determination. Am Stat. 2001 Aug 1;55(3):187–93.

6. Surovell TA, Toohey JL, Myers AD, LaBelle JM, Ahern JCM, Reisig B. The end of archaeological discovery. Am Antiq. 2017;1–13.

7. Surovell TA, Byrd Finley J, Smith GM, Brantingham PJ, Kelly R. Correcting temporal frequency distributions for taphonomic bias. J Archaeol Sci. 2009 Aug;36(8):1715–24.

8. Nickerson RS. Confirmation bias: A ubiquitous phenomenon in many guises. Rev Gen Psychol. 1998;2(2):175–220.

9. Nakhaeizadeh S, Dror IE, Morgan RM. Cognitive bias in forensic anthropology: Visual assessment of skeletal remains is susceptible to confirmation bias. Sci Justice. 2014 May;54(3):208–14.

10. Contreras DA, Meadows J. Summed radiocarbon calibrations as a population proxy: a critical evaluation using a realistic simulation approach. J Archaeol Sci. 2014 Dec;52:591–608.

11. Durand JD. Historical Estimates of World Population: An Evaluation. Popul Dev Rev. 1977;3(3):253–96.

12. Martindale AR, Marsden S. Defining the Middle Period (3500 bp to 1500 bp) in Tsimshian History through a Comparison of Archaeological and Oral Records. BC Stud Br Columbian Q. 2003;(138/9):13–50.

13. Henige D. Impossible to disprove yet impossible to believe: the unforgiving epistemology of deep-time oral tradition. Hist Afr. 2009;36:127–234.

14. Mason RJ. Inconstant companions: archaeology and North American Indian oral traditions. University of Alabama Press; 2006.

15. Ames KM, Maschner HG. Peoples of the northwest coast: their archaeology and prehistory. 1999 [cited 2017 Jan 22]; Available from: https://works.bepress.com/kenneth_ames/1/

16. Ames KM. Sedentism: a temporal shift or a transitional change in hunter-gatherer mobility patterns? Bands States Cent Archaeol Investig Occas Pap No 9 [Internet]. 1991 [cited 2017 Jan 22]; Available from: http://pdxscholar.library.pdx.edu/anth_fac/62/

17. Rick JW. Dates as Data: An Examination of the Peruvian Preceramic radiocarbon record. Am Antiq. 1987;52(1):55–73.

18. Collard M, Vaesen K, Cosgrove R, Roebroeks W. The empirical case against the 'demographic turn' in Palaeolithic archaeology. Philos Trans R Soc B Biol Sci. 2016 Jul 5;371(1698):20150242.

19. Collard M, Edinborough K, Shennan S, Thomas MG. Radiocarbon evidence indicates that migrants introduced farming to Britain. J Archaeol Sci. 2010;37(4):866–870.

20. Edinborough K. Population history and the evolution of Mesolithic arrowhead technology in south Scandinavia. 2009 [cited 2017 Jan 21]; Available from: http://discovery.ucl.ac.uk/1452288/

21. Buchanan B, Collard M, Edinborough K. Paleoindian demography and the extraterrestrial impact hypothesis. Proc Natl Acad Sci. 2008;105(33):11651–11654.

22. Rick JW. Dates as Data: An examination of the Peruvian Preceramic radiocarbon record. Am Antiq. 1987;52(1):55–73.

23. Edinborough KSA. Evolution of bow-arrow technology. [Internet]. University of London; 2005 [cited 2017 Jan 21]. Available from: http://discovery.ucl.ac.uk/1444653/1/U591962.pdf

24. Edinborough K. Population history and the evolution of Mesolithic arrowhead technology in south Scandinavia. 2009 [cited 2017 Jan 21]; Available from: http://discovery.ucl.ac.uk/1452288/

25. Edinborough K. Weapons of maths instruction: A thousand years of technological stasis in arrowheads from the south Scandinavian Middle Mesolithic. Pap Inst Archaeol [Internet]. 2005 [cited 2017 Jan 21];16. Available from: http://www.pia-journal.co.uk/article/view/pia.248/

26. Shennan S, Edinborough K. Prehistoric population history: from the Late Glacial to the Late Neolithic in Central and Northern Europe. J Archaeol Sci. 2007;34(8):1339–1345.

27. McEvedy C, Jones R. Atlas of world population history. Penguin Books Ltd, Harmondsworth, Middlesex, England.; 1978.

28. Downey SS, Bocaege E, Kerig T, Edinborough K, Shennan S. The Neolithic demographic transition in Europe: correlation with juvenility index supports interpretation of the Summed Calibrated Radiocarbon Date Probability Distribution (SCDPD) as a valid demographic proxy. PloS One. 2014;9(8):e105730.

29. Crema ER, Habu J, Kobayashi K, Madella M. Summed Probability Distribution of 14 C Dates Suggests Regional Divergences in the Population Dynamics of the Jomon Period in Eastern Japan. PloS One. 2016;11(4):e0154809.

30. Torfing T. Neolithic population and summed probability distribution of 14C-dates. J Archaeol Sci. 2015 Nov;63:193–8.

31. Durand JD. Historical Estimates of World Population: An Evaluation. Popul Dev Rev. 1977;3(3):253–96.

32. Marsden S. Defending the mouth of the Skeena: perspectives on Tsimshian Tlingit relations. Perspect North Northwest Coast Prehistory. 2001;(160):61–106.

33. Ames KM, Martindale A. Rope bridges and cables: a synthesis of Prince Rupert Harbour archaeology. Can J Archaeol. 2014;38(1):140–78.

34. Letham B, Martindale A, McLaren D, Brown T, Ames KM, Archer DJ, et al. Holocene settlement history of the Dundas islands archipelago, Northern British Columbia. BC Stud. 2015;(187):51.

35. Martindale A, Letham B, Supernant K, Brown TJ, Edinborough K, Duelks J, et al. Monumentality and Urbanism in Northern Tsimshian Archaeology. In Hunters and Gatherers; 2017.

36. Stuiver M, Reimer PJ, Braziunas TF. High-precision radiocarbon age calibration for terrestrial and marine samples. Radiocarbon. 1998;40(03):1127–1151.

