Statistical issues with using herbarium data for the estimation of invasion lag-phases

R. J. Hyndman · M. B. Mesgaran · R. D. Cousens

R. J. Hyndman (✉)
Department of Econometrics and Business Statistics, Monash University, Victoria 3800, Australia
e-mail Rob.Hyndman@monash.edu

M.B. Mesgaran · R.D. Cousens
School of BioSciences, The University of Melbourne, Victoria 3010, Australia

Abstract Current methods for using herbarium data as time series, for example to estimate the length of the invasion lag phase, often make assumptions that are both statistically and logically inappropriate. We present an alternative statistical approach, estimating the lag phase based on annual rather than cumulative data, a generalized linear model incorporating a log link for overall collection effort, and piecewise linear splines. We demonstrate the method on two species representing good and poor data quality, then apply it to two data sets comprising 448 species/region combinations. Significant lags were detected in only 28% and 40% of time series, a much lower level than the 95% and 77% found in previous analyses of the same data. In a case with high quality data, a lag was concluded even though during the "lag" the locations of herbarium collections indicated that it was spreading rapidly at a continental scale. In species with few records, results were sensitive to the way in which zeroes were included. Overall, our method gives very good fit to the data, avoids unrealistic assumptions of other methods and gives more reliable estimates of confidence. However, given the poor representation of herbarium samples in the early stages of invasions and the fact that they do not constitute a structured survey of abundance, we warn against over-reliance on statistical analysis of such data to reach conclusions about the dynamics of invasions.

Keywords Lag phase · invasion · herbarium · statistical analysis

Introduction

Invasion ecologists often comment on the existence of lag phases in invasions. Although precise definitions are seldom given, a lag is usually stated to be a period of negligible or slow range expansion after species introduction (perhaps also reflecting slow increase in abundance or restriction to just a small number of locations) and preceding a phase of more rapid increase (Pyšek and Hulme 2005; Aikio et al. 2010; Larkin 2012; Aagaard and Lockwood 2014). There are many plausible reasons for an acceleration of spread in this manner, including intrinsic (population genetic, demographic, dispersal and developmental) and extrinsic (changes in the physical or biotic environment) factors, which may also interact with landscape structure (e.g. Cousens and Mortimer 1995; Pyšek and Hulme 2005). Given that rate of population spread is the product of dispersal and population growth rate (Kot et al. 1996), the change from a low rate of expansion to a much greater one must necessarily involve a change in either, or both, of these factors [we do not subscribe to the view (Crooks and Soulé 1999) that an exponential growth curve has an "inherent" lag phase]. Direct evidence linking a particular invasion lag to a given mechanism (e.g. Wangen and Webster 2006) is, however, rare.

Formal and informal analyses, as well as comments in reviews and books, have suggested that lag phases may be not only very lengthy in some cases but also common (Pyšek and Prach 1993; Kowarik 1995; Mack et al. 2000; Ellstrand and Schierenbeck 2000; Sakai et al. 2001; Lee 2002; Parker 2004; Pyšek and Hulme 2005; Holt et al. 2005; Williamson et al. 2005; Keller and Taylor 2008; Daehler 2009; Dogra et al. 2010; Lenda et al. 2012; Booth et al. 2011; Larkin 2012; Aagaard and Lockwood 2014). Unfortunately, lack of formal surveys during the early years of invasions make it difficult to observe rates of spread directly (by the time we realise the species is spreading, it is too late to initiate monitoring). For plants, herbarium data are increasingly available in digital format, and potentially allow us to attempt to reconstruct invasion histories (e.g. Rodman 1986; Cousens et al. 2013), to determine periods of species influx (e.g. Fuentes et al. 2008), to chart invaded area and range limits over time and to calculate rates of spread (Pyšek and Hulme 2005), to predict eventual invasion extents (using species distribution models based on climate: e.g. Webber et al. 2011) and to examine invasion niche shifts (Gallagher et al. 2010). They may also allow us to estimate lag phase lengths.

There are a number of problems with the use of herbarium data to measure population spread. First, they are seldom – if ever – an entirely systematic or random

sample of species distributions, either spatially or temporally. Most commonly they are collected at the discretion of botanists (who vary in their motivations, times, frequencies and regions of travel, and their reasons for being in the field). They may sometimes be part of a general survey or an intensive survey of a particular species, and they will be subject to the current priorities of their organisation. The abundance of collectors also changes considerably over time. As a result, collection "effort" and therefore the probability of both detection and collection, will change over time (Aikio et al. 2010). Although the number of specimens collected over a particular period may well reflect, in part, the abundance and distribution of a species, they will reflect these other factors as well. That is, there is a huge amount of "noise" and potential bias.

Second, as we have mentioned, data are either absent or extremely sparse during the early stages of most invasions. We seldom target species invasions for study until after they have started to spread actively. Moreover, the number of observers/collectors (and hence the probability of being collected) was very low during the historical periods when our current major invaders were just establishing themselves (Fig. 1). This means that it is very difficult to distinguish statistically – or visually – between a true lag phase (where there is a distinct change in rate of increase in species records) and the early parts of a continuous, rapid increase (Cousens and Mortimer 1995). An absence of data between the first occurrence and the start of the rapid increase phase will mean that regression models incorporating a lag phase appear to fit extremely well, simply because there are no data to indicate lack of fit in that region (e.g. Larkin's 2012 Fig 2b). Cousens and Mortimer (1995) also pointed out that it is common for observation effort to increase once a species has been identified as an issue (e.g. if targeted for eradication), inflating the apparent rate of increase and introducing apparent phase changes where they may not exist.

Third, the data are "censored". The date of an herbarium specimen only tells us the date by which the species was already present at a site; it could have appeared many years earlier, but it could not have appeared later. If there is documentary evidence for the date of first deliberate planting, this may not be a problem for the first occurrence (Kowarik 1995), but it remains an issue for accidental introductions and for every further observation as the species spreads. Low sampling effort and low abundance can lead to large errors in first detection at a location. Even using a statistically appropriate method, we will still obtain biased (underestimated) lag phases because the date of first arrival may be many years (or decades) earlier than the first sample is collected. This error is again likely to be more pronounced in historical times when there were few collectors to cover

very large areas. For example, on the first Australian herbarium specimen of the alien *Cakile edentula* it is stated that the species has been "*known there wild since 20 years*". In South Africa, the first herbarium specimen of *Nassella trichotoma* was from 1952, yet farmers had known about it prior to 1930 (Wells 1974).

