Regional variations in the European Neolithic dispersal: the role of the coastlines

Daniel A. Henderson[1], Andrew W. Baggaley[2], Anvar Shukurov[1], Richard J. Boys[1], Graeme R. Sarson[1] & Andrew Golightly[1]

[1]*School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne NE1 7RU, UK (Email: daniel.henderson@ncl.ac.uk; anvar.shukurov@ncl.ac.uk; richard.boys@ncl.ac.uk; g.r.sarson@ncl.ac.uk; andrew.golightly@ncl.ac.uk)*

[2]*School of Mathematics & Statistics, University of Glasgow, Glasgow G12 8QW, UK (Email: andrew.baggaley@glasgow.ac.uk)*

The mechanisms by which agriculture spread across Europe in the Neolithic, and the speed at which it happened, have long been debated. Attempts to quantify the process by constructing spatio-temporal models have given a diversity of results. In this paper, a new approach to the problem of modelling is advanced. Data from over 300 Neolithic sites from Asia Minor and Europe were used to produce a global picture of the emergence of farming across Europe which also allows for variable local conditions. Particular attention is paid to coastal enhancement: the more rapid advance of the Neolithic along coasts and rivers, as compared with inland or terrestrial domains. The key outcome of this model is hence to confirm the importance of waterways and coastal mobilities in the spread of farming in the early Neolithic, and to establish the extent to which this importance varied regionally.

This paper is published in full in *Antiquity* 88 no. 342 December 2014. Here we publish supplementary material.

Keywords: Europe, Asia Minor, Neolithic dispersal, modelling, waterways, propagating population front, wave of advance

---

Henderson, D.A., A.W. Baggaley, A. Shukurov, R.J. Boys, G.R. Sarson & A. Golightly. 2014. Regional variations in the European Neolithic dispersal: the role of the coastlines. *Antiquity* 88: 1291–1302. http://antiquity.ac.uk/ant/088/ant0881291.htm

Radiocarbon data

As in our previous work (Baggaley *et al.* 2012b), we here use a compilation of first-arrival dates from 302 sites in Southern and Western Europe. This is derived from the calibrated dates of Steele & Shennan (2000), Gkiasta *et al.* (2003) and Thissen *et al.* (2006); but whereas the original data contain multiple dates per site, we determine a single representative first-arrival date for each site, t_i, using the method of Davison *et al.* (2007). Briefly, for sites with at least eight date measurements, a χ^2 statistical test is used to determine the most likely first arrival date from a coeval sub-sample; for sites with fewer measurements, we use a weighted mean of these measurements. Our data is supplied as part of the supplementary material as an Excel spreadsheet.

The same compilation of original dates was also used in Baggaley *et al.* (2012a), where an alternative method of determining a representative date for each site was used. We believe that our results are robust with respect to the two slightly different procedures. Baggaley *et al.* (2012a) also discussed and implemented more sophisticated models of data errors, including an allowance for local spatial inconsistencies. Those alternative procedures did not alter our basic conclusions.

Front propagation

Our wavefront model is described in detail in Baggaley *et al.* (2012b); here we briefly summarise some details of this method, to make the current work more self-contained.

Our wavefront model tracks the time-evolving position of the propagating front. The front is defined via an indexed set of suitably separated points; we track the geographical coordinates of these points as the front expands and evolves. We start with the intial front defined using points on a small circle centred on our source of spread in the Levant.

At each time step, we move point *i*, at position $\boldsymbol{x}_i = (\phi_i, \lambda_i)$ — where ϕ_i and λ_i are respectively the polar and azimuthal coordinates—a small amount in both the ϕ and λ directions using the local velocity $\mathbf{u}_i$, via

$$\frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} = \mathbf{u}_i, \qquad \mathbf{u}_i = \mathbf{U}_i + \mathbf{V}_{\mathrm{R},i} + \mathbf{V}_{\mathrm{C},i}. \tag{1}$$

Henderson, D.A., A.W. Baggaley, A. Shukurov, R.J. Boys, G.R. Sarson & A. Golightly. 2014. Regional variations in the European Neolithic dispersal: the role of the coastlines. *Antiquity* 88: 1291–1302. http://antiquity.ac.uk/ant/088/ant0881291.htm

The global propagation velocity $\mathbf{U}_i$ is directed along the outward normal to the front, as calculated from the coordinates of the neighbouring points on the front. The advective velocities $\mathbf{V}_{\mathrm{R},i}$ and $\mathbf{V}_{\mathrm{C},i}$ are directed along the relevant river or coastline, in the sense closest to the outward normal.