37. Reimer PJ, Bard E, Bayliss A, Beck JW, Blackwell PG, Bronk Ramsey C, et al. IntCal13 and Marine13 radiocarbon age calibration curves 0-50,000 years cal BP. 2013 [cited 2017 Jan 23]; Available from: https://waikato.researchgateway.ac.nz/handle/10289/8955

38. Edinborough K, Martindale A, Cook GT, Supernant K, Ames KM. A Marine Reservoir Effect ΔR Value for Kitandach, in Prince Rupert Harbour, British Columbia, Canada. Radiocarbon. 2016 Dec;58(04):885–91.

39. Burley DV, Edinborough K. Discontinuity in the Fijian archaeological record supported by a Bayesian radiocarbon model. Radiocarbon. 2014;56(1):295–303.

40. Burley D, Edinborough K, Weisler M, Zhao J. Bayesian modeling and chronological precision for Polynesian settlement of Tonga. PloS One. 2015;10(3):e0120795.

41. Ramsey CB. Radiocarbon calibration and analysis of stratigraphy: the OxCal program. Radiocarbon. 1995;37(02):425–430.

42. Riede F, Edinborough K. Bayesian radiocarbon models for the cultural transition during the Allerød in southern Scandinavia. J Archaeol Sci. 2012;39(3):744–756.

Supplementary Information

Case study 1 results.

500 samples

For the sample size of 500, 7 out of 10 times we detect the Black Death signal either by the Shennan-Timpson method alone (when the empirical curve hits or crosses the 95% CI) or by the extended method (testing for the difference between A and B, when marked on graph). Cases in which neither method produces significant results are marked by NS in the graph. In cases where the SCPD curve clearly goes under the lower 95% CI limit the extended method was not applied.

Insert Fig 7. 500 samples.

1000 samples

For the sample size of 1000, 9 out of 10 times we detect the Black Death signal either by the Shennan-Timpson method alone (when the empirical curve hits or crosses the 95% CI) or by the extended method (testing for the difference between A and B). Cases in which neither method produces significant results are marked by NS in the graph. In cases where the SCPD curve clearly goes under the lower 95% CI limit the extended method was not applied.

Insert Figure 8. 1000 samples.

Case study 2.

1. Oral records provide evidence for an occupation gap that may be recoverable in the radiocarbon dated archaeological record.

2. Available radiocarbon dates for Prince Rupert Harbour are audited and calibrated. Please note, radiocarbon results must always be calibrated using the latest calibration curve (37), or they will be inaccurate and imprecise.

3. The calibrated results are examined for visually obvious gaps in Prince Rupert Harbour settlement history that may correspond to the oral historical record.

4. OxCal radiocarbon calibration software (41) is used to construct two archaeological groups (phases) of terrestrial radiocarbon for dates sequenced around a putative occupation gap (42). This allows an interval (or putative gap) between these two groups of dates to be calculated in calendar years. A standard two-phase sequence OxCal model is used with a uniform prior distribution, and the interval command (see Supplementary Information SI_CS2_OxCal_20_7_2017.xls file for OxCal Code, and uncalibrated radiocarbon data with calibrated result output). This model assumes that two defined archaeological 'phases', are ordered sequentially, meaning that phase-1 is older than phase-2. Dates within each of these phases are not assumed to have any order. An OxCal model "Agreement Index" lower than 60 is considered insufficient, following OxCal protocol (41), and is rejected here. Agreement indices for individual dates in this model typically varied between 90-105, with a single date at the extreme old age of the pre-hiatus phase was highlighted as inconsistent, but had no effect on the estimate of the hiatus. Our model's overall Agreement Index was 101.6, indicating an internally consistent model. Using the Boundary End of phase-1 as the estimate for the beginning of the hiatus and the Boundary Start Phase-2 as the estimate for the ending of the hiatus, our model indicates that phase-1

ended sometime between 1240-1060 (median 1166) and phase-2 began sometime between 1070-945 (median 994). The interval command, which finds the highest likelihood estimate for the space between the "Boundary End 1" and "Boundary start 2 commands", returned a result of 42-259 years, estimates with 95% confidence that a gap of occupation between 42-259 years separates the ending of phase-1 and the beginning of phase-2 (See Figure 9).

Insert Figure 9. OxCal result.

R-Code – UCL method replicated

5. Radiocarbon dates are then summed using a published and widely cited method (1,4,29) to see if this gap can be detected by a more conservative simulation test. This summing and simulation method uses bespoke computer code written in open-sourced R statistical software and works as follows;

Load the following R Packages;

```
library(caTools)
library(Bchron) #version 3.3.0
data(intcal13) # if marine calibration curve is used intcal13 needs to be replaced with marine13 in all lines of the code

empdata <- read.table("clipboard") # Imports copied data from a clipboard (copy spreadsheet data in attached file SI_CS2_OxCal_20_7_2017. xls in three grey columns - radiocarbon dates, std. errors, site/phase codes); site phase codes need to be integers; before copying spreadsheet data need to be sorted first by site phase code in ascending order and then by site phase code in descending order

Mastergrid <- c(0:13000) #sets the global calibration range (in years BP where present is year 1950 AD)
Masterdensitybin <- c(rep(0, length(Mastergrid)))
```

5a. BINNING RADIOCARBON DATES at defined intervals (e.g., 100 year bins) sets the threshold for the separation of bins within a site defined by the archaeologist. Binning the data in this way accounts for “over-sampling”, where one particular site may have a great deal of radiocarbon dates, as opposed to another site with far less radiocarbon dates.

```
bindif = 100        #sets the threshold between radiocarbon dates for the separation of bins within site phases
MASTERBIN <- c()
r = 500000
for(q in 1:length(levels(factor(empdata[,3])))) {
```

```
a <- empdata[which(empdata[,3]==q),1]

BIN <- c(1:length(a))

if(length(a)==1){BIN=r; r = r+1} else {for(i in 1:(length(a)-1)) {
                                        if((a[i]-a[i+1]) < bindif) {BIN[i+1] <- BIN[i];} else
{BIN[i+1]=BIN[i]+1;}}
}

MASTERBIN <- c(MASTERBIN, BIN*(-1)^q)

}

BIN <- c(1:length(MASTERBIN))
h = 1
for(i in 2:length(MASTERBIN)) {
if(MASTERBIN[i]==MASTERBIN[i-1]) {BIN[i] = BIN[i-1]} else {h = h +1; BIN[i] = h;}
}
```