Notwithstanding such limitations of the data, herbarium databases have been used on numerous occasions to extract information about rate of invasion and to search for any points at which the invasion rate changes abruptly (e.g. where the lag phase ends). A recent approach developed by Aikio et al. (2010) uses time series of numbers of specimens to model collection rate of a species (number of herbarium specimens collected per year); others have also used time series of numbers of specimens or have converted the data to time series of occupied grid cells in order to model area of occurrence (e.g. Williamson et al. 2005). Both approaches make a range of assumptions, some of which are shared and others are particular to one type of data or the other. These assumptions can be argued for or against and their relative strengths are in many cases unclear. For example, in modelling the collection rate, resampling of the same locations may give information on how abundance changes over time; in modelling area of occurrence, multiple occurrences within an area must be eliminated to avoid overestimation of area. In this paper we will focus on the Aikio et al. (2010) approach since most aspects of our statistical methods, and the statistical issues that they seek to overcome, are generic. Rather than argue about what to model, we focus on how to model.

Aikio et al.'s (2010) method (and the application of it by Larkin 2012) is based on a two-segment regression of the cumulative number of herbarium specimens against time. If a two-segment model, with the first phase having a lesser slope, explains a significant amount of variance in comparison with a one-segment model (using the Akaike Information Criterion, AIC), then a change in rate of collection (and by inference a change in rate of invasion) is supported and the intersection of the two models is regarded as the end of the lag phase. The first use of this approach appears to be by Pyšek and Prach (1993) although few details are given in that paper. In order to estimate the length of the lag, an assumption also needs to be made that the first record is an accurate representation of the start of the invasion (see above discussion on censored data). In recognition of the fact that collection effort changes over time, Aikio et al. adjusted the collection rate of a species to allow for temporal variation in overall alien species rate of collection (assumed to be proportional to the number of all alien herbarium samples in that time interval). Both Aikio et al. and Larkin's studies concluded that lag phases were common,

statistically significant in at least 95% and 77% of cases respectively for which there were sufficient data (defined as >15 herbarium specimens).

Aikio et al.'s (2010) method, as inferred by its title, aims to separate true lag phases from artefacts of the data and to give an accurate estimate of the length of the lag phase. However, their analysis relies on fitting statistical models (in this case parametric ones). Models, even ones that appear to fit the data well, are abstract representations of reality: if the model is inappropriate or makes implausible assumptions, then it is quite possible that the statistical analysis will itself produce artefacts and give incorrect estimates/inferences (either within or outside the range of the data). Although the parametric functions used in their analysis appear to describe the cumulative data reasonably well, they are unlikely to be generally applicable. The four models fitted by Aikio et al. (2010) all assume that there is a sudden jump in collection rate of a species (even after adjustment for changes in overall collection effort) immediately the lag phase ends (Fig 2a-d). If collection rate is simply proportional to the species' abundance or distribution (a fundamental assumption in this use of herbarium data to estimate lag phases), the collection rate an infinitesimally short time after the end of the lag phase should barely change: the species will still be at a very low abundance even though increasing more rapidly (Fig 2e). There would only be a sudden jump in collection rate if collector behaviour changed (rather than rate of spread/population increase), for example as a result of the recognition of the species as an increasing threat that needs to be monitored (Cousens and Mortimer 1995), or if the species was spread throughout a region in a single dispersal event acting over an *extremely* short period. Two of the models – the ones that provided the best fit to the data in both Aikio et al. and Larkin's studies – also assume that after the second (more rapid) phase of invasion, the annual rate of specimen collection will fall to zero (Fig 2a,c). This would only be the case if collectors completely lost interest in the species for some reason and not because the species stops increasing. Herbarium specimens are often collected as part of plant community or regional surveys, so that it is unlikely that collection would ever stop completely.

There are also other statistical and logical problems with Aikio et al.'s (2010) method. First, the use of cumulative data in a least squares regression makes the statistically unrealistic assumptions that residuals are independent and that the variance is homogeneous. Because the next observation is added to the previous one, the errors in the first observation contribute to errors in the second, and so on. As the cumulative number becomes larger, so will the variance. Both of these assumptions will lead to incorrect

estimates of error and hence may result in incorrect statistical inferences; for example, the use of cumulative data can underestimate the true standard errors considerably (Mesgaran et al. 2013). This phenomenon is well-known in econometrics, where it is called "spurious regression" (Davidson and MacKinnon, 2004, chapter 14). Two of the assumptions implicit in using the usual AIC for regression models — that the errors are both independent and homoscedastic — are clearly inappropriate with these cumulative models. Another implicit assumption of using cumulative data is that all occasions of zero records in a year are true estimates of abundance, rather than missing values. If there are no new specimens of any alien species collected in that year, this could have arisen because no effort was put into collecting alien species in general, rather than because the target species was too low in abundance to have been collected. A zero therefore may indicate a complete absence of information rather than a failure to find the species in a structured survey. Of course, an absence of new records for the focal species could also have occurred because no botanists interested in aliens visited its habitat and current location in that year, again indicating that it may be more appropriate to treat that year as a missing value. Larkin (2012) also noted that lag lengths were correlated with the date of first appearance of the species and interpreted this as being because, for short time series of recently introduced species, insufficient time would have passed for long lags to have ended. However, it may also indicate that even after correction for temporal variation in sample collection rate, the lower historical collection effort may exaggerate the length of the lag phase.

The aim of our paper is to demonstrate an alternative, “semi-parametric”, statistical method based on annual collection rates rather than cumulative specimen number and with more appropriate assumptions about the data. The method also avoids the assumptions implicit within the regression models used previously. We illustrate our method using an example taken from a database representing the combined collections of the major herbaria in Australia and one example of a smaller data set from New Zealand. We then compare our results with the previous studies of Aikio et al. (2010) and Larkin (2012) and show how our approach leads to very different conclusions.

Material and Methods

Data

We illustrate our method in detail using *Cakile maritima* in Australia and *Holcus lanatus* in the South Island of New Zealand. The reason for their selection is somewhat arbitrary: *C. maritima* is a coastal species that we have studied closely (Cousens et al. 2013) and is represented by 603 records in the Australian Virtual Herbarium (AVH), a database combining the records of all major Australian herbaria (The Council of Heads of Australasian Herbaria 1999). This is a large sample size for non-European collections. *H. lanatus* is a pasture species estimated to have a long lag phase by Aikio et al. (2010). It suffers from a marked gap in records during the early part of its spread and contains only 50 records; it therefore represents a data set whose analysis represents a statistical challenge. Number of *C. maritima* specimens per year and total number of alien specimens collected per year in Australia were obtained from the AVH data by filtering with a species list of Australian alien plants supplied by R. P. Randall. Duplicate specimens (same location, same collector, same year) were removed. The same procedure was used for the database of specimens in the Allan Herbarium (CHR), New Zealand used by Aikio et al. (2010).

