# “Facts” may still be artefacts, since models can make unrealistic assumptions: statistical methods for the estimation of invasion lag-phases from herbarium data

Rob J. Hyndman[1], Mohsen B. Mesgaran[2] and Roger D. Cousens[2]

*[1]Department of Econometrics & Business Statistics, Monash University, VIC 3800, Australia (Rob.Hyndman@monash.edu) [2]Department of Resource Management & Geography, The University of Melbourne, Victoria 3010, Australia*

**Abstract**

We present a new method for estimating the length of the invasion lag phase from simple time series of counts of herbarium records. This is 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. 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.

**Synthesis**

It is a widely held view that invasive species commonly undergo a lag phase prior to a more rapid phase of expansion. This conclusion, however, is not based on a systematic analysis of data. A previous method of estimating the length of the lag phase from herbarium data, based on unrealistic statistical and model assumptions, concluded that lags occurred in the majority of species. Our new method overcame these methodological problems and led to the conclusion that statistically significant lags only occurred in a much smaller proportion of the cases previously analysed. While statistics can help to separate facts from artefacts, the reality is that any method – including ours - relies on the validity of its assumptions: inappropriate assumptions can themselves introduce artefacts. Despite the temptations introduced by the easy accessibility of herbarium databases, herbarium data are mostly inadequate for the estimation of the lengths of invasion lag phases.

It is an almost universal belief amongst invasion ecologists that there is often a lag phase in the dynamics of an invasion. During this phase, a species undergoes only slow or negligible range expansion and population growth; a period of more rapid expansion then follows. Discussions of invasion dynamics vary in their interpretation of the ubiquity of the phenomenon, from "some species" (Frappier et al. 2003), "often/in many instances" (Mack et al. 2000, Lenda et al. 2011), "commonly/generally" (Parker 2004, Pysek 2005), to "most" (Radosevich 2007) or all (implied by the use of the term "there *is* a lag phase": Dogra et al. 2010).

There are many plausible reasons why a lag phase might occur. Given that rate of population spread is the product of dispersal and population growth rate (Kot et al. 1996), the change from a low or negligible rate of increase to a much greater one must necessarily involve a change in either, or both, of these factors. Causes postulated in the literature include:

- A threshold period before the founders are physiologically able to reproduce;
- A threshold population size below which pollination/reproduction is inefficient;
- Abiotic habitat change (e.g. climate change, altered land management, novel elements introduced into landscape, extreme events such as fire or tornado);
- Biotic habitat change (e.g. loss of herbivore, competitor or predator; introduction of pollinators or mycorrhizae);
- Long period required to evolve genotypes more suited to alien habitats (due to genetic bottleneck);
- Chance introduction of additional genotypes that are either more fit or allow novel, fitter recombinations to be selected;
- Introduction of the opposite sex (in dioecious species);
- Dispersal across a localised barrier, around the location of introduction, into a larger region of suitable habitat (perhaps because population size had to build up to achieve significant propagule pressure);
- After a slow initial spread because of poor dispersal, reaching a location in which additional, long-distance dispersal vectors become available.

Direct evidence linking a particular invasion to a given mechanism (e.g. Wangen and Webster 2006) is, however, rare.

Some have argued that the early part of an exponential increase in population extent/abundance is, in essence, another type of lag phase (an "inherent" lag phase, Crooks and Soulé 1999) because it is a period where the species is still highly restricted and the rate of invasion increases as population size grows. Indeed, authors sometimes draw arrows on graphs where the exponential is clearly an excellent fit, indicating the putative end of a lag phase (Cousens and Mortimer 1995). Mathematically, this is misleading: the absolute rate of increase grows steadily, rather than step-wise, and the per capita rate of increase of an exponential is constant. Thus it can be argued mathematically that there is no lag phase if population growth is exponential: it is merely an optical illusion.

There are highly documented instances in which the limited data are persuasive in arguing for the existence of a prolonged lag phase before the main invasion occurred. This has led to an interest in measuring the lengths of lags. Where there is documentary evidence for the date of first deliberate planting, it is common to take the lag phase as being the time between then and when it is first reported to have established away from the origin (Kowarik 1995). Using this method, lag phases appear to be common and lengthy (a mean of 147 years for woody species in one study: Kowarik 1995). The results from another study, with a finer scale of observation than most others, led to the hypothesis that lags are inherently shorter in the tropics (Daehler 2009). Such estimates are, however, always problematic: the data are censored, as we only observe that the lag has ended after the event has occurred (Daehler 2009), and often we do not know the date of introduction, only the date of first observation, which may be decades later.