5b. SUMMING WITHIN BINS combines all dates in each defined interval (e.g., each 100 year bin) automatically to one uncalibrated date per bin.

```
BINfreq <- as.vector(summary(factor(BIN)))
BINfreq
numofbins <- length(levels(factor(BIN)))
BINNEDDATE <- c()
for(w in 1:numofbins){
binindex <- which(BIN==w)
        for(s in 1:length(binindex)){
                agesbin = BchronCalibrate(ages=empdata[binindex[s],1],
ageSds=empdata[binindex[s],2], calCurves='intcal13')
                yearsbin <- agesbin$date1$ageGrid
```

```
                densitybin <- agesbin$date1$densities
                for (m in 1:length(yearsbin)) {
                indexbin <- which(Mastergrid==yearsbin[m])
                Masterdensitybin[indexbin] <- Masterdensitybin[indexbin] + densitybin[m]
                };
                }
if(sum(Masterdensitybin) > 0) {BINNEDDATE <-
cbind(BINNEDDATE,(Masterdensitybin/sum(Masterdensitybin)))} else
{BINNEDDATE=BINNEDDATE}
Masterdensitybin <- c(rep(0, length(Mastergrid)))
}
```

5c. SUMMING BETWEEN BINS creates one calibrated date per bin, accounting for the fluctuations in the latest radiocarbon calibration curve. All the calibrated data is then summed over one creating a calibrated Summed Probability Distribution Frequency (SPDF), of the archaeological radiocarbon data. This result is the solid black line on figures 5 and 6. A focal range needs to be defined (i.e. the time range of the data) so that the null model is constructed only over this specifc time interval.

```
SUMOFSUMS <- apply(BINNEDDATE, 1, sum)
SUMOFSUMS <- SUMOFSUMS/sum(SUMOFSUMS)   #Empirical SPD

Empirical_density <- SUMOFSUMS

SUMrunmean <- runmean(SUMOFSUMS, 200)

#SETTING BOUNDARIES OF THE RESTRICTED (FOCAL) RANGE

start=-9200     #start point of focal interval in years BC/AD (if BC put minus sign in front)
end=1850                #end point of focal interval in years BC/AD (if BC put minus sign in
front)
```

```
sgrid <- which((Mastergrid-1950)==-start)
egrid <- which((Mastergrid-1950)==-end)

s = sgrid
e = egrid
```

5d. CREATING A NULL MODEL This exponential SPDF model accounts for an expected exponential loss of archaeological data over time by site formation processes (Surovell et al., 2009).

```
Weights=5.726442*(10^6)*((Mastergrid[s:e]+2176.4)^(-1.3925309))
Weights=Weights/sum(Weights)
```

5e. SIMULATING DATES FROM A NULL MODEL then creates a huge number of simulated calibrated and summed datasets based on the real SPDF dataset (see 5c above), using a Monte Carlo based method. This allows us to see if the "real SPDF dataset" crosses the simulated dataset at a 95% Confidence Interval. If it does, we are provided with good evidence for a positive (line through the upper CI) or negative (line through the lower CI) population signal.

```
zscoretrans <- function (x) {

z = (x - mean(x))/sd(x)
return(z)
}

mround <- function(x,base){
      base*round(x/base)
}
```

```
empsderr <- as.numeric(empdata[,2])

iter = 10000      #sets the number of simulations from the null model
ndates = numofbins
Masterdensity <- c(rep(0, length(Mastergrid)))
MASTERSPD <- matrix(0, length(Mastergrid), iter)
SIM <- c(1:ndates)
SIMcal <- c(1:ndates)


for(j in 1:iter) {
cat(j,'\n')
Null <- sample(Mastergrid[s:e], ndates, replace = TRUE, prob = Weights)

for(es in 1:length(Null)) {
if(Null[es] < 13900) {SIMcal[es] <- mround(Null[es],5)} else {if(Null[es]<25000) {SIMcal[es] <-
round(Null[es],-1)} else{if(((round(Null[es],-1)-signif(Null[es],3))/10)%%2==0) {SIMcal[es] <-
round(Null[es],-1)} else {SIMcal[es] <- round(Null[es],-1)-10}}}
}


SIMSD <- sample(empsderr, ndates, replace=TRUE)


for(u in 1:ndates) {                                    #Reverse calibration
ind <- which(intcal13[,1]==SIMcal[u])
SIM[u] <- rnorm(1, intcal13[ind,2], intcal13[ind,3])
if(SIM[u] < 0){SIM[u] <- abs(SIM[u])} else {we=1}
SIM[u] <- rnorm(1, SIM[u], SIMSD[u])
if(SIM[u] < 0){SIM[u] <- abs(SIM[u])} else{wer=2}

}
```

```
for(i in 1:ndates) {

ages1 = BchronCalibrate(ages=SIM[i], ageSds=SIMSD[i], calCurves='intcal13')
years <- ages1$date1$ageGrid
density <- ages1$date1$densities


for (m in 1:length(years)) {
index <- which(Mastergrid==years[m])
Masterdensity[index] <- Masterdensity[index] + density[m]
};

}

Masterdensity <- Masterdensity/sum(Masterdensity)
MASTERSPD[,j] <- Masterdensity                #Matrix containing simulated SPDs in columns
(iterations)
Masterdensity <- c(rep(0, length(Mastergrid)))

}
```