To make a direct comparison of the results of the original analytical method and our own, we then reanalysed all the other time series (species × region combinations) used by Aikio et al. (2010) and Larkin (2012) for the north and south islands of New Zealand and the midwest US respectively. Further details on the sources of the numbers of specimens collected in a given year can be found in the above references. Although the New Zealand and US data sets cover a restricted geographic region and thus, in theory, may be able to focus on invasions as they first occur at a local level, they will also be highly limited in statistical power due to the restricted number of samples. Aikio et al. (2010) and Larkin (2012) included data sets with as few as 15 specimens, often spread over many decades. In such cases, it is quite possible for the statistically most parsimonious model to be biologically inappropriate due to high residual variance or large gaps in collection. As we have discussed, in the lag phase and the early parts of a phase of more rapid increase, the data will be of extremely poor quality even in large data sets and even if later phases are well-defined.

Methods of Analysis

Instead of analysing cumulative data, we analysed the number of herbarium specimens of the focal species collected in each year (n_t); following Aikio et al. (2010), we refer to this as

collection rate. As the abundance or distribution of a species increases, so will the probability of a sample being collected. The use of herbarium records as a surrogate for abundance or distribution relies on an assumption of constant collection effort. Clearly, this is not the case. The number of collectors (Fig. 1), the general collection strategy, attitudes towards aliens (in some periods some herbaria did not welcome collections of "weeds", while at other times there have been deliberate changes in policy to focus temporarily on aliens), changes in collection effort for the focal species, and other factors together affect the probability of a species sample being collected. Unfortunately, information on most of these is not available. We therefore used the total number of samples of alien species collected in that year (N_t) as a multiplier in the expected collection rate, to allow for changes in overall alien collection effort (assuming in the absence of other information that effort is independent of species, cf. Aikio et al., 2010). We assume that n_t, the number of specimens of the focal species collected in year *t,* has a Poisson distribution with mean equal to $N_t \exp[f(t)]$ where $f(t)$ is a function of time allowing the number of specimens of the focal species in a particular year to change over time. If $f(t)$ is a linear function of time, then this would be a generalized linear model with a log-link function, and with $\log(N_t)$ as an offset term (Dobson 2008, p.152). The Poisson assumption is consistent with count data and allows for heterogeneity. It would be a simple matter to allow over-dispersion in this model in the usual way if required. Note that $f(t)$ will take negative values, but the exponential constrains the mean of the distribution to be positive. The equation for our model is

$$n_t \sim \text{Poisson}(N_t \exp[f(t)]) \qquad (1)$$

We assume that a lag-phase exists if $f(t)$ is constant until some year τ, and then takes some other higher values thereafter. This is exactly the same definition of the lag phase used by Aikio et al. (2010): they argued that a constant search rate of an unchanging population would result in a linear accumulation of records (i.e. a constant average specimen collection rate). Therefore, we fit $f(t)$ using piecewise linear functions where the first segment is constant (up to year τ). The value of τ, and the number and position of knots after year τ, are selected by minimizing the AIC corrected for small samples. We also fit the model where $f(t)$ is constant for all t, in order to test whether a lag phase is justified. We classify a species as having a lag phase if the model chosen has at least one knot, and if the slope of $f(t)$ is positive after the first knot . Our model may be considered a special case of a generalized additive model (Wood 2006) using piecewise linear splines, a log-link function, and an offset $\log(N_t)$. R code for our analysis can be found at www.robjhyndman.com/lagphase and as an online supplement to this paper. Aikio et al.

(2010) adjusted the number of specimens recorded in a given year by using the quotient n_t/N_t prior to cumulation. This is analogous to what we have done, except that we have allowed for the count distribution and the changing variance by using a Poisson distribution, and we have estimated the lag phase without making clearly inappropriate assumptions about the data.

As the data form a time series, it is possible for some remaining autocorrelation to be present in the data. In our analysis, we tested that the residuals of the model were indistinguishable from white noise (via an ACF plot).

Without knowing the detailed habits of all collectors in each year, it is impossible to know whether to treat all zeroes as valid estimates, all as missing values, or as some intermediate combination. We treated all zeroes of the focal species in a year as valid observations. We then compared these results with models in which all zeroes were considered as missing values.

A somewhat similar approach to modelling lag phases was used by Aagaard and Lockwood (2014) who also used a piecewise linear approach with the first segment being constant. However, they did not account for the natural heterogeneity due to dealing with count data (which we have allowed for by using a Poisson distribution), and they allowed only two segments whereas our more flexible approach allows for non-linear growth (and even subsequent decline) in the period after the end of the lag phase. While we could use our method to extend discussion to the slopes, lengths and numbers of post-lag phases, in this paper we focus only on our estimate of the first knot, in order to make a direct comparison with the two previous applications of Aikio et al.'s (2010) method.

Results

Case studies: *Cakile maritima*

We estimated a significant lag phase ending in 1949, 52 years after the first herbarium record (Fig. 3). During this estimated lag phase, however, specimens were being collected from Western Australia, South Australia and Victoria, suggesting that its geographic extent was already increasing rapidly along the south coast, even though the collection rate remained very low. Treating all zeroes as missing values had little effect on the estimate of the end of the lag phase; in this case the estimate was 51 years after the first collection.

298 Case studies: *Holcus lanatus*

299 Aikio et al. (2010) estimated a lag of 91 years after the first record for this species on the
300 South Island. We found a similar estimate of 92 years (Fig. 4). However, if all years with
301 zero collections were treated as missing observations, then we found no significant lag phase.
302 In this case, there is insufficient data to accurately determine whether a lag phase exists and
303 how long it is. A period of over six decades with no collections meant that there was no
304 information on species abundance during the possible lag period.

305

306 Reanalysis of all species in the New Zealand data set

307 We found that the inclusion of a lag phase into our analysis was justified statistically in only
308 53 of 191 (28%) of cases (Fig. 5a). The mean lag time for these significant cases was 25
309 years (SE 2.3; range 3 to 92 years), with the mode of the distribution in the 0-20 year
310 histogram bin. Taking only those species with significant lags, there was a significant,
311 negative correlation between the year in which the first record occurred and the length of the
312 lag phase (r = -0.62, $p<0.00001$). We note that all of the instances of very long lag phases
313 reported by Aikio et al. (2010), and which were not significant in our analysis, had a single
314 collection followed by several decades of no records at all. Considering all zero values as
315 missing observations reduced (to 12%) the number of cases with significant lags, but with an
316 increased mean of 39 years (SE 2.9 range 13-57).

317

318 Reanalysis of the midwest US data

319 A lag phase was only justified in 102 (40%) of 257 of the time series. The mean for these
320 was 33 years (SE 2.2; range <2 to 86) after the first collection (Fig. 5b), again with a mode
321 less than 20 years. For these 102 cases there was a strong negative correlation (r= -0.70;
322 $p<0.001$) between the year of the first record and the length of the lag phase. Treating all
323 zeroes as missing values, the lag was significant in just 63 (25%) cases with a mean for those
324 cases of 37 years (SE 3.2; range <1 to97).