This also highlights a more general problem, that data during the lag phase (if indeed one exists) are either absent or extremely sparse. 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 between a true lag phase and the early parts of a rapid (e.g. exponential) 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 2010 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, inflating the apparent rate of increase and introducing phase changes where they may not exist.

Notwithstanding these potential problems, ecologists would still like to determine, in a given case, whether there is evidence that a lag phase exists and, if so, to estimate its length with as much accuracy as possible. Herbarium data, increasingly available in digital format, allow us to attempt to reconstruct invasion histories (e.g. Rodman,1985; Cousens et al. 2013), to determine periods of species influx (e.g. Fuentes et al. 2008), eventual invasion extents (using species distribution models based on climate: e.g. Webber et al. 2011) and invasion niche shifts (Gallagher et al. 2010). Aikio et al. (2010) estimated the end of the lag phase as the point of intersection between two regression lines relating cumulative number of herbarium specimens and time (the first use of this method appears to be by Pyšek and Prach 1993, although few details are given). Thus, rate of increase of a species in herbaria was used as a surrogate for abundance and/or the size of the invaded area. Prior to cumulation, the number of specimens was detrended for temporal variation in overall invasive species collection effort (assumed to be proportional to the number of herbarium samples), which would otherwise bias inferences about invasion rates. The method was repeated by Larkin (2012): both 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 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. 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).

There are, however, both 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). Aikio et al. (2010), unlike Larkin (2012), made no attempt to estimate the confidence intervals in their lag time estimates, though both studies compared the fit of models with/without lag phases statistically using the Akaike Information Criterion (AIC). Again, 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 exotic species collected for that year, this could have arisen because no effort was put into collecting exotic 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 exotics 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 sampling frequency may exaggerate the length of the lag phase.

Second, the equations that they fit, though appearing to describe the cumulative data reasonably well, make some unlikely assumptions if the data are being used to measure the progress of the invasion. The four models fitted by Aikio et al. (2010) all assume that there is a sudden jump in specimen collection rate (even after adjustment for changes in collection effort) immediately the lag phase ends (Fig 2a-d). If collection rate is simply proportional to the abundance or distribution of the species (a fundamental assumption in this use of herbarium data to estimate lag phases), the rate of collection 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 a very 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 once a species reaches its greatest abundance, 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, however, are often collected as part of plant community or regional surveys, so that it is unlikely that collection would ever stop completely.

The aim of this paper is to propose a statistical method, based on annual collection rates rather than cumulative specimen number, making more appropriate assumptions about the data. The method also avoids the assumptions implicit within the regression models used previously. We apply our method to an example taken from a database representing the combined collections of the major herbaria in Australia. We then compare our results with previous studies of smaller data sets.

**Material and Methods**

*Data.* We illustrate our method in detail using two species: *Cakile maritima* in Australia, a coastal species that we have studied closely (Cousens et al. 2013) and represented by 603 specimens; and *Holcus lanatus* in the South Island of New Zealand, a species estimated to have a long lag phase by Aikio et al. (2010) but represented by only 50 specimens. We then reanalysed the entire data sets used by Aikio et al. (2010) and Larkin (2012) for the north and south islands of New Zealand and the Midwest US respectively. For *C. maritima*, we accessed data from the Australian Virtual Herbarium (AVH), a database combining the records of all major Australian herbaria. Total number of invasive specimens collected per year was 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. 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 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 collected in each year ($n_t$), as a measure of the collection rate. As the abundance or distribution of a species increases, so will the probability of a sample being collected. We assume that $n_t$ is Poisson distributed. The use of herbarium records as a surrogate for abundance or distribution relies on an assumption of constant sampling 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 invasives), changes in collection effort for the focal species, and other factors together affect the probability of a species sample being collected. We therefore used the total number of samples of alien species collected in that year ($N_t$) as a term in the following model:

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

where $f(t)$ allows 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). It would be a simple matter to allow over-dispersion in this model in the usual way if required. A lag-phase exists if $f(t)$ is constant until some year $\tau$, and then takes some other higher values thereafter (Aikio et al. 2010). Therefore, we fit $f(t)$ using piecewise linear splines where the first segment is constant (up to year $\tau$). The value of $\tau$, and the number and position of knots after year $\tau$, 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 . 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, and we have estimated the lag phase without making clearly inappropriate assumptions about the data.