5f. GLOBAL *P*-VALUE TEST determines the area of the z score Confidence Intervals (CI's) for the simulated summed probabiliy distributions (SPDFs), and returns a global probability (*P*-value) to determine if the model is merely correct by chance, or instead provides us with a globally statistically significant result.

```
for(x in 1:iter) {
zeroindex <- which(MASTERSPD[,x]== 0)
MASTERSPD[zeroindex,x] = runif(length(zeroindex), 0.0000000000,0.00000000001)

}
```

```
MASTERZSCORES <- t(apply(MASTERSPD, 1, zscoretrans))            #Z scores of simulated
SPDs
CI <- t(apply(MASTERZSCORES, 1, quantile, probs = c(.025, .975)))  #95% confidence interval
limits
CIraw <- t(apply(MASTERSPD, 1, quantile, probs = c(.025, .975)))

statisticupper <- c(rep(0,iter))
statisticlower <- c(rep(0,iter))
globalstatistic <- c(rep(0,iter))    #sums of Z scores outside 95% CI intervals for simulated
SPDs

for(j in 1:iter) {
        difflower <-  MASTERZSCORES[s:e,j] - CI[s:e,1]
        diffupper <-    MASTERZSCORES[s:e,j] - CI[s:e,2]
        indexlower <- which(difflower < 0 )
        indexupper <- which(diffupper > 0)
        statisticlower[j] <- sum(difflower[indexlower])
        statisticupper[j] <- sum(diffupper[indexupper])
        globalstatistic[j] <- abs(statisticlower[j])+statisticupper[j]
}

Zscore_empirical <- (Empirical_density[s:e] - apply(MASTERSPD[s:e,], 1,
mean))/apply(MASTERSPD[s:e,], 1, sd)

########################################empirical statistic#########
difflower <-  Zscore_empirical - CI[s:e,1]
        diffupper <-    Zscore_empirical - CI[s:e,2]
        indexlower <- which(difflower < 0 )
        indexupper <- which(diffupper > 0)
        statisticlower <- sum(difflower[indexlower])
        statisticupper <- sum(diffupper[indexupper])
        empirical_statistic <- abs(statisticlower)+statisticupper
```

```
perc.rank <- function(x, xo)  length(x[x <= xo])/length(x)
prank <- perc.rank(globalstatistic, empirical_statistic)
pvalue <- 1-prank  #global statistic p value
pvalue
```

5g. MAKING A GRAPH WITH THE EMPIRICAL CURVE, ROLLING MEAN AND 95% CONFIDENCE INTERVALS

```
Mastergrid <- Mastergrid – 1950 # if  timescale in BC/AD format is preferred, run this line, if not, skip it.

plot(Mastergrid[s:e], Empirical_density[s:e], type="l", col="black",
xlim=rev(range(Mastergrid[s:e])), xlab="calBP", ylab="SPD")
lines(Mastergrid[s:e], SUMrunmean[s:e], type="l", col="red", lwd =2,
xlim=rev(range(Mastergrid[s:e])))
lines(Mastergrid[s:e], Weights, type="l", lty = 2, col ="gray",
xlim=rev(range(Mastergrid[s:e])))
lines(Mastergrid[s:e], CIraw[s:e,1],type="l", lty = 1, col ="grey",
xlim=rev(range(Mastergrid[s:e])))
lines(Mastergrid[s:e], CIraw[s:e,2],type="l", lty = 1, col ="grey",
xlim=rev(range(Mastergrid[s:e])))
```

6. Extension of UCL Method

Short description

The significance of the relative changes in the SPD curve can be tested by statistically comparing the difference between two points on the empirical SPD curve to the distribution of differences between the points with the same coordinates in calibrated time on the SPD coming from the null model. The statistical test is based on drawing a large number of samples from the probability distribution of calendar dates given by the null model, back-calibrating them, re-calibrating them and summing them, and calculating the vertical difference between points A and B on the simulated SPD curve for each sample draw. This will produce the distribution of vertical differences between two fixed points in calibrated time under the null model. Then we just compare the empirical difference to the distribution of differences under the null model.

identify(Mastergrid, Empirical_density) #read the x coordinate of the point of interest by clicking on it on the graph

index1 = # insert x coordinate of the higher point
index2 = # insert x coordinate of the lower point

teststat <- MASTERSPD[index1,] - MASTERSPD[index2,]

empstat<-Empirical_density[index1]-Empirical_density[index2]

prank <- perc.rank(teststat, empstat)
pvalue <- 1-prank #global statistic p value
pvalue

Figures.

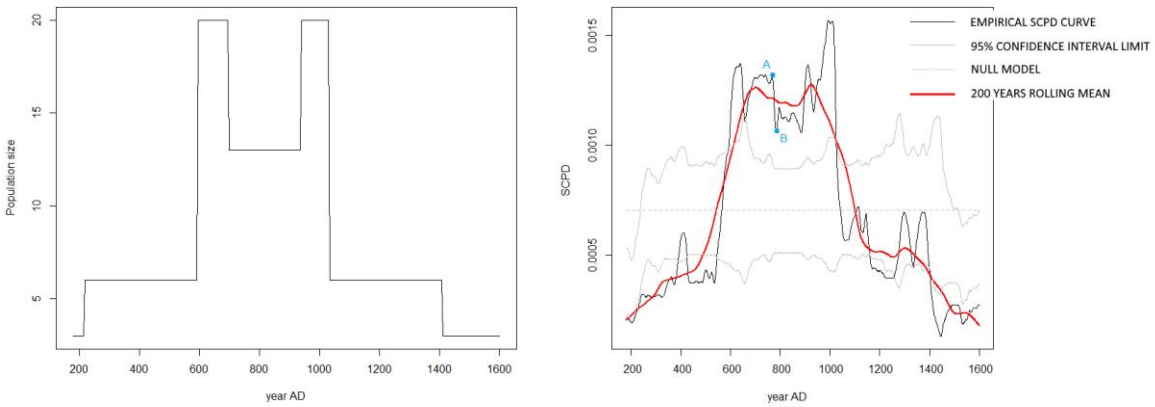


Figure 1. If the true population scenario looked like the left panel, the UCL method would pick up the general deviation from the null model. If applied to a set of 350 randomly simulated radiocarbon dates from the hypothetical population model, it would not be able to tell us whether the changes in the part of the curve which is already outside the confidence intervals are significant (right panel).

