Geoinformatica (2012) 16:675–690
DOI 10.1007/s10707-012-0159-6

Assessing the changing flowering date of the common lilac in North America: a random coefficient model approach

Chris Brunsdon · Lex Comber

Received: 5 October 2011 / Revised: 1 May 2012 /
Accepted: 14 June 2012 / Published online: 1 July 2012

Abstract A data set consisting of *Volunteered geographical information* (VGI) and data provided by expert researchers monitoring the first bloom dates of lilacs from 1956 to 2003 is used to investigate changes in the onset of the North American spring. It is argued that care must be taken when analysing data of this kind, with particular focus on the issues of lack of experimental design, and Simpson's paradox. Approaches used to overcome this issue make use of random coefficient modelling, and bootstrapping approaches. Once the suggested methods have been employed, a gradual advance in the onset of spring is suggested by the results of the analysis. A key lesson learned is that the appropriateness of the model calibration technique used given the process of data collection needs careful consideration.

Keywords Phenology · Random effects models · Citizen science

1 Introduction

There is a long tradition of volunteers collecting and reporting different types of information about the environment we live in [35]. Robert Marsham started to formally note the arrival of the first swallow in 1736. A recent description notes that

> This popular science really took off when he reported his records to the Royal Society in 1789 and many other country gentlemen took up the pastime. [34]

Considerably more recently, the term 'volunteered geographical information' (VGI) was coined by [15], to describe geographical information collated from a

C. Brunsdon (✉)
School of Environmental Sciences, University of Liverpool, Liverpool, UK
e-mail: Christopher.Brunsdon@liverpool.ac.uk

L. Comber
Department of Geography, University of Leicester, Leicester, UK
e-mail: ajc36@le.ac.uk

broad group of private citizens, typically without formal training, and on a voluntary basis. A related idea is that of *citizen science* (CS)—see for example [8] or [20]. Here, information is collated from a large group of citizens—and in several cases this information is also geographically referenced. The two concepts are not identical—but in general, both activities are seen as activities involving the collection of data by the public at large, rather than by officially sanctioned agencies. One characteristic of CS is that the degree of prior understanding required of volunteers in a CS-based project can vary greatly. In some cases virtually no skills required of the volunteers, whereas in others some degree of volunteer instruction—and possibly selection—is necessary, so that the input of information has some degree of formal control. One example of the latter is given by Goodchild in the above reference who cites the US Christmas Bird Count, and states that

> Participants require a fairly high level of skill, and over the years a number of protocols have been established to ensure that the resulting data have high quality.

There is an increasing amount of such data that could be, and in some cases is being, incorporated into formal scientific analyses. This includes spatially referenced and geo-located data such as the data referred to above, as well as explicitly map-based data (e.g. Openstreetmap). Also, much historical volunteered information is held by public organisations and agencies who have an obligation to make their data holdings publicly available [18].

However, in all of the above situations the data collection process differs from that of a prescribed scientific experiment—and this has to be taken into consideration when analysing the data, calibrating models or testing hypotheses. The ideal from a scientific viewpoint is perhaps the 'designed experiment' [24] seen as a desirable situation for reliable statistical modelling. In an ideal world, one has a great deal of control over data collection, and indeed it is possible to deduce strategies for data collection giving optimal calibrations of statistical models. However VGI can sometimes provide a very different situation from this, as even when training is provided for the volunteers to improve the reliability of the observations, one has little control of the spatial distribution of the locations: this depends on the locations of the individuals volunteering the information, and a higher level of spatial control in which locations of observations are pre-specified—such as that considered in the design aspects of Myers et al.'s [23] review of methods for response surfaces over data collection strategies that are needed for such optimally designed experiments is lacking.

Nevertheless, despite the above observation there are other benefits to using the public participation approach. Arguably, an advantage is that data is collected by a potentially very large unpaid workforce—for example, although noting the importance of the correct training of staff, [6] observe that the use of CS has resulted in the saving of $30,000 per year on one particular project, and at the time of writing, a large amount of data for a diverse range of applications is collected in this way [17]. However, the aim of this paper is not simply to report that this phenomenon is occurring, as discussion is quite widespread, see for example [7, 16] or more generally a special edition of Geoinformatica from March 2010. However, here we consider a data analysis application, but highlight aspects that need to be considered since data of the above kind is being used. In particular, we consider the analysis of CS data

relating to *phenology*—see for example [29]. Phenology is the study of periodical plant and animal life cycle events and their relation to climate. In recent years, phenology has been used as an instrument to assess evidence of climate change. In particular [5] and [32] use observations of the first bloom and first leaf dates of particular plants recorded over a number of years to assess gradual advances in the onset of spring. The data analysed in these studies contains the recorded first bloom and first leaf dates of lilacs (*Syringa*) from a series of observation locations, with the data provided by a mixture of trained scientists and others.

A key aim here is to provide a demonstration of how data of this kind may be analysed, and in addition to reporting preliminary finding, to discuss and outline some of the issues that were encountered when carrying out the analysis. In the next section, the data set used is described in more detail. The following section outlines an example of the problems that may occur when an inappropriate data analysis technique is employed. A remedy to this problem is then proposed, and from this stand point a number of further directions for analysis are explored. The paper is then concluded with a short summary and discussion.