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 compared these results with models in which all zeroes were considered as missing values.

**Results**

*Case studies: Cakile maritima*. We estimated a significant lag phase ending in 1949, 52 years after the first 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 frequency 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.

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

*Reanalysis of all species in the New Zealand data set*. We found that the inclusion of a lag phase into our analysis was justified statistically in only 53 of 191 (28%) of cases (Fig. 5a). The mean lag time for these significant cases was 25 years (SE 2.3; range 3 to 92 years), with the mode of the distribution in the 0-20 year histogram bin. Taking only those species with significant lags, there was a significant, negative correlation between the year in which the first record occurred and the length of the lag phase ($r$ = -0.62, $p<0.000001$). We note that all of the instances of very long lag phases reported by Aikio et al. (2010), and which were not significant in our analysis, had a single collection followed by several decades of no records at all. Considering all zero values as missing observations reduced (to 12%) the number of cases with significant lags, but with an increased mean of 39 years (SE 2.9 range 13-57).

*Reanalysis of Midwest US data.* A lag phase was only justified in 102 (40%) of 257 of the time series. The mean for these was 33 years (SE 2.2; range <2 to 86) after the first

collection (Fig. 5b), again with a mode less than 20 years. For these 102 cases there was a strong negative correlation ($r = -0.70$; $p<0.001$) between the year of the first record and the length of the lag phase. Treating all zeros as missing values, the lag was significant in just 63 (25%) cases with a mean for those cases of 37 years (SE 3.2; range <1 to97).

**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 x island time series did *not* have a significant lag phase, 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 x 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. 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. 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); 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 rate over time (in our case using this as an offset), 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 notable 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.

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 early years when there were few collectors to cover very large areas. For example, on the first Australian specimen of *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). Thus an estimate of a 20 year lag phase determined from herbarium samples could easily be only half the true value. How useful, then, is such an estimate?

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 19$^{th}$

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 assumptions in order to carry out our statistical analyses. We believe we have made fewer unreasonable assumptions than previous analyses, but we cannot take account of collector behaviour, and so our results may also be artefacts of the assumptions made.

**Acknowledgements**

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.

**References**

Aikio, S. et. al. 2010. Lag-phases in alien plant invasions: separating the facts from the artefacts. - Oikos 119: 370-378.

Crooks, J. A. and Soulé, M. E. 1999. Lag times in population explosions of invasive species: causes and implications. - In: Sandlund, O. T. et al. (eds), Invasive species and biodiversity management. Kluwer, pp. 103-125.

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

Cousens, R. and Mortimer, M. 1995. Dynamics of weed populations. - Cambridge University Press.

Cousens, R. et al. 2013. Reassessment of the invasion history of two species of Cakile (Brassicaceae) in Australia. - Cunninghamia 13: 275-290.

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

Delisle, F. et al. 2003. Reconstructing the spread of invasive plants: taking into account biases associated with herbarium specimens. - J. Biogeog. 30: 1033-1042.

Dobson, A.J. 2008. An introduction to generalized linear models, 3r$^{d}$ ed. - Chapman and Hall/CRC.

Dogra, K. S. et al. 2010. Alien plant invasion and their impact on indigenous species diversity at global scale: a review. J. Ecol. Nat. Environ. 2: 175-186.

Frappier, B. et al. 2003. Small-scale invasion pattern, spread rate, and lag-phase behavior of *Rhamnus frangula* L. – Forest Ecol. Manag. 186: 1-6.

Fuentes. N. et al. 2008. Alien plants in Chile: inferring invasion periods from herbarium records. – Biol. Invas. 10: 649-657.

Hobbs, R. J. and Humphries, S. E. 1995. An integrated approach to the ecology and management of plant invasions. - Conserv. Biol. 9: 761-770.

Kot, M. et al. 1996. Dispersal data and the spread of invading organisms. - Ecology 77: 2027-2042.