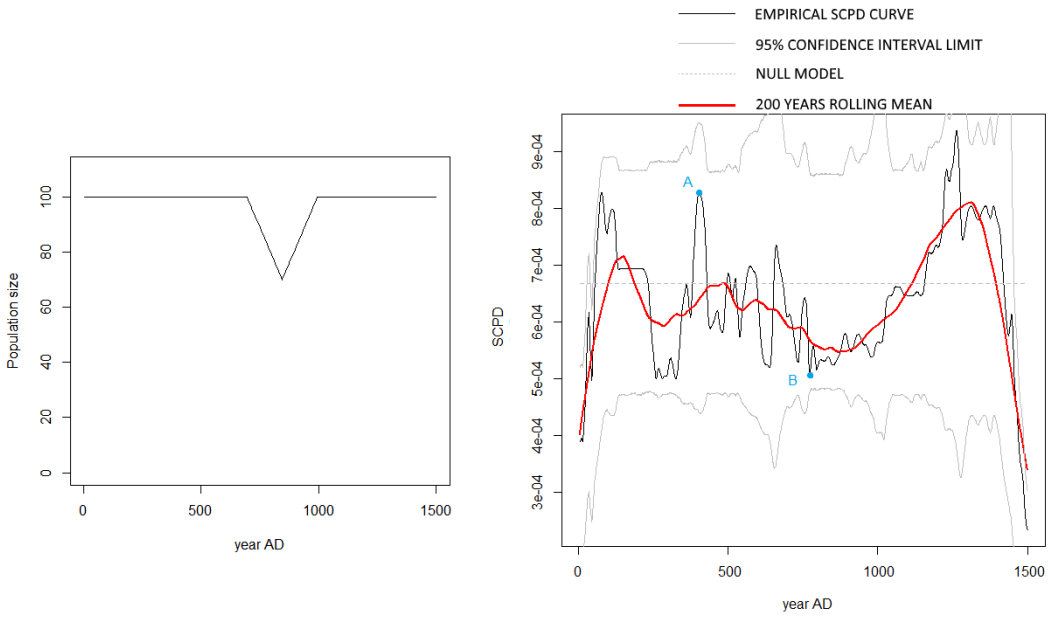


Figure 2. A set of 350 randomly simulated dates are sampled from the underlying hypothetical model and summed. The SPD curve is consistent with the null model when it comes to the range of variation for each calendar year; however, we know that the shape of the true underlying model is different from the uniform model.

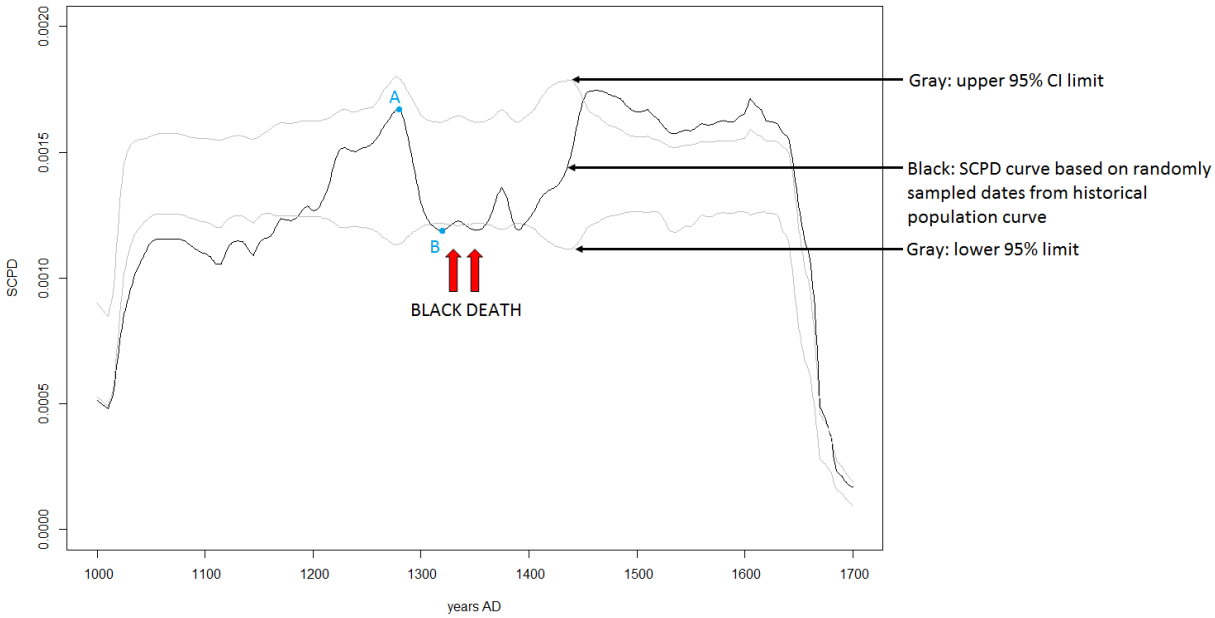


Figure 3. 1000 randomly sampled radiocarbon dates from the period between 1000 and 1700 AD, with the standard error of dates between 30 and 40 radiocarbon years.