The factor $F(a,\phi)$, which controls the environmental dependence in the global propagation velocity $\mathbf{U}$, as introduced in the main text, is given by

$$F(a,\phi)=\left(\frac{5}{4}-\frac{\phi}{100^\circ}\right)\begin{cases}\dfrac{1}{2}-\dfrac{1}{2}\tanh\{10(a/1\mathrm{km}-1)\}, & a>0,\\ \exp(-d_{\mathrm{c}}/10\mathrm{km}), & a<0.\end{cases} \qquad (2)$$

Here d_{c} gives the shortest distance to land for points in the sea; this is used to allow limited sea travel. This environmental dependence is based on geographical altitude data from the ETOPO1 1-minute Global Relief database (Geophysical Data System 2011), using a spatial resolution of 4 arc-minutes with approximate longitudinal boundaries of 15° W and 60° E and latitudinal boundaries of 25° N and 75° N.

The river and coastline vectors used in this study are taken from World DataBank II (2004). At each point on our mesh, the vectors $\mathbf{V}_{\mathrm{R}}$ and $\mathbf{V}_{\mathrm{C}}$ are calculated using contributions from each of the irregularly spaced vector data segments defining the waterway, weighted by $\exp(-d_{\mathrm{vec}}/15\mathrm{km})$, where d_{vec} is the distance between the grid point and the river/coastal vector (in km). To apply these environmental dependences to points on our front (which normally fall between points on the grid), we use bilinear interpolation from the values at the four closest mesh-points.

After every time-step, the separation of points on the wavefront is monitored, to maintain a roughly constant resolution. If two neighbouring points are more than δ apart, a new point is inserted between them. If two points which are not neighbours are less than δ apart, then the encroaching points are removed and the ordering of the loops is switched, to merge the fronts. We take δ to be our environmental grid spacing of 4 arc-minutes.

Our wavefront model can be compared with partial differential equation (PDE) models such as the Fisher–Kolmogorov–Petrovsky–Piskunov (FKPP) equation, whose solutions are most simply obtained by discretising to a grid of points in space, and time-stepping the solution forward in time

Henderson, D.A., A.W. Baggaley, A. Shukurov, R.J. Boys, G.R. Sarson & A. Golightly. 2014. Regional variations in the European Neolithic dispersal: the role of the coastlines. *Antiquity* 88: 1291–1302. http://antiquity.ac.uk/ant/088/ant0881291.htm

at each point. Here, where we are only interested in the first arrival time at any point, time-stepping at points behind and ahead of the front is unnecessary, and adds greatly to the computational expense. The stability of the wavefront model, as with a PDE model, depends upon the time-step used. For a given front velocity, point spacing δ, and length-scale of environmental variation (which controls the heterogeneities in the velocity), the time-step cannot exceed a certain level, for the front to expand in a stable, sensible manner. The time-step used is therefore in practice rather similar to that used in a PDE model. However, as expected, the wavefront model requires considerably less computation, as only the relatively small number of points on the front need to be time-stepped.

Statistical model

The statistical aspects of the model and the approach used to fit the model are essentially the same as in Baggaley *et al*. (2012b) and so here we provide only brief details to supplement the description in the main article.

The time at which the wavefront arrives at site *i* is denoted $\tau_i(\boldsymbol{\theta})$, for *i*=1,…,*n*, where *n* = 302 is the number of radiocarbon sites in our data set. The parameters controlling the mathematical model are denoted $\boldsymbol{\theta} = (U_0, V_1, V_2, V_3, V_4, V_R)$. The observed arrival time at site *i* is denoted by t_i and these times are displayed in Figure 1 in the main article.

Our statistical model assumes that these data are generated by the wavefront model subject to (spatially) independent normal errors, that is

$$t_i = \tau_i(\boldsymbol{\theta}) + \sigma\varepsilon_i, \quad i = 1,..., n, \tag{3}$$

where the ε_i are independent and identically distributed standard normal random variables and σ is the spatially homogeneous standard deviation, allowing for a mismatch between the wavefront model and the observations. The likelihood function, that is the joint probability density of the observed arrival times, regarded as a function of the parameters, is

$$\pi(\boldsymbol{t} \mid \boldsymbol{\theta}, \sigma) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \{t_i - \tau_i(\boldsymbol{\theta})\}^2 \right] \tag{4}$$

where $\boldsymbol{t} = (t_1, t_2, ..., t_n)$ denotes the collection of observed arrival times. Note that the data enter the likelihood function through $\sum_{i=1}^{n} \{t_i - \tau_i(\boldsymbol{\theta})\}^2$, the sum of squared errors (SSE), and the parameter σ effectively scales this SSE.

We adopt a Bayesian approach (Bernardo & Smith 1994) to the problem of inferring the model parameters in which prior beliefs about the likely parameter values are expressed through the joint prior distribution, denoted $\pi(\boldsymbol{\theta}, \sigma) = \pi(\boldsymbol{\theta})\pi(\sigma)$. Our specific choices for the component prior distributions are detailed in Table 1 in the main article. The posterior distribution for the parameters given the observed arrival times is then given by Bayes' theorem as

$$\pi(\boldsymbol{\theta}, \sigma \mid \boldsymbol{t}) \propto \pi(\boldsymbol{\theta}, \sigma)\pi(\boldsymbol{t} \mid \boldsymbol{\theta}, \sigma). \quad (5)$$

This represents our beliefs about the parameter values after observing the data.