Kowarik, I. 1995. Time lags in biological invasions with regard to the success and failure of alien species. - In: Pyšek, P. et al. (eds), Plant invasions – General aspects and special problems. SPB Academic Publishing, pp. 15-38.

Larkin, D. J. 2012. Lengths and correlates of lag phases in upper-Midwest plant invasions. – Biol. Invasions 14: 827-838.

Lenda, M. et al. 2011. Plant establishment and invasions: an increase in a seed disperser combined with land abandonment causes an invasion of the non-native walnut in Europe. - Proc. R. Soc. B 279: 1491-1497.

Mack, R. N. et al. 2000. Biotic invasions: causes, epidemiology, global consequences, and control. - Ecol. Appl. 10: 689-710.

Mesgaran, M. B. et al. 2013. Importance of frequency distribution selection in hydrotime and hydrothermal time models of seed germination. - Weed Res. 53: 89-101.

Mihulka, S. and 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, I. M. 2004. Mating patterns and rates of biological invasion. - PNAS 101: 13695-13696.

Pyšek, P. and 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.

Radosevich, S. R. 2007. Plant invasions and their management. Chapter 3 - In: CIPM (ed.), Invasive Plant Management. Center for Invasive Plant Management. http://www.weedcenter.org/textbook/index.html (accessed 18 February 2014).

Wangen, S. R. and Webster, C. R. 2006. Potential for multiple lag phases during biotic invasions: reconstructing an invasion of an exotic tree *Acer platanoides*. - J. Appl. Ecol. 43: 258-268.

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

Wood, S.N. 2006. Generalized additive models: an introduction with R. - Chapman and Hall/CRC.

**Legends to Figures**

Fig. 1. Changes in herbarium sample collection over time, based on the combined collections of the major Australian herbaria. Dotted line shows the number of collectors in each year; solid line shows the number of samples that they collected.

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

Fig. 3. Analysis of Australian *Cakile maritima* samples collected in Australia. (a) number of samples collected per year (open symbols), and the fitted model (line, equation 1); (b) Frequency of collections adjusted by collection rate of all invasive species, showing estimated knots (the first being the statistically significant lag phase ending in 1908), confidence intervals (grey shading) and years in which collections were >0 (dashes on x-axis); (c) cumulative number of records, shown only in years when there were new collections.