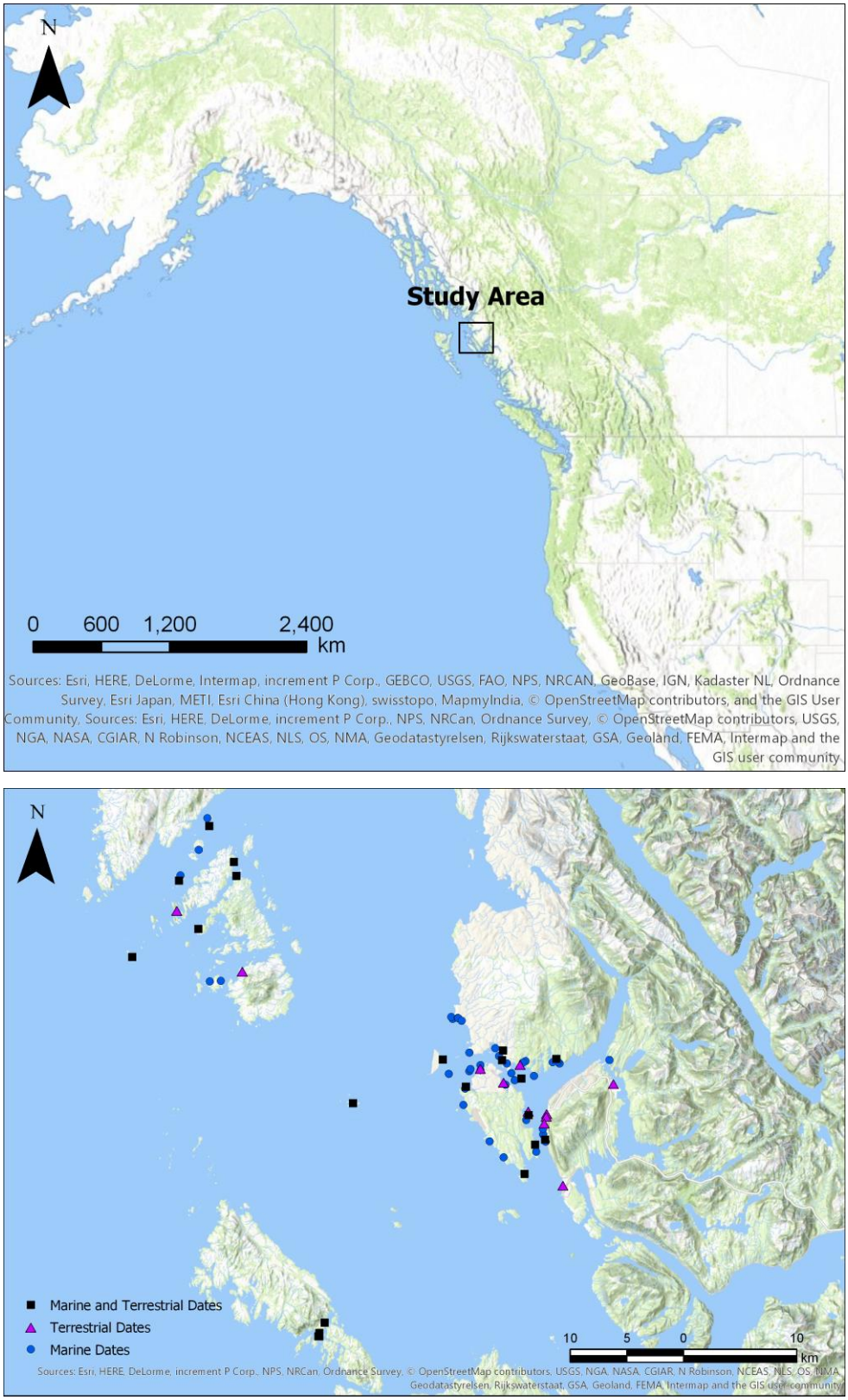
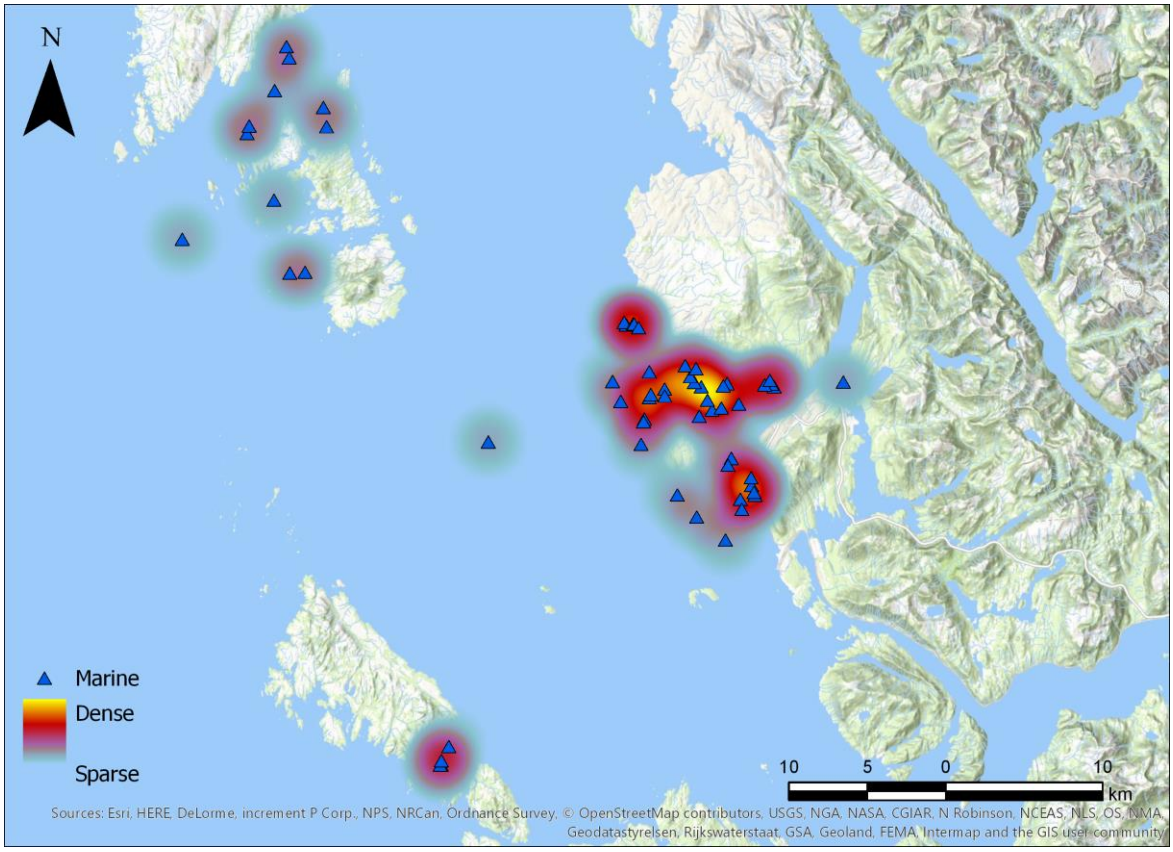


Figure 4. The Prince Rupert Harbour area (upper panel), showing archaeological sites with terrestrial and marine based radiocarbon samples (lower panel).