2 The data

The data here was downloaded from the web site provided by [31],[1] and is derived from two main studies, one in the western states of the US, and one based mainly in the eastern states, with a small number of locations in Canada included also. The former is described in detail in [5], the latter in [29] and [28]. In each case, the data records the first leaf and first bloom dates of the common lilac, expressed as a number of days since the start of the year. The [32] paper combines data from both of these studies to obtain a data set for the whole of the US. The locations are shown in Fig. 1 on a backdrop of national boundaries.[2] Clearly, the east/west divide is not perfect, and the density of observations also changes in the two studies (particularly in the mid west). The change in pattern serves to illustrate the split in the data, which will prove to be significant later on in this paper.

Both studies were implemented via networks of observers. The most detailed description of this is in the Cayan et al. paper—who state that the western survey was initiated by [3]—who describes how the Montana Agricultural Experimental Station set up a network of observers, with contributors from the US Weather Bureau and local garden clubs, to monitor various stages in the annual cycle of the lilac—a very early example of data collected using the CS paradigm. This is quite a complex data set in terms of its provenance—perhaps those data provided by the US Weather Bureau would be more reliable than those of other contributors, although since all of the data is combined it is difficult to assess this hypothesis. A history of this and the eastern network is provided by the [37]. Over time this activity extended geographically, and continued until 1994. There was a subsequent revival of interest in 1999, until the last observations of the data recorded in 2003. The eastern network was initiated later—in 1961—and was at its most active during the 1960s and 1970s.

[1] ftp://ftp.ncdc.noaa.gov/pub/data/paleo/phenology/north_america_lilac.txt

[2] http://www.naturalearthdata.com/downloads/50m-cultural-vectors/50m-admin-0-countries-2/

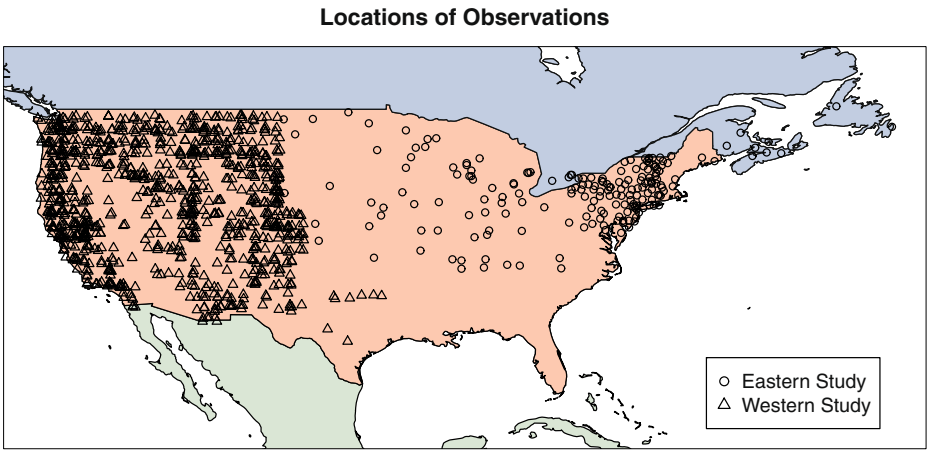


Fig. 1 Locations of lilac observation points

The numbers of observations from each network in five year periods are tabulated in Table 1. Note that in total there are 15072 observations (although only 14265 have recorded the first bloom date) from 1126 distinct locations.

From the table it can be seen that the balance between eastern and western observations in the data set changes over time—in the period from 1995 onwards, the data is dominated by eastern observations—while in the first five year interval, all of the data is from the western network. A more balanced pattern in data collection would be desirable. Going back to the notion of experimental design, the ideal situation would be a uniform coverage of all observation points across the USA over the entire time period. In contrast, this is an example of 'real world' data, where events such as the cessation of public funding, or the loss of a key organising individual—or the emergence of a new key player—can bring about unexpected changes in the pattern of data collection. Whilst not achieving the ideals of the designed experiment, the collection and distribution of this kind of data is at least achievable in terms of resources.

Table 1 Numbers of observations by five year periods for eastern and western lilac phenology networks

	Eastern	Western
1955–1959	0	1997
1960–1964	97	2548
1965–1969	449	2420
1970–1974	640	1965
1975–1979	643	1049
1980–1984	578	778
1985–1989	248	664
1990–1994	231	411
1995–1999	183	8
2000–2004	120	43

Table 2 Regression analysis for the model $B_i = a + b\,Y_i + \varepsilon_i$

	Estimate	Std. error	t value	Pr ($>\|t\|$)
a	126.393	0.246	514.392	0.000
b	0.215	0.018	12.108	0.000

3 An initial analysis and a cautionary tale

Reading the early [3] article suggests that the original interest was in mapping the date of onset of spring itself, rather than in changes in this date over time. This is perhaps not surprising, since when this network began climate change was not an issue. However, it will be argued that this data can be used to investigate climate change, if analysed appropriately.