325 In no cases did we find any significant autocorrelation in the residuals of our models.

326

Discussion

Using a more statistically appropriate approach, we found far fewer significant lag phases than in previous analyses of the same data: 72% of the New Zealand species × island time series did *not* have a significant lag phase after the first herbarium record, compared with only 5% in Aikio et al.'s original analysis. Similarly, for the US data we concluded no significant lags in 60% of regions × species compared with Larkin's 23%. The percentages with significant lag phases were even lower if all zeros were treated as missing values. Of course, it cannot be inferred that the lags in all these non-significant cases were equal to zero, since Type II statistical errors could be high in small data sets and estimates can only be positive in value. The data sets that we re-analysed are often small, with long gaps, and thus the tests have low statistical power: any short (real) lags would be unlikely to be detected. This is not a fault with our method, but a statistical fact of life affecting any analysis of such data. However, a statistically significant lag can be biologically meaningless if based on a significant difference from zero or an AIC value justifying a more parsimonious model. For the US data, for example, we observed three cases that exhibited statistically significant lags but whose length was estimated at very much less than one year.

We must, however, conclude from our analyses that there is little evidence in these data to support an interpretation that lag phases are the norm in invasive species. It is possible that our method will identify a "phantom" lag in some cases where no true lag exists especially if the rate of expansion of the population is very slow (Supplementary Material). Like any use of statistical inference, our method is capable of making errors, although these are likely to be less frequent than with previous methods. However, the possibility of phantom lags will have little impact on our overall qualitative conclusions. Even with the addition of some false positives, our proportion of species with lags is still very much lower than in the previous studies. The high frequency of significant lags recorded by Aikio et al. (2010) and Larkin (2012) could have been because of invalid statistical assumptions, with cumulative data tending to give unrealistically low estimates of standard errors (Mesgaran et al. 2013) and too many instances where significant lags were concluded; it could also be because the parametric models that they used were inappropriate, even though they fitted the cumulative data quite well. For example, the von Bertalanffy model that fitted best in about 70% of Aikio et al.'s analyses and in 17%

of Larkin's analyses is difficult to justify on logical grounds, as is their second choice model, the logistic (Fig. 2).

While it is intuitively appealing to deal with cumulative data – and straightforward to find empirical models that appear to fit well – there are real dangers. We note that it has become common in invasion ecology to analyse cumulative data (e.g. Mikhulka and Pyšek 2001; Delisle et al. 2003; Fuentes et al. 2008). The error structures of such data must be recognised and appropriate actions taken for any statistical analysis to be valid. More generally, extreme caution must be used in the analysis of herbarium data. Even though we can allow mathematically for changes in general botanical collection effort over time (in our case using this as an offset in a log-linear model), this is only approximate and may even exacerbate errors in very old invasions – for example where the only data come from rare, intensive collecting by the few botanists of the time. It also does not account for changes in search effort for the focal species in response to awareness of its spread. Increased search effort will inevitably result in a greater collection rate for that species, perhaps considerably (Cousens and Mortimer 1995), even if total rate of collection of species as a whole barely alters. Herbarium collections are irregular in both time and space and are only a very crude indicator of abundance or area invaded. To treat them as if they were equivalent to a formal survey could easily lead to erroneous inferences. We should therefore not expect too much from any calculations made from the increasingly available herbarium databases: they, too, will be crude.

The significant negative correlation of lag length with the first year of collection in both data sets is almost certainly also an artefact. As concluded by Larkin, there is insufficient time for long lag phases to have ended within recent, shorter runs of data. There is also insufficient time to have detected the end of long lag phases, unless they concluded some years ago. Consequently, the longest lag phases must be associated with early years of collecting. Further, poor sampling frequency in the earliest years of botanical collecting means there are large gaps in the data, which may result in the best model having a lag phase simply because there are no intermediate data to demonstrate systematic lack of fit. The longest lag estimates for the New Zealand data obtained by the original authors were all first recorded in the 19th century: the early decades after first discovery in these particular time series were characterised by long periods without any collection of a given species. While long periods with no collection are to be expected for very restricted species in a long lag phase, the outcome is that we have no data to test for systematic departure from a constant collection rate during that period. It is also possible

that the first occurrence may have died out or been removed and the next occurrence was a new invasion.

While we have argued that our statistical approach is less problematic than the approach adopted by Aikio et al (2010) and Larkin (2012), we do not consider it a panacea for modelling herbarium data. No statistical method can deal with the fundamental underlying problem that herbarium data are subject to the changing policies, choices and values of collectors over the years, and do not constitute a structured survey of species abundance. Like all previous analyses of these data, we have had to make some statistical assumptions in order to carry out our statistical analyses. We believe we have made fewer unreasonable assumptions than previous analyses.

Although the focus on our paper has been the statistical assumptions of the analytical methods, the interpretation of change points in the collection rate of a species as the end of an invasion lag phase and the time from first collection to this change point as the length of the lag phase require considerable leaps of faith. As discussed in the Introduction, many other factors can lead to a change in collection rate for an individual species. In the absence of any additional information, Aikio et al.'s (2010) method – with or without our modifications – has no option but to ignore all other plausible causes of change in collection rate. What proportion, then, of our reduced number of significant results are not true changes in invasion rate, even though they may be true changes in specimen collection rate? For that matter, what proportion of the non-significant results in fact hide true lags? The actual date of introduction will always be underestimated (except in the exceptional case of an invasion being found the same year in which it occurs), perhaps by decades, and this bias may easily equal or exceed the lag estimated from herbarium samples (cf. Aikio et al. 2010).

Acknowledgments We thank Sami Aikio and Ines Schonberger for supplying data from the Allan Herbarium (CHR), Dan Larkin for his mid-west USA data and Alison Vaughan, National Herbarium of Victoria (MEL), for supplying data from Australia's Virtual Herbarium. We also thank Rod Randall for supplying a list of invasive species for Australia. We appreciate comments on a previous version of this paper by Richard Duncan, Sami Aikio and Dan Larkin, although we have only incorporated some of their suggestions.

Conflict of Interest: The authors declare that they have no conflict of interest.

426

427

428 **References**

429 Aagaard K, Lockwood J (2014) Exotic birds show lags in population growth. Divers Distrib
430 20:547–554.

431 Aikio S, Duncan RP, Hulme PE (2010) Lag-phases in alien plant invasions: separating the
432 facts from the artefacts. Oikos 119:370-378.

433 Booth, BD, Murphy SD, Swanton CJ (2011) Invasive plant ecology in natural and
434 agricultural systems (2nd edition). CABI, Cambridge, Massachusetts, USA.