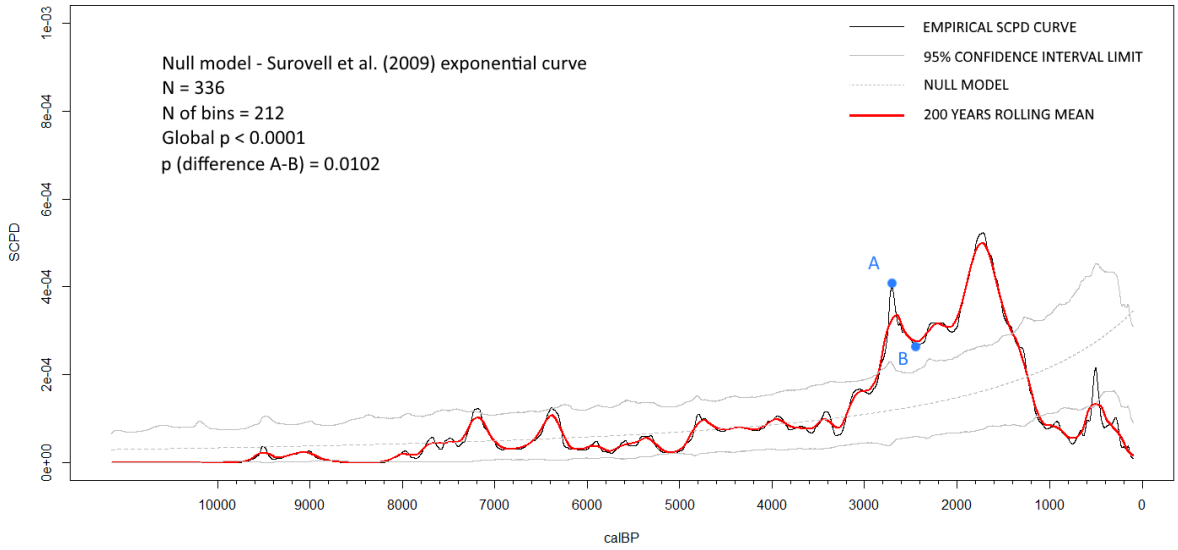
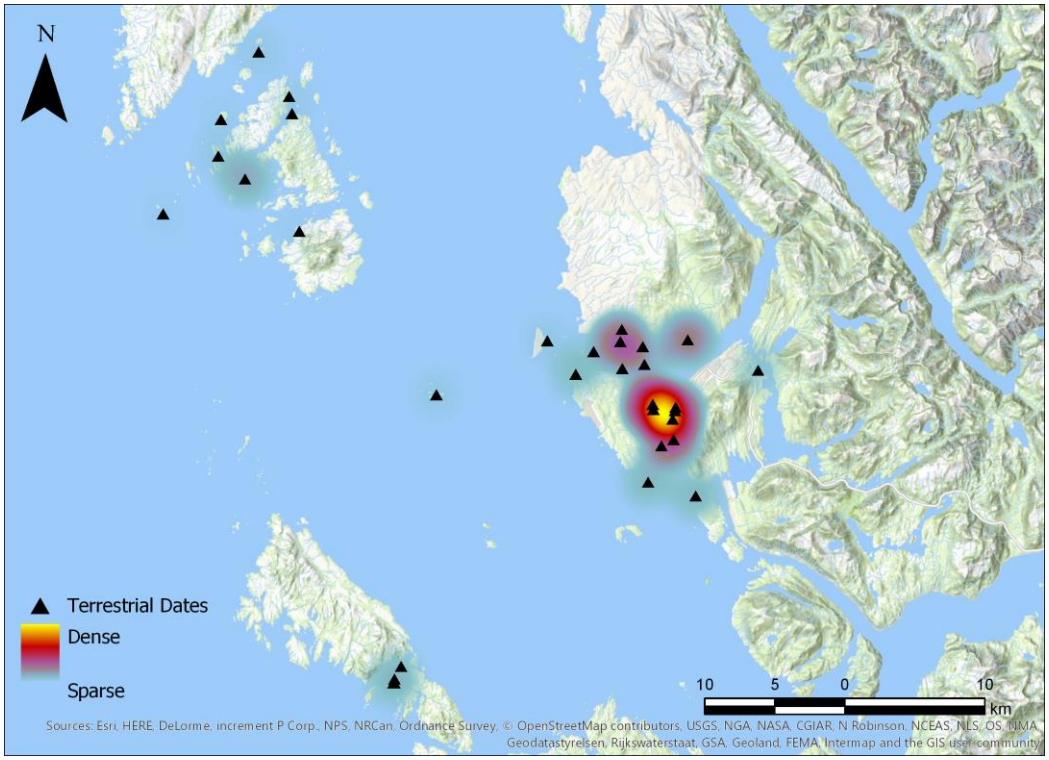


Fig. 5. Prince Rupert Harbour area (above plot), with an illustrative kernel density heat map showing both distribution and relative intensity of marine based radiocarbon results . Below plot: Prince Rupert Harbour marine based radiocarbon data summed with extended UCL method with 100 year data bins. Points 'A' and 'B' in blue, show a significant drop outside the 95% confidence intervals.


N
Terrestrial Dates
Dense
Sparse
10 5 0 10
km
Sources: Esri, HERE, DeLorme, increment P Corp., NPS, NRCan, Ordnance Survey, © OpenStreetMap contributors, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community