In Schwartz and Reiter's [32] paper, this data was analysed on a site-by-site basis, with a regression model applied at each location. However, since no site can have more than 46 observations (on for each year between 1957 and 2003)—and many have fewer—the aim here is to attempt an analysis of the data pooled for all of the sites. This then allows a pooled estimate based on over 14,000 observations. An initial approach to analysis is to fit a simple linear model of the form

$$B_i = a + b\,Y_i + \varepsilon_i \tag{1}$$

where B_i is the day of first bloom for observation i, Y_i is the year in which the ith observation was made[3] and ε_i is a normally distributed random error term for each observation i, with variance σ^2. a is the intercept term of the model, and b is the slope, which may be interpreted as the rate of advance ($-$ve b) or retreat ($+$ve b) of the date of first bloom, in days per year. Given that a relatively slow rate of change is likely, one would expect b to be a fairly small quantity, with an absolute value less than 1. The results of the analysis are given in Table 2.

This result suggests that b differs significantly from zero—suggesting that there is evidence that the day of first bloom varies over the study period. However, and rather surprisingly, the estimate for b is positive indicating that the first bloom date is retreating. That is, it is getting later in the year. A plot of Y_i and B_i is given in Fig. 2. The plot format is a binned hexagon plot, in which scatterplot points are allocated to small hexagonal regions, and the size of hexagon drawn in each region is proportional to the count of points contained there—this method is preferable to a standard scatterplot when there are a large number of points (in this case more than 10000)—see for example [4]. The regression line is superimposed on the plot.

Both the plot and the regression analysis show the first bloom day to be getting later—suggesting that the accumulation of thermal time, as the driver of plant development, is getting *slower* over time. This runs counter to expectation, but more importantly from the viewpoint of data analysis, it also contradicts the findings of the Schwartz and Reiter analysis of the same data mentioned earlier. Their multiple analyses of each of the individual observation locations suggest a general trend in which the first bloom day gets earlier.

[3] Centred on 1980—the midpoint of the time interval, as this reduces rounding error when calibrating the model.

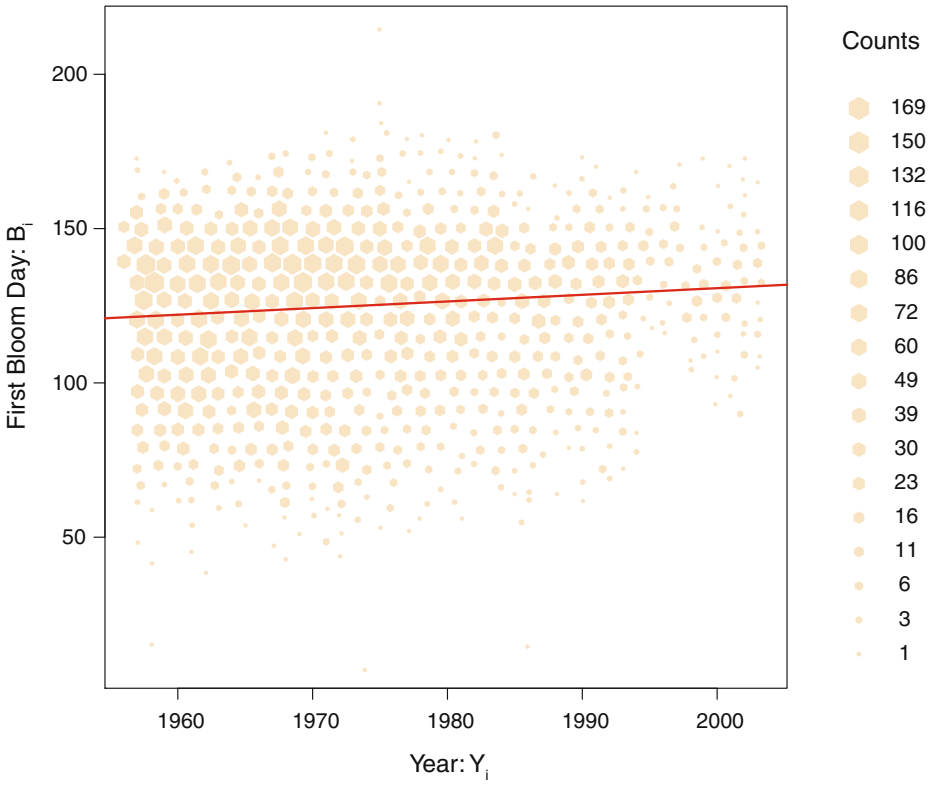


Fig. 2 Plot of first bloom day (B_i) vs. year (Y_i)

The reason for this discrepancy may be understood by further consideration of the data collection process, and in particular the change in geographical emphasis seen over the data collection period. In Fig. 3 box plots are given for the first bloom dates of both the eastern and western data. One notable pattern is that the western data has earlier first, second and third quartiles for the first bloom date than the eastern data. Note that this phenomena is not unique to VGI or CS data—similar situations have been noted with authoritative data; for example [10] explore measurements of sea surface temperature (SST) monitored by sensors on buoys and note that computed annual mean SSTs for 1990 and 1996 were based on quite distinct distributions of measurement locations. Those in 1990 were 'almost exclusively restricted to the tropical Pacific and the northern North Atlantic' whilst those in 1996, while not uniform certainly covered a greater area. Similarly [26] combine several data sets in order to examine patterns in SST, sea ice and night marine air temperature since the late nineteenth century, and in an appendix note variation in geographical coverage over time when comparing these sources of information.