Due to the complex dependence of the likelihood on the model parameters $\boldsymbol{\theta}$, the posterior density $\pi(\boldsymbol{\theta}, \sigma \mid \boldsymbol{t})$ does not admit a simple form. Values of the parameters $\{\boldsymbol{\theta}, \sigma\}$ are therefore sampled from the posterior distribution via fairly standard Markov chain Monte Carlo (MCMC) methods (Gammerman & Lopes 2006). In brief, a Markov chain is constructed by generating candidate parameter values from a suitable proposal distribution. A proposed value is accepted as the next value in the chain with a probability that ensures the Markov chain has invariant distribution given by $\pi(\boldsymbol{\theta}, \sigma \mid \boldsymbol{t})$. If a proposal is not accepted then the next value for that parameter is taken to be its current value. The acceptance probability requires that the target density can be evaluated up to proportionality. We refer the reader to Baggaley *et al.* (2012b) for further details on the construction of the acceptance probability and suitable choices of proposal distribution. The algorithm results in a collection of parameter values that give model output that is consistent with the observed first arrival times, rather than a single 'history' that best matches the data.

Our sampling approach as described above requires evaluation of the likelihood term at each iteration. Therefore, the wavefront model must be run at many thousands of potential parameter values. Each individual wavefront model evaluation is too time-consuming to be used in the MCMC scheme and so we replace it with a fast approximation. The output from the mathematical model at site i, $\tau_i(\boldsymbol{\theta})$, is replaced in the likelihood function (4) by an accurate but computationally

more efficient approximation $\hat{\tau}_i(\boldsymbol{\theta})$. This approximation is based on a Gaussian process (Williams & Rasmussen 2006) that is fitted to carefully chosen runs of the mathematical model, as is standard practice in the statistical literature on computer models (O'Hagan 2006). Full details are provided in Baggaley *et al.* (2012b). Replacing the output from the model in this way makes the MCMC scheme outlined above computationally practicable.

Acknowledgements

This work was supported by the Leverhulme Trust under Research Grant F/00 125/AD.

References

BAGGALEY, A.W., R.J. BOYS, A. GOLIGHTLY, G.R. SARSON & A. SHUKUROV. 2012a. Inference for population dynamics in the Neolithic period. *Annals of Applied Statistics* 6: 1352–76. http://dx.doi.org/10.1214/12-AOAS579

BAGGALEY, A.W., G.R. SARSON, A. SHUKUROV, R.J. BOYS & A.GOLIGHTLY. 2012b. Bayesian inference for a wave-front model of the neolithization of Europe. *Physical Review E* 86: 016105. http://dx.doi.org/10.1103/PhysRevE.86.016105

BERNARDO, J.M. & A.F.M. SMITH. 1994. *Bayesian theory*. Chichester: Wiley. http://dx.doi.org/10.1002/9780470316870

DAVISON, K., P.M. DOLUKHANOV, G.R. SARSON, A. SHUKUROV & G.I. ZAITSEVA. 2007. A pan-European model of the Neolithic. *Documenta Praehistorica* 34: 139–54. http://dx.doi.org/10.4312/dp.34.10

GAMMERMAN, D. & H. LOPES. 2006. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. London: Taylor & Francis.

Geophysical Data System. 2011. GEODAS-NG desktop software. Available at: http://www.ngdc.noaa.gov/mgg/geodas (accessed 11 November 2014).

GKIASTA, M., T. RUSSELL, S.J. SHENNAN & J. STEELE. 2003. Neolithic transition in Europe: the radiocarbon record revisited. *Antiquity* 77: 45–62.

O'HAGAN, A. 2006. Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering and System Safety* 91: 1290–300. http://dx.doi.org/10.1016/j.ress.2005.11.025

Henderson, D.A., A.W. Baggaley, A. Shukurov, R.J. Boys, G.R. Sarson & A. Golightly. 2014. Regional variations in the European Neolithic dispersal: the role of the coastlines. *Antiquity* 88: 1291–1302. http://antiquity.ac.uk/ant/088/ant0881291.htm

STEELE, J. & S.J. SHENNAN. 2000. *Spatial and chronological patterns in the neolithisation of Europe [data-set]*. York: Archaeology Data Service [distributor].

THISSEN, L., A. REINGRUBER & D. BISCHOFF. 2006. CANeW, ^{14}C databases and chronocharts (2005–2006) updates. Data now available within the CONTEXT database: http://context-database.uni-koeln.de/ (accessed 19 November 2014).

WILLIAMS, C.K.I. & K.E. RASMUSSEN. 2006. *Gaussian processes for machine learning*. Cambridge (MA): MIT Press.

World DataBank II. 2004. CIA World DataBank II. Available at: http://www.evl.uic.edu/pape/data/WDB/ (accessed 11 November 2014).