435 Cousens R, Mortimer M (1995) Dynamics of weed populations. Cambridge University Press,
436 Cambridge, UK.

437 Cousens R, Ades PK, Mesgaran MB, Ohadi S (2013) Reassessment of the invasion history
438 of two species of Cakile (Brassicaceae) in Australia. Cunninghamia 13:275-290.

439 Crooks JA, Soulé ME (1999) Lag times in population explosions of invasive species: causes
440 and implications. In: Sandlund OT, Schei PJ, Viken A (eds), Invasive species and
441 biodiversity management. Kluwer, pp. 103-125.

442 Daehler CC (2009) Short lag times for invasive tropical plants: evidence from experimental
443 plantings in Hawai'i. PLoS ONE 4(2):e4462.

444 Davidson R, MacKinnon JG (2004) Econometric theory and methods. Oxford University
445 Press, Oxford, UK.

446 Delisle F, Lavoie MJ, Lachance D (2003) Reconstructing the spread of invasive plants:
447 taking into account biases associated with herbarium specimens. J Biogeog 30:1033-
448 1042.

449 Dobson AJ (2008) An introduction to generalized linear models, 3r^{d} ed. Chapman and
450 Hall/CRC Press, Boca Raton, Florida, USA.

451 Dogra KS, Sood SK, Dobhal PK, Sharma S (2010) Alien plant invasion and their impact on
452 indigenous species diversity at global scale: a review. J Ecol Nat Environ 2:175-186.

453 Ellstrand NC, Schierenbeck KA (2000) Hybridization as a stimulus for the evolution of
454 invasiveness in plants? Proc Natl Acad Sci USA 97:7043–7050.

455 Fuentes N, Ugarte E, Kuhn I, Klotz S (2008) Alien plants in Chile: inferring invasion periods
456 from herbarium records. Biol Invas 10:649-657.

457 Gallagher RV, Beaumont LJ, Hughes L, Leishman MR (2010) Evidence for climatic niche
458 and biome shifts between native and novel ranges in plant species introduced to
459 Australia. J Ecol 98:790–799.

460 Hobbs RJ, Humphries SE (1995) An integrated approach to the ecology and management of
461 plant invasions. Conserv Biol 9: 761-770.

462 Holt RD, Barfield M, Gomlkiewicz R (2005) Theories of niche conservatism and evolution:
463 could exotic species be potential tests? In: Sax DF, Stachowicz JJ, Gaines SD (eds),
464 Species invasions: Insights into ecology, evolution, and biogeography. Sinauer,
465 Sunderland, Massachusetts, USA pp. 259–290.

466 Keller SR, Taylor DR (2008) History, chance and adaptation during biological invasion:
467 separating stochastic phenotypic evolution from response to selection. Ecol Lett 11:852–
468 866.

469 Kot M, Lewis MA, van den Driessche (1996) Dispersal data and the spread of invading
470 organisms. Ecology 77:2027-2042.

471 Kowarik I (1995) Time lags in biological invasions with regard to the success and failure of
472 alien species. In: Pyšek P, Prach K, Rejmánek M, Wade M (eds), Plant invasions –
473 General aspects and special problems. SPB Academic Publishing, Amsterdam, The
474 Netherlands pp. 15-38.

475 Larkin DJ (2012) Lengths and correlates of lag phases in upper-Midwest plant invasions. Biol
476 Invasions 14:827-838.

477 Lee CE (2002) Evolutionary genetics of invasive species. Trends Ecol Evol 17:386–391.

478 Lenda M, Skórka P, Knops JMH, Moroń D, Tworek S, Woyciechowski M (2012) Plant
479 establishment and invasions: an increase in a seed disperser combined with land
480 abandonment causes an invasion of the non-native walnut in Europe. Proc R Soc B
481 279:1491-1497.

Mack RN, Simberloff D, Lonsdale WM, Evans H, Clout M, Bazzaz FA (2000) Biotic invasions: causes, epidemiology, global consequences, and control. Ecol Appl 10:689-710.

Mesgaran MB, Mashhadi HR, Alizadeh H, Hunt JR, Young KR, Cousens RD (2013) Importance of frequency distribution selection in hydrotime and hydrothermal time models of seed germination. Weed Res 53:89-101.

Mihulka S, Pyšek P (2001) Invasion history of Oenothera congeners in Europe: a comparative study of spreading rates in the last 200 years. J Biogeog 28:597-609.

Parker IM (2004) Mating patterns and rates of biological invasion. Proc Natl Acad Sci USA 101:13695-13696.

Pyšek P, Hulme PE (2005) Spatio-temporal dynamics of plant invasions: Linking patterns to process. Écoscience 12:302-315.

Pyšek P, Prach K (1993) Plant invasions and the role of riparian habitats: a comparison of four species alien to central Europe. J. Biogeogr. 20:413-420.

Rodman JE (1986) Introduction, establishment and replacement of sea-rockets (Cakile, Cruciferae) in Australia. J Biogeogr 13:159-171.

Sakai AK, Allendorf FW, Holt JS, Lodge DM, Molofsky J, With KA, Baughman S, Cabin RJ, Cohen JE, Ellstrand NC, McCauley DE, O'Neil P, Parker IM, Thompson JN, Weller SG (2001) The population biology of invasive species. Annu Rev Ecol Syst 32:305-332.

The Council of Heads of Australasian Herbaria (1999) Australia's Virtual Herbarium www.chah.gov.au/avh [Accessed 19 April 2013]

Wangen SR, Webster CR (2006) Potential for multiple lag phases during biotic invasions: reconstructing an invasion of an exotic tree *Acer platanoides*. J Appl Ecol 43:258-268.

Webber, B. L. Yates CJ, Le Maitre DC, Scott JK, Kriticos DJ, Ota N, McNeill A, Le Roux JJ, Midgley GF (2011) Modelling horses for novel climate courses: insights from projecting potential distributions of native and alien Australian acacias with correlative and mechanistic models. Divers Distrib 17:978-1000.

511 Wells MJ (1974) *Nassella trichotoma* (Nees) Hack. in South Africa. Proc 1st Nat Weeds
512 Conf South Africa, pp. 125-137

513 Williamson M, Pyšek, P., Jarosík, V. & Prach, K. 2005. On the rates and patterns of spread
514 of alien plants in the Czech Republic, Britain, and Ireland. Écoscience 12:424-433.

515 Wood SN (2006). Generalized additive models: an introduction with R. Chapman and
516 Hall/CRC Press, Boca Raton, Florida, USA.

517

518

519

520 **Legends to Figures**

521 **Fig. 1** Changes in herbarium sample collection over time, based on the combined collections
522 of the major Australian herbaria. Dotted line shows the number of collectors in each year;
523 solid line shows the number of samples – aliens and natives - that they collected. The steep
524 drop in recent years is partly due to the time-lag in processing samples and adding them to
525 databases.