Returning to the phenology data, recall from Section 2, that the eastern data is more prominent towards the end of the study. The extreme values are more expanded for the western data, but this could be attributable to the fact that in general, there were more samples provided by this network. This suggests that

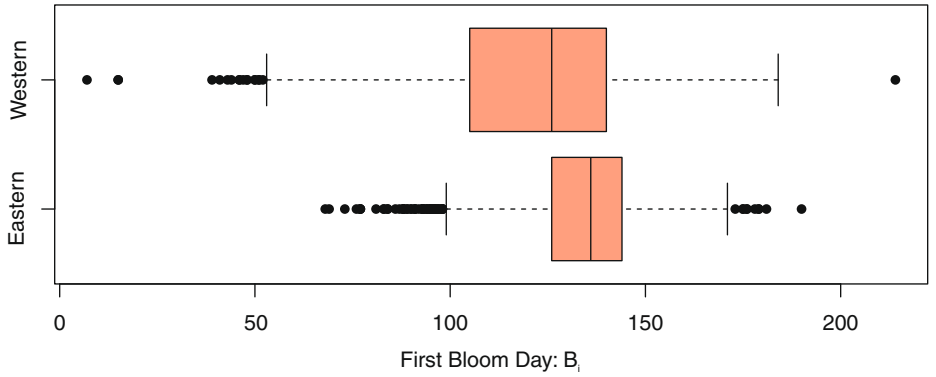


Fig. 3 Box plots comparing first bloom dates for eastern and western data

one possible explanation for the surprising estimate for b is the fact that the later blooming eastern data dominates the pooled set towards the end of the study period, so that, if location is ignored the average first bloom date may indeed increase with year of observation.

This is an example of *Simpson's Paradox* [33]. The paradox is very succinctly stated by [1], who refer to

> ... the dangers of ignoring a covariate that is correlated to an outcome variable and an explanatory one.

In this case, the covariate being ignored is the location of the observation, the explanatory variable is Y_i and the outcome is B_i. The situation here is unusual, in that examples of the paradox more usually involve probabilities or rates estimated using categorical data rather than regression analysis applied to continuous data (see for example [39]), but nevertheless it clearly fits the situation described by Appleton et al.

One step towards addressing this problem is to include an indicator variable in the regression model, giving the updated form:

$$B_i = a + b_Y Y_i + b_N N_i + \varepsilon_i \tag{2}$$

where, in addition to the previously defined variables, b_Y is the regression coefficient for the year, N_i is the network indicator variable for observation i (0 for Eastern, 1 for Western), and b_N is the regression coefficient for this variable. Thus, this coefficient is a measure of the difference in first bloom date (on aggregate) between the eastern and western networks. Implicit in this model is a uniformity of rate of change of first bloom date across both networks, and a uniformity of the intercept term within each

Table 3 Regression analysis for the model $B_i = a + b_Y Y_i + b_N N_i + \varepsilon_i$

	Estimate	Std. error	t value	Pr ($>\|t\|$)
a	134.009	0.405	331.198	0.000
b_Y	0.036	0.019	1.918	0.055
b_N	−11.593	0.495	−23.439	0.000

Table 4 Regression analysis for models fitted separately to each network

	Estimate	Std. error	t value	Pr ($>\|t\|$)
a (Eastern)	134.017	0.406	330.473	0.000
a (Western)	122.382	0.315	388.282	0.000
b_Y (Eastern)	0.048	0.041	1.164	0.244
b_Y (Western)	0.033	0.021	1.555	0.120

network, if the intercept is considered to be a for the eastern network, and $a + b_N$ for the western one.

The results of fitting this model are given in Table 3. In this case, the slope for Y_I is still positive, but no longer significantly different from zero $p > 0.05$. This suggests the possibility that Simpson's paradox is occurring here—as allowing for geography even in a fairly crude way does change the results relating to the slope term.

A further investigation is possible by splitting the analysis into two regression models—one for each of the two data collection networks. This gives the results in Table 4. In his case the two slopes are both positive, although differing slightly—and, as with the above result, neither are now significantly different from zero.

However, it could be argued that some geographical effects are still ignored. In the paper by [3] maps of the first bloom date for the western area show notable geographical patterns exist *within* that area—and it is a reasonable expectation that similar variation may also occur in the eastern area. Thus, it may be the case that effects due to Simpson's paradox may still be influencing the results given in Table 3. In the next section, approaches to modelling the data allowing for more comprehensive variational effects in the model coefficients will be considered.

4 Proposed alternative analyses

The issue relating to Simpson's paradox in the previous section arises essentially from the failure to incorporate sufficient information about geographical variation in the model. As a starting point to address this, suppose initially that the slope is the same everywhere, but that each observation station has a different intercept. That is, we assume there is a 'green wave' [30] across the US so that some regions see the first bloom of lilac before others, but the rate of change of onset of this wave is uniform across the US. We can model this by

$$B_{ij} = a_j + b_Y Y_{ij} + \varepsilon_{ij} \tag{3}$$

where the extra subscript j denotes a quantity relating to observation i at station j—thus B_{ij} is the ith first bloom date at station j, which was observed in the year Y_{ij} and so on. Note that a_j only has the j subscript, suggesting that there is a different first bloom date associated with each station, but not a unique one for each year at each station in the model proposed by Eq. 3.