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

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

Fig. 1

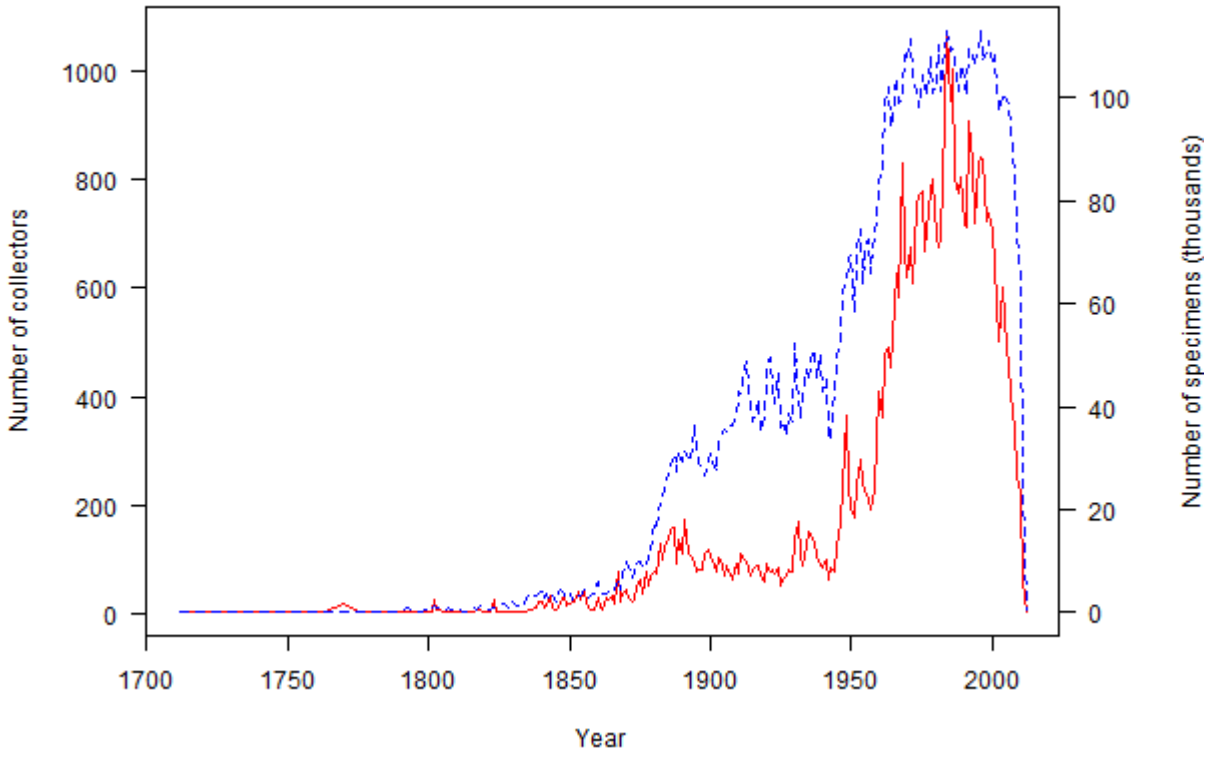

Number of collectors
Number of specimens (thousands)
Year
0
200
400
600
800
1000
0
20
40
60
80
100
1700
1750
1800
1850
1900
1950
2000

Fig. 2.