526 **Fig. 2** Alternative models for sample collection rate as a function of time (after having
527 adjusted for temporal changes in collection effort of all alien plants). a-d are models fitted by
528 Aikio et al. (2010). The centre column shows the basic models that Aikio et al. formulated
529 for collection rate as a function of cumulative number of samples; the right hand column
530 shows the integrated form of those models, relating cumulative sample number to time,
531 which they then fitted to data; the left hand column is the relationship between rate of
532 collection (samples per year) and time that emerges from each model. Red lines show the lag
533 phase; blue lines show the post-lag period during which collection rate increased. For the
534 right hand column, the increase phase is (a) the von Bertalanffy model, (b) linear model, (c)
535 logistic model, (d) exponential model. In (a) and (c) K is the maximum number of samples
536 that will ever be collected and r is the maximum rate of collection (an hypothetical,
537 extrapolated value in the case where a linear lag phase is combined with the von Bertalanffy).
538 (e) shows what would be expected if population size increases in a sigmoidal fashion after the
539 end of the lag phase (i.e. the pattern accepted by most invasion biologists for non-lag
540 populations, e.g. Hobbs and Humphries, 1994) and collection rate is directly proportional to
541 population size: note that there would be no sudden jump in collection rate (left column) and
542 a linear rather than asymptotic final trajectory for cumulative sample number (right column).

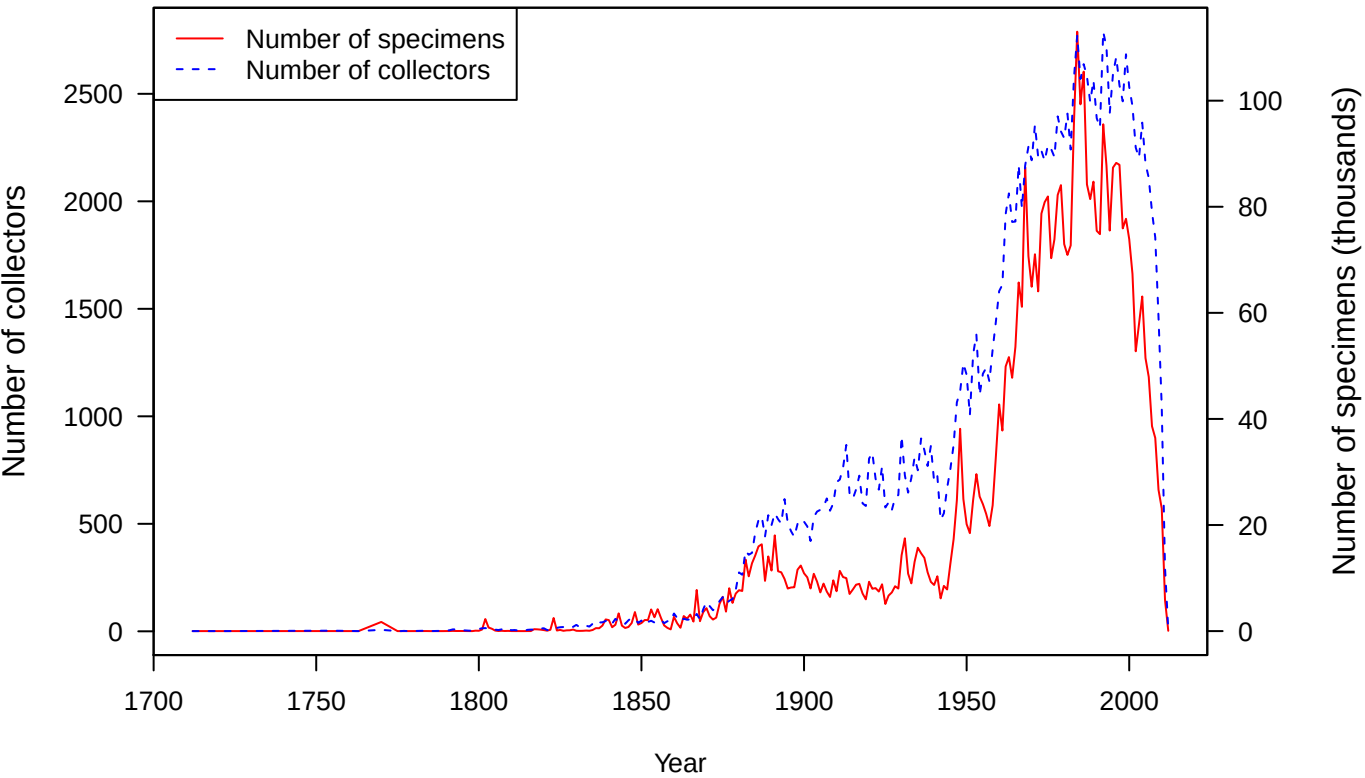
543 **Fig. 3** Analysis of *Cakile maritima* samples collected in Australia. (a) Annual collection
544 rates (i.e., numbers of samples of *C. maritima* collected per year) are shown as open symbols,
545 and the line is the mean collection rate for each year given by $N_t \exp[f(t)]$. (b) Collection rate
546 adjusted by collection effort for all invasive species. The adjusted collection rate (shown as a
547 solid line) is $\exp[f(t)]\ \bar{N}$, where $\bar{N}$ is the average number of samples of alien species collected
548 per year. This is equal to the estimated mean collection rate in each year normalized by the
549 total number of alien species collected in each year. The estimated knots are breakpoints of
550 the line (the first being the statistically significant lag phase ending in 1949), confidence
551 intervals are shown using grey shading, and years in which collections were >0 are shown as