Such a model could be calibrated using ordinary regression, treating the a_j's as series coefficients for dummy variables indicating which station each observation occurred at. However, recalling that there are 1126 distinct locations in the data, this would require a large number of coefficients to be calibrated, with a resulting increase of degrees of freedom in the model, and a resultant increase in the standard error of the estimate for b_Y. An alternative approach is to use a *random coefficient*

model [19] where the a_j's are themselves assumed to be random variables: for example

$$a_j \sim \mathrm{N}(a, \sigma_a^2) \tag{4}$$

where a is the distribution mean of the a_i's and σ_a^2 is the variance. Thus the likelihood of the observed data can be written in terms of just four parameters: a, σ_a^2, b_Y, and σ^2, rather than over 1100 parameters as in the model considered at the beginning of this section. Another justification of this approach is that since focus is based on the estimation of b_Y, rather than attempting to calibrate every a_j exactly, the aim here is more simply to take into account the fact that a_j varies, and to characterise this variability using a small number of parameters, namely a and σ_a^2.

Note that, by writing a_j as the sum of a and a zero-centred random variable υ_j (with variance σ_a^2) Eq. 3 can be re-written as

$$B_{ij} = a + b_Y Y_{ij} + \upsilon_j + \varepsilon_{ij} \tag{5}$$

—this is very similar to model 1, except that now the random part of the model consists of two terms, representing variation at the observation level and at the location level. For this reason, the model can also be described as a *multi-level model* [12–14]. In this paper the convention is adopted that random coefficints in models predicting the first bloom date will be denoted with Greek letters, and fixed coefficients will be denoted with Roman letters.

In model 1 the random part of the model is independent for each observation, but for model 5 it may be checked that for two obversations ij and kl (i.e. the first observation is the ith at station j, and the second is the kth observation at station l), the correlation between the random terms, $\rho_{ij:kl}$, is given by:

$$\rho_{ij:kl} = \begin{cases} \dfrac{\sigma_a^2}{\sigma_a^2 + \sigma^2} & \text{if } j = l \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Arguably, this is a crude representation of Tobler's first law [36]:

> Everything is related to everything else, but near things are more related than distant things.

—here, 'near things' are considered to be observations taken from the same location, regardless of time. Observations at the same location are correlated, whereas those from different locations are not.

The results of fitting this model are shown in Table 5.

From this it may be seen that the estimate for b_Y is now negative, and is significantly different from zero. With an estimated value of around −0.177 this suggests that the onset of spring advances by around one day every six years. This

Table 5 Regression analysis for the model $B_{ij} = a + b_Y Y_{ij} + \upsilon_j + \varepsilon_{ij}$

	Estimate	Std. error	t-value	Pr (> $\lvert t \rvert$)
a	121.892	0.646	188.555	0.000
b_Y	−0.177	0.010	−17.184	0.000

may be compared with some other results, such as [21] who find the same rate of advancement based on a data set covering a range of phenological indicators from data from the International Phenological Gardens (IPG), a Europe-wide network.

Although the focus of this study has been the estimation of b_Y it is still possible to estimate the individual a_j's via the multi-level model. Effectively, the estimate for each a_j is achieved by computing an estimate of $\mathrm{E}(a_j|B, Y, a, b_Y, \sigma^2, \sigma_a^2)$, where B and Y are the respective vectors of all B_{ij} and Y_{ij} observations, and estimates of a, b_Y, σ^2 and σ_a^2 are obtained when calibrating model 5 above. These are shown in map form in Fig. 4. From this, it can be seen that earlier first bloom dates tend to occur along the west coast of the US, and also that elsewhere there is a north-south trend, with spring arriving later in the north.

As well as obtaining a map of the 'green wave' as experienced through the first bloom dates of lilacs, this demonstrates that even the approach of the model of Eq. 2 failed to reflect the full geographical variability in first bloom dates—Fig. 4 suggests that geographical variation also occurs within both networks. This is manifested in the change in the estimate of b_Y as each of the models from Eqs. 1, 2 and 5 are calibrated in sequence.

Thus, in this study an estimate of b_Y (with standard error) taking into account the variability of an intercept term. However, this assumes that the relationship between the first bloom date and the year of observation is linear, so that the change in B per year is fixed over the entire study period. A more flexible approach is to estimate a general time effect, so that rather then modelling the temporal change in first bloom date with the regression term $b_Y Y_{ij}$, an alternative model replaces this by an effect for each year, say c_i for each year indexed by i. As with locations, these year-wise effects can be modelled as random effects, so that

$$c_i = c + \tau_i \text{ where } \tau_i \sim \mathrm{N}(0, \sigma_c^2) \tag{7}$$