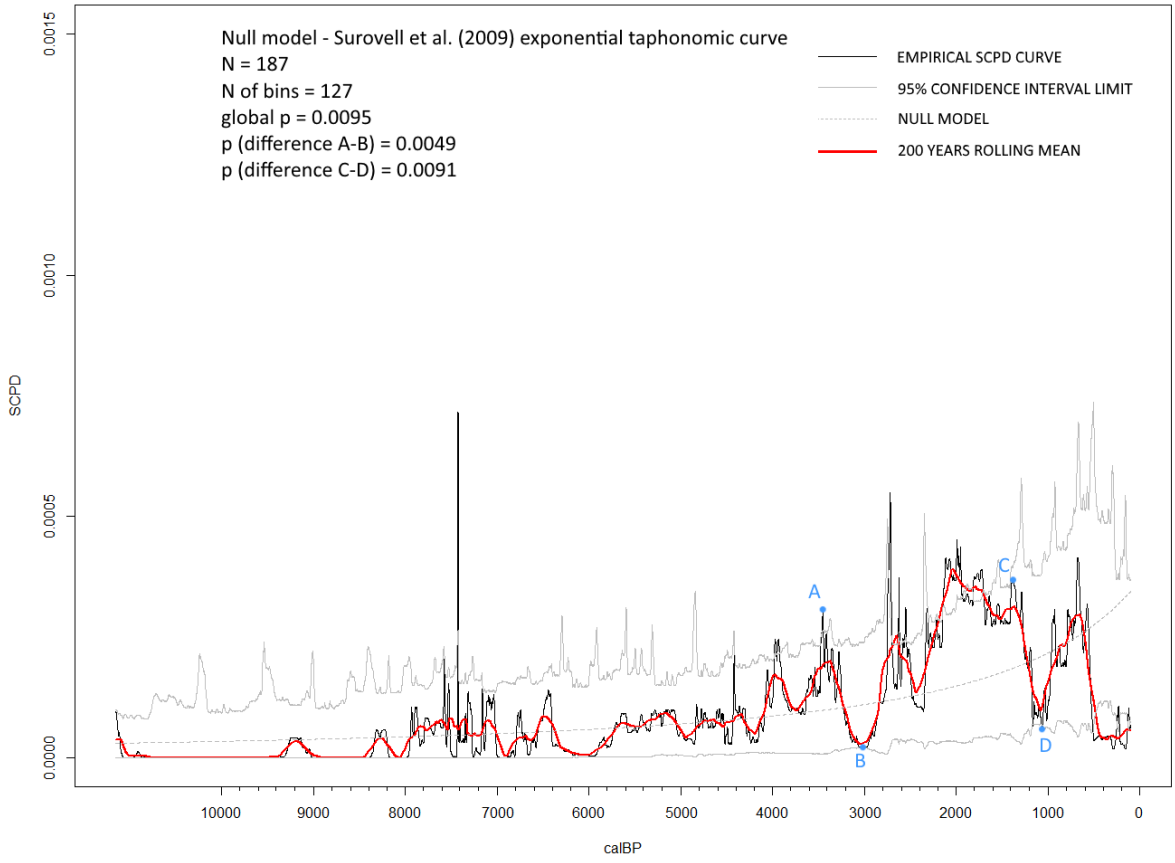

Null model - Surovell et al. (2009) exponential taphonomic curve
N = 187
N of bins = 127
global p = 0.0095
p (difference A-B) = 0.0049
p (difference C-D) = 0.0091
EMPIRICAL SCPD CURVE
95% CONFIDENCE INTERVAL LIMIT
NULL MODEL
200 YEARS ROLLING MEAN
SCPD
0.0000
0.0005
0.0010
0.0015
A
B
C
D
10000 9000 8000 7000 6000 5000 4000 3000 2000 1000 0
calBP

Figure 6. Prince Rupert Harbour area (above plot), with an illustrative kernel density heat map showing both distribution and relative intensity of terrestrial based radiocarbon results. Below plot: Prince Rupert Harbour terrestrial data summed with extended UCL method, using 100 year radiocarbon data bins. Points 'A' and 'B', and 'C' and 'D' in blue, show a significant drop outside the 95% confidence intervals.

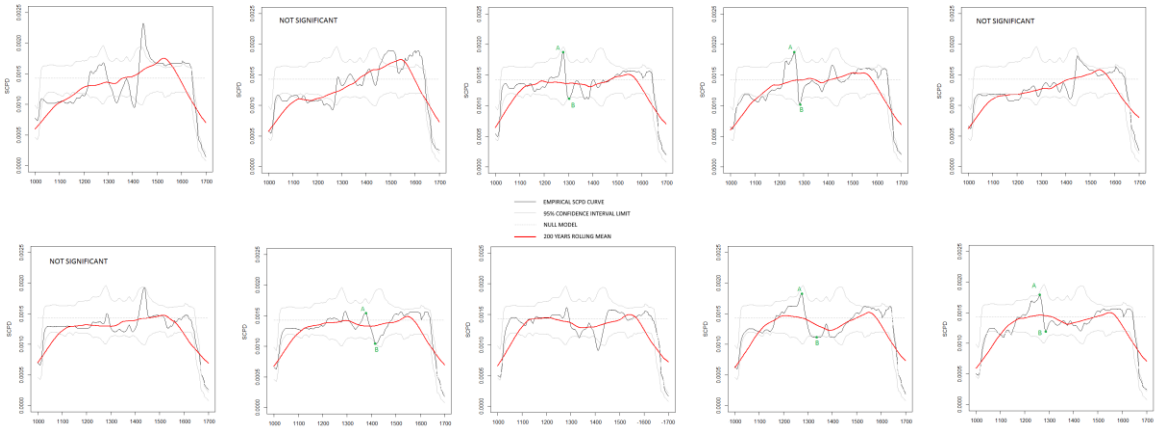


Fig 7. 500 samples.

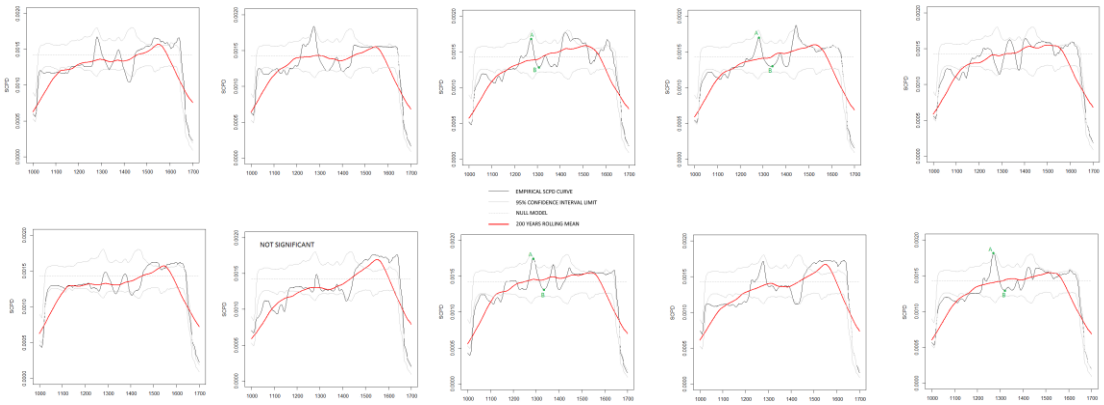


Figure 8. 1000 samples.

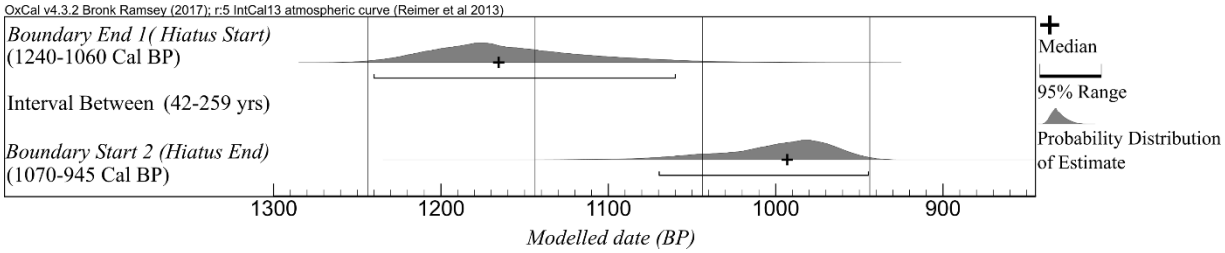


Figure 9. OxCal result.