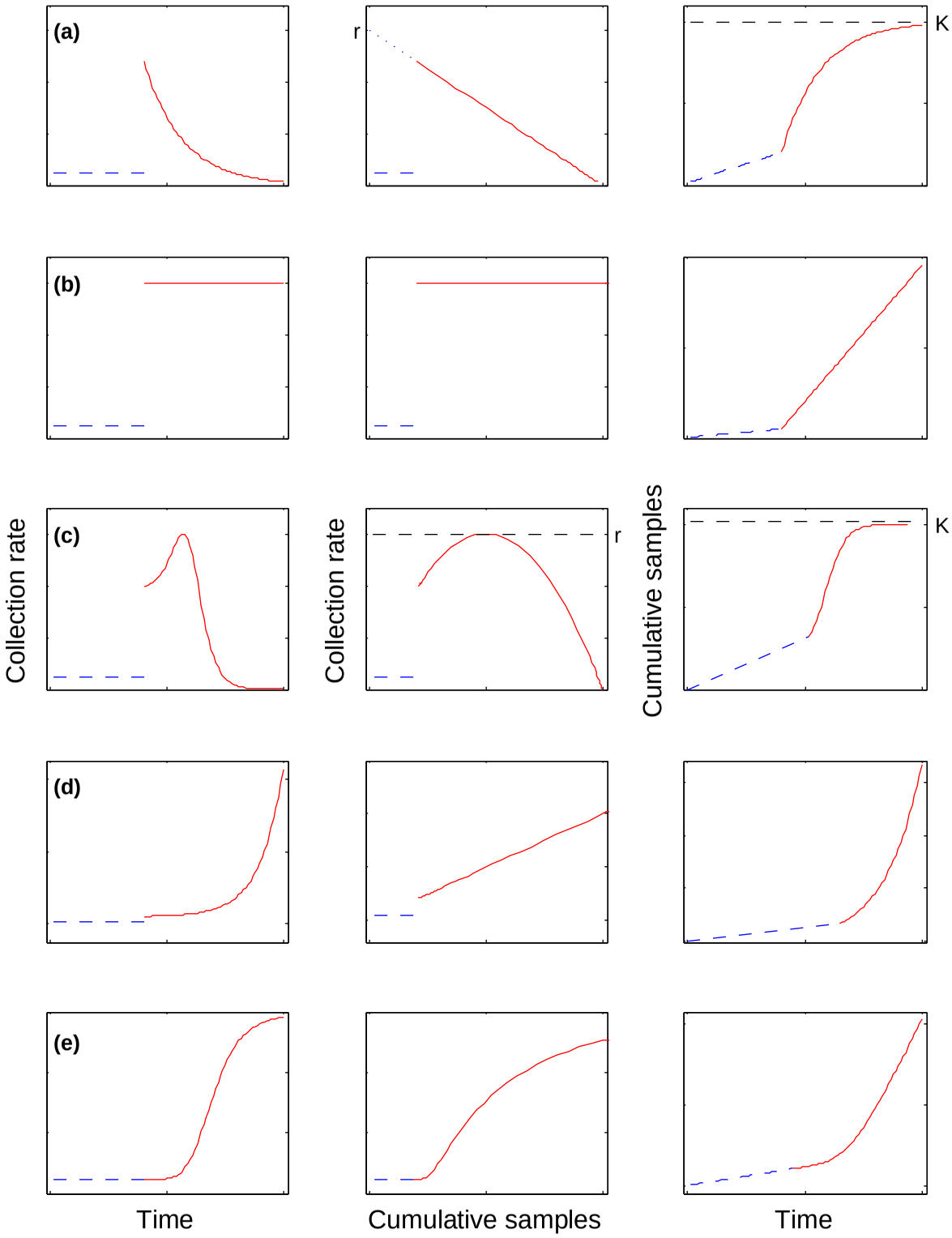
552 dashes on the horizontal axis. (c) Cumulative number of records, shown only in years when
553 there were new collections.

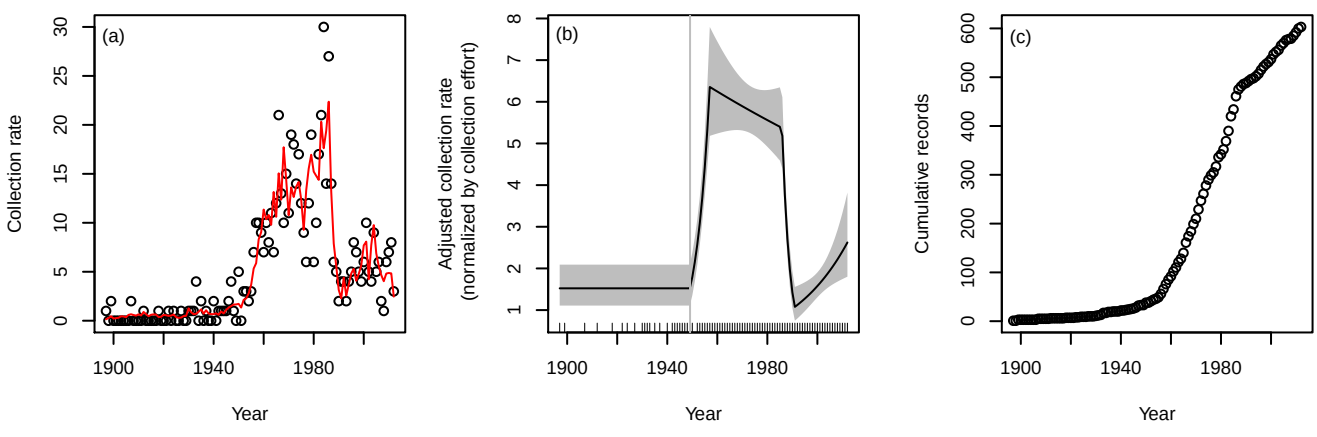
554 **Fig. 4** Analysis of *Holcus lanatus* data from the South Island of New Zealand (from the
555 Allan Herbarium, CHR). Graphs are equivalent to Fig. 3, with the lag phase ending in 1964.

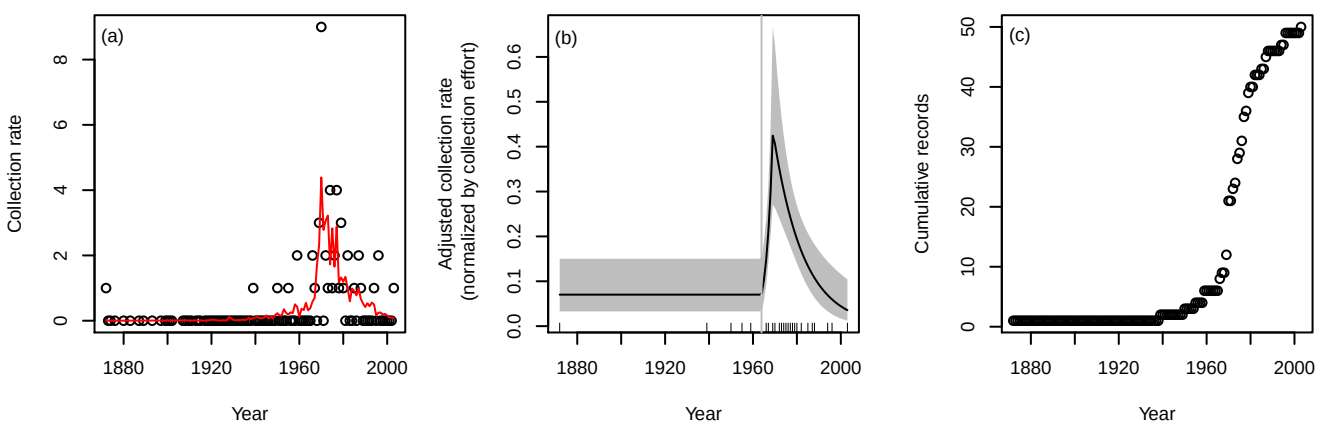
556 **Fig. 5** Frequency distributions of the lengths of significant lags in (a) New Zealand and (b)
557 mid-west USA herbarium data.