and therefore

$$B_{ij} = A + \tau_i + \upsilon_j + \varepsilon_{ij} \tag{8}$$

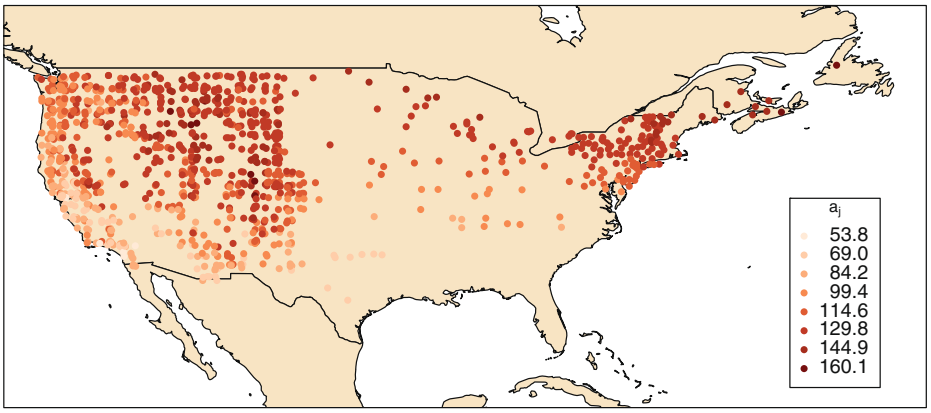


Fig. 4 Map showing estimated a_j

where the single term A replaces $a + c$ to avoid redundancy in the model (any pair of a and c adding up to A would give the same model likelihood for a given data set). This model is still a random coefficient model, but it can no longer be described as a multi-level model—as the effects are no longer nested—effects for time do not nest within locations. Models of this kind are referred to as *crossed-effects models* [2]. The result of fitting a model of this kind is shown in Fig. 5. This plot shows the variation in τ_j over the time period 1956–2003, also giving bands showing standard errors for the estimates, based on bootstrapping techniques [9]—see "Appendix" for detail.

The pattern seen in Fig. 5 shows a general advance in the first bloom date over the study period, although it does suggest that the trend is more complex than the linear model used until now. In particular, in the second half of the study period, it appears that there is some degree of oscillation around a general downward trend, with the oscillation period being just over a decade or so. Also of note is the fact that the standard error bands are notably larger from around 1986—this was the time that funding of these networks begin to reduce, and in turn, the numbers of observations were also reduced. In terms of the model calibration, this is reflected in greater uncertainty of parameter estimation.

One final parameter of interest may be of use here—this is used to measure the trend of τ_i of time. Clear, in strict terms, it cannot be said that τ_i is decreasing *every* year—the oscilatory effect suggests that there will be some pairs of consecutive years when τ_i increases. However, it is helpful to consider whether there is an *overall trend* towards lower τ_i values. A straightforward way to measure this is to compute the difference in mean values of τ_i for the first and second halves of the study period. That is, the respective periods 1956–1979 and 1980–2003. If this statistic is called Δ then

$$\Delta = \frac{1}{24}\left[\sum_{i=1,24} \tau_i - \sum_{i=25,48} \tau_i\right]. \tag{9}$$

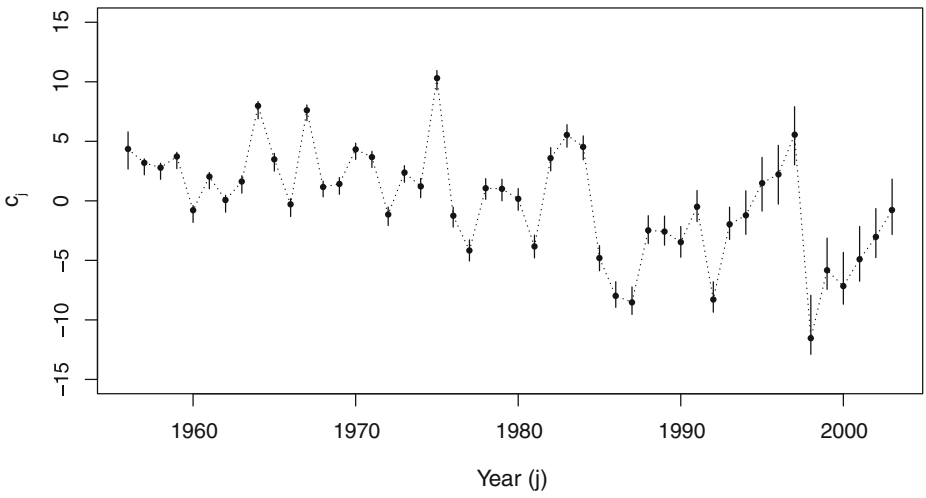


Fig. 5 Graph showing estimated τ_j against time, with upper and lower pointwise confidence intervals for the individual τ_j values

Table 6 Estimate and bootstrap confidence intervals for Δ

	Estimate	Lower 95% CI	Upper 95% CI
Δ	4.65	3.71	4.76

Thus, positive values of Δ suggest a trend of spring getting earlier, and negative values suggest it is getting later. The estimate of Δ, together with its 95% bootstrap confidence intervals are given in Table 6. From this there is fairly strong evidence of a trend of spring getting earlier. That is, despite some fluctuation around the trend, the average first bloom date in the second half the the study period is earler than that of the first half.

5 Concluding discussion

In this paper, a number of models have been made to estimate the change in the onset of spring over a time period from 1956 to 2003, making use of data collated from a number of networks, making use of voluntary data. The data consisted of observed first bloom dates of lilacs. These dates are strongly linked to the accumulated thermal time of the plants, and hence act as a proxy for patterns in seasonal temperature. From an initial analysis that was distorted due to Simpson's paradox, the final analysis takes into account the changes in geographical distribution of the networks of people collecting data, and identifies both general trends and fluctuations in the onset of spring.

The uncertainties and biases associated with analysing such data due to the voluntary nature of its collection have long been recognised. For example, hoopoe are birds that are occasionally seen in the UK with a preference for long grass habitats with tree cover. Much historical data in the UK relating to the sighting of these birds records them in the gardens of vicars—a combination of habitat and a 19th century predilection among the clergy for recording nature. Moreover, the situation is not unique to VGI or CS data—for example, in more 'official' projects, such as those set out in the examples discussed in Section 3 where geographical distribution of sea surface temperature sensors changes varies over time, over- or under-representing certain regions or varying in resolution.