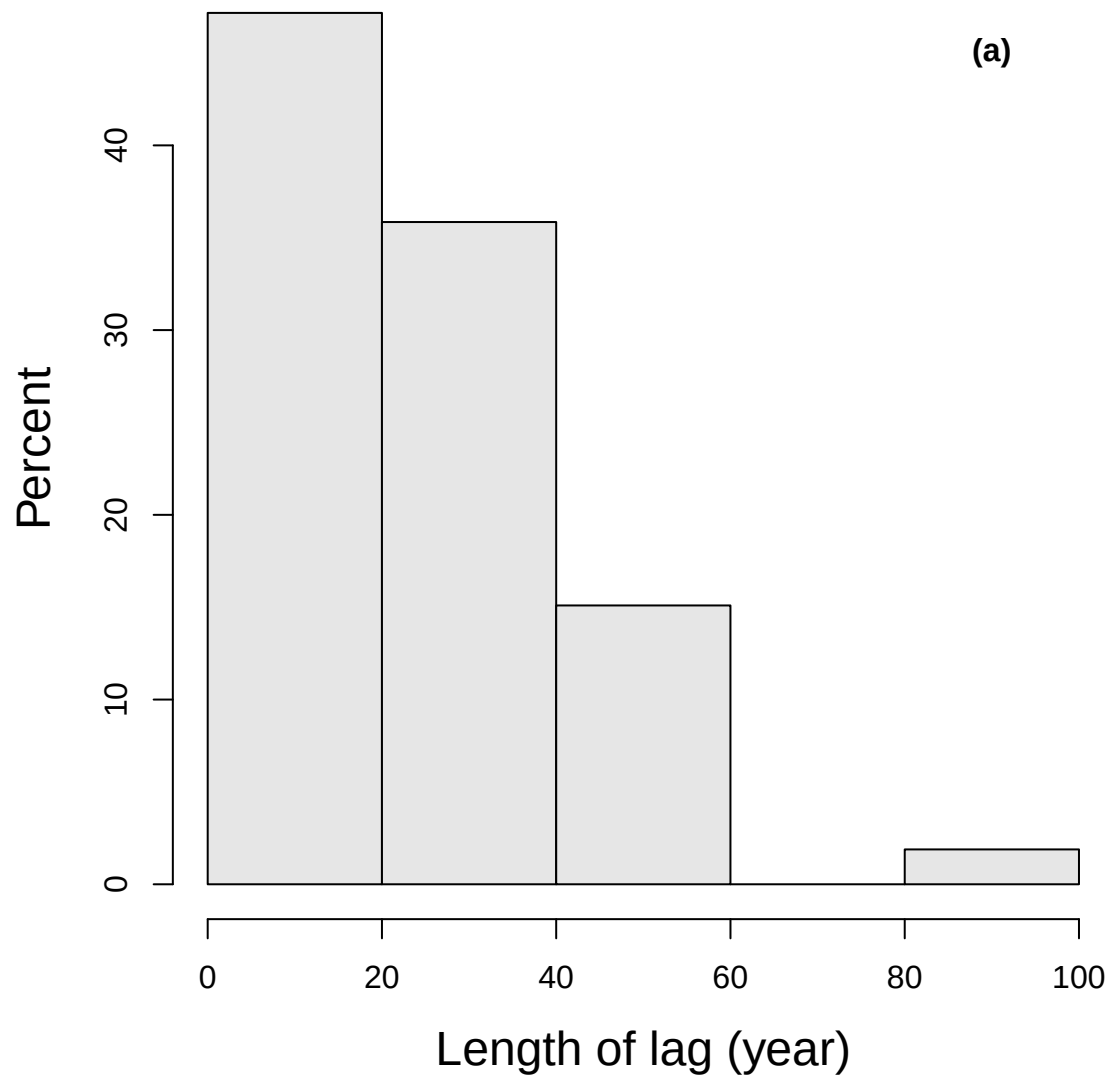
558

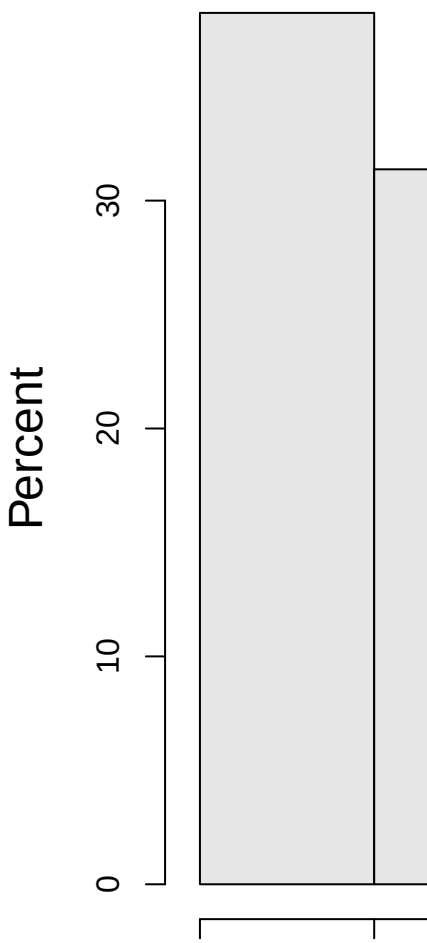

Number of specimens
Number of collectors
Number of collectors
Number of specimens (thousands)
Year
0
500
1000
1500
2000
2500
0
20
40
60
80
100
1700
1750
1800
1850
1900
1950
2000


(a)
r
K
(b)
(c)
r
K
(d)
(e)
Collection rate
Collection rate
Cumulative samples
Time
Cumulative samples
Time


(a)
Collection rate
0
5
10
15
20
25
30
1900
1940
1980
Year
(b)
Adjusted collection rate
(normalized by collection effort)
1
2
3
4
5
6
7
8
1900
1940
1980
Year
(c)
Cumulative records
0
100
200
300
400
500
600
1900
1940
1980
Year


(a)
Collection rate
0
2
4
6
8
1880
1920
1960
2000
Year
(b)
Adjusted collection rate
(normalized by collection effort)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1880
1920
1960
2000
Year
(c)
Cumulative records
0
10
20
30
40
50
1880
1920
1960
2000
Year


(a)
Percent
0
10
20
30
40
0
20
40
60
80
100
Length of lag (year)


Percent
0
10
20
30
0
20