However as yet little work has explored the sensitivity and reliability of phenology data. Some ecological research has explored the variation and uncertainties associated with the use of phenological data. Robbirt et al. [27] compared plant specimens (herbarium data) with field observations and found that the response of flowering time to variation in mean spring temperature to be identical and much of the variation in the results to be due to the geographic location of the collection sites—a factor which we have also found to be important in the analysis above. Also, there are other factors which may need to be considered: Miller-Rushing et al. [22] compared herbarium data with phenological events as recorded in dated photographs. They suggested that first flowering dates may not be ideal measures of plant responses to climate change due the the extremes of flowering distributions being more susceptible to confounding effects than central values. This is perhaps

another situation where there is a trade off between the ideal situation, and what may be achieved in practice. Central values, such as means, would require observation of *all* bloom dates at a given location, which may require more observational effort than can be realistically provided by a volunteer network. A compromise may be to obtain a central measure such as the mid-point between the first and last blooms (although this may still suffer from the problem of being sensitive to extremes). Of course, any such recommendations can only apply to future data, as recording the first bloom date is already a well established convention—and a great deal of data using this convention already exists.

A further issue relates to the linkage between phenological event timing and temperature: van Oort et al. [38] explored the sensitivity of phenological events and the possible correlation between temperature and phenology prediction error of rice and found that phenological models were not as sensitive as thought at the higher end of the temperature range. As this study concentrates more on the timing of the phenological events, this finding perhaps has less direct bearing on the analysis, however, it does perhaps have implications when interpreting the observed patterns.

In this paper, methods for addressing this issue of geographical variation were considered—the adoption of these being largely guided by considering the process used to collect the data. However, there is room for further work to improve on this. For example, in situations where [22]'s concerns regarding the use of first bloom dates are likely to affect the outcomes of analysis, robust regression techniques or a distribution model for residuals having heavier tails than the Gaussian could be applied.

Another modification to the model could take into account the temporal auto-correlation of the τ_i coefficients. Currently these are assumed to be random, but independent, but could be assumed to follow a multivariate normal distribution with a variance/covariance matrix reflecting this temporal structure. Similarly the values of υ_i could be assumed to exhibit a spatially autocorrelated structure. Exploiting the latter structure would allow values of υ at points other than the observation points to be estimated—as it provides information as to the degree to which observations of υ values near to a point of interest influence the value *at* that point. Estimating or visualising the υ values was not the key focus of this study—which placed more emphasis on change of the first bloom date, but in a study where spatial variation was the main subject, this spatial modelling approach could be used to create a pixel-by-pixel surface of estimated υ values, providing a visualisation of the 'green wave'. This could also be extended to allow for spatial variation in the b_Y coefficient.

However, addressing some of these more advanced issues may call for more advanced computing tools. The R package `lme4` used here only offers random coefficient models where the distributions for coefficients are independent; an alternative package `lmer` allows for non-independent random coefficients, but at the time of writing does not offer calibration of crossed-effects models. One way of overcoming this is to use Markov chain Monte Carlo (MCMC) approaches to calibrate the model, in the manner of [11], but this would require considerably more computing effort, the use of notably more complex software, and a change in statistical inferential paradigm from classical to Bayesian. These changes may well all be justified, but it is hoped that one of the main messages in this paper is that the analytical techniques used when working with any data need to reflect issues arising from the process of data acquisition, and that any data set is a reflection of both the

underlying natural process and the process of data collection and organisation, and a useful analysis of this data needs to reflect this.

Appendix: Computational considerations

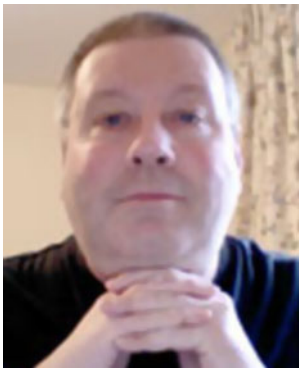
In this section, some more detail is supplied about the software tools and techniques that were used to carry out this analysis. All of the statistical modelling was carried out using the R statistical programming language [25]. In particular, the random coefficient models were calibrated using the `lme4` package.

The functions supplied in the R base library and `lme4` were sufficient for all of the computations, except for the standard errors associated with the τ_i values, and Δ. For these, a regression bootstrap approach as set out in [9] is used. Briefly, this estimates the sampling variation of parameters of interest by simulating data sets drawn from the model that is being fitted to the data (in this case the model given by Eq. 8). The sampling variation simulated is just that due to the variability in ε_{ij}—so that rather than randomly assigning new values for the τ_j's and υ_i's for each simulated sample, it is assumed they are fixed at the estimated values. By simulating a large number of data sets in this way (say 1000, as in this paper), and applying the random coefficient estimation function supplied by `lme4` to each simulated data set, an estimate of the sampling variability of the τ_j's is obtained.

References

1. Appleton D, French J, Vanderpump M (1996) Ignoring a covariate: an example of Simpson's paradox. Am Stat 50(4):340–341
2. Baayen R, Davidson D, Bates D (2008) Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang 59:390–412
3. Caprio J (1957) Phenology of lilac bloom in Montana. Science 126:1344–1345
4. Carr DB (1991) Looking at large data sets using binned data plots. In: Buja A, Tukey P (eds) Computing and graphics in statistics. Springer, Berlin
5. Cayan D, Kammerdiener S, Dettinger M, Caprio J, Peterson D (2001) Changes in the onset of spring in the western united states. Bull Am Meteorol Soc 82(3):399–415
6. Cohn JP (2008) Citizen science: can volunteers do real research? BioScience 58(3):192–197. doi:10.1641/B580303
7. Coleman D (2010) The potential and early limitations of volunteered geographic information. Geomatica 64(2):27–39
8. Cooper CB, Dickinson J, Phillips T, Bonney R (2007) Citizen science as a tool for conservation in residential ecosystems. Ecol Soc 12(2):11
9. Davison A, Hinkley D (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge
10. Emery W, Baldwin D, Schlüssel P, Reynolds R (2000) Accuracy of in situ sea surface temperatures used to calibrate infrared satellite measures. J Geophys Res 106:2387–2405. doi:doi:10.1029/2000JC000246
11. Gelfand A, Banerjee S, Sirmans C, Tu Y, Eng Ong S (2007) Multilevel modeling using spatial processes: application to the Singapore housing market. Comput Stat Data Anal 51:3567–3579
12. Goldstein H (1986) Multilevel mixed linear model analysis using iterative generalised least squares. Biometrika 73:43–56
13. Goldstein H (1987) Multilevel covariance component models. Biometrika 74:430–431
14. Goldstein H (1987) Multilevel models in educational and social research. Griffin, London
15. Goodchild M (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221. doi:10.1007/s10708-007-9111-y

16. Haklay M (2010) How good is volunteered geographical information? A comparative study of openstreetmap and ordnance survey datasets. Environ Plann B, Plann Des 37(4):682–703
17. Hand E (2010) Citizen science: people power. Nature 466(7307):685–687. doi:10.1038/466685a
18. Lister M Adrian, the Climate Change Research Group (2011) Natural history collections as sources of long-term datasets. Trends Ecol Evol 26(4):153–154
19. Longford N (1993) Random coefficient models. Clarendon Press, Oxford
20. McCaffrey RE (2005) Using citizen science in urban bird studies. Urban Habitats 3(1):70–86
21. Menzel A, Fabian P (1999) Growing season extended in europe. Nature 397:659
22. Miller-Rushing A, Primack R, Primack D, Mukunda S (2006) Photographs and herbarium specimens as tools to document phenological changes in response to global warming. Am J Bot 93:1667–1674
23. Myers R, Montgomery D, Vining G, Borror C, Kowalski S (2004) Response surface methodology: a retrospective and literature survey. J Qual Technol 36:53–78
24. Myers JL, Well A, Lorch RF (2010) Research design and statistical analysis, 3rd edn. Routledge, New York
25. R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
26. Rayner N, Parker D, Horton E, Folland C, Alexander L, Rowell D, Kent E, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108:4407. doi:doi:10.1029/2002JD002670
27. Robbirt K, Davy A, Hutchings M, Roberts D (2010) Validation of biological collections as a source of phenological data for use in climate change studies: a case study with the orchid ophrys sphegodes. J Ecol 99(1):235–241
28. Schwartz M (1997) Phenology of seasonal climates. In: Lieth H, Schwartz M (eds) Spring index models: an approach to connection satellite and surface phenology. Backhuys, Netherlands, pp 23–38
29. Schwartz MD (1994) Monitoring global change with phenology—the case of the spring green wave. Int J Biometeorol 38(1):18–22
30. Schwartz M (1998) Green-wave phenology. Nature 394(6696):839–840
31. Schwartz M, Caprio J (2003) North american first leaf and first bloom lilac phenology data. IGBP PAGES/World Data Center for Paleoclimatology Data; Contribution Series # 2003-078; NOAA/NGDC Paleoclimatology Program, Boulder CO, USA
32. Schwartz M, Reiter B (2000) Changes in north american spring. Int J Climatol 20(8):929–932
33. Simpson E (1951) The interpretation of interaction in contingency tables. J R Stat Soc, Ser B Stat Methodol 13(2):238–241
34. The Guardian (2011) Spring's here: skylarks overhead, moles in the garden, moths in the bathroom. URL http://www.guardian.co.uk/environment/2011/mar/27/spring-wildlife-black-mountains-wales
35. The Guardian (2011) Weatherwatch: phenology in the UK. URL http://www.guardian.co.uk/news/2011/apr/11/weatherwatch-phenology
36. Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46:234–240
37. USA National Phenology Network (2011) History of lilac and honeysuckle phenological observations in the USA. http://www.usanpn.org/?q=node/36
38. van Oort P, Zhang T, de Vries M, Heinemann A, Meinke H (2011) Correlation between temperature and phenology prediction error in rice (*Oryza sativa L.*). Agric For Meteorol 151(12):1545–1555
39. Wagner CH (1982) Simpson's paradox in real life. Am Stat 36(1):46–48. URL http://www.jstor.org/stable/2684093



Chris Brunsdon is professor of human geography at the University of Liverpool. His interests include the methodologies underlying spatial statistical analysis and geographical information systems, and their application in a number of subject areas.

Lex Comber is a senior lecturer in the Department of Geography in the University of Leicester. His research interests are in two primary areas: geocomputation and spatial analysis of policy.