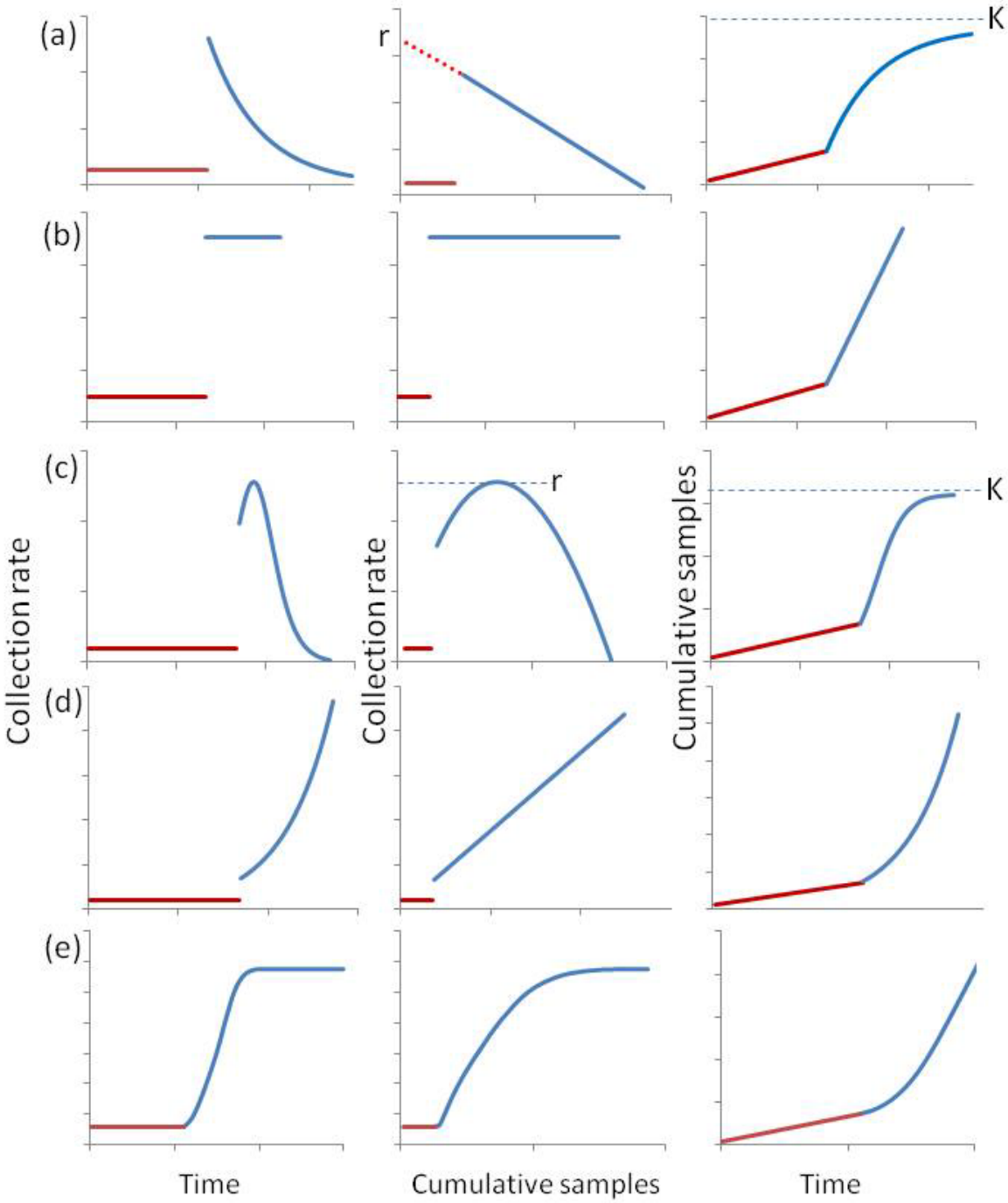

(a)
r
K
(b)
(c)
r
K
Collection rate
Collection rate
Cumulative samples
(d)
(e)
Time
Cumulative samples
Time

Fig 3.

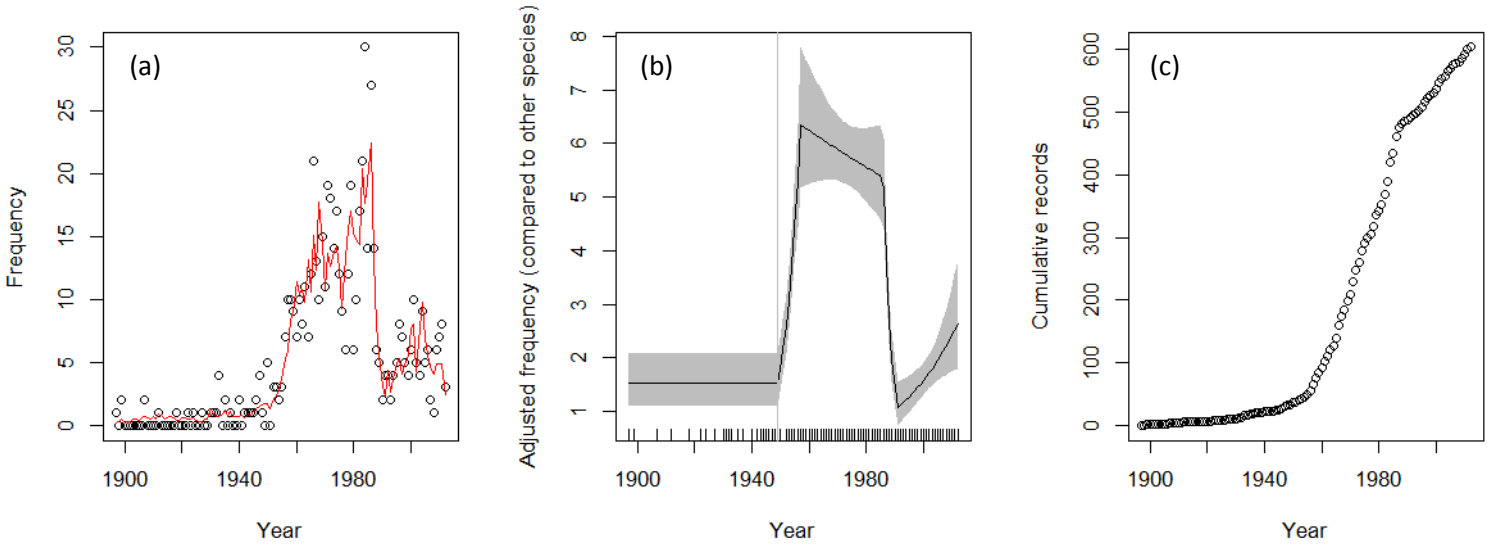

(a)
Frequency
0
5
10
15
20
25
30
1900
1940
1980
Year
(b)
Adjusted frequency (compared to other species)
1
2
3
4
5
6
7
8
1900
1940
1980
Year
(c)
Cumulative records
0
100
200
300
400
500
600
1900
1940
1980
Year

Fig 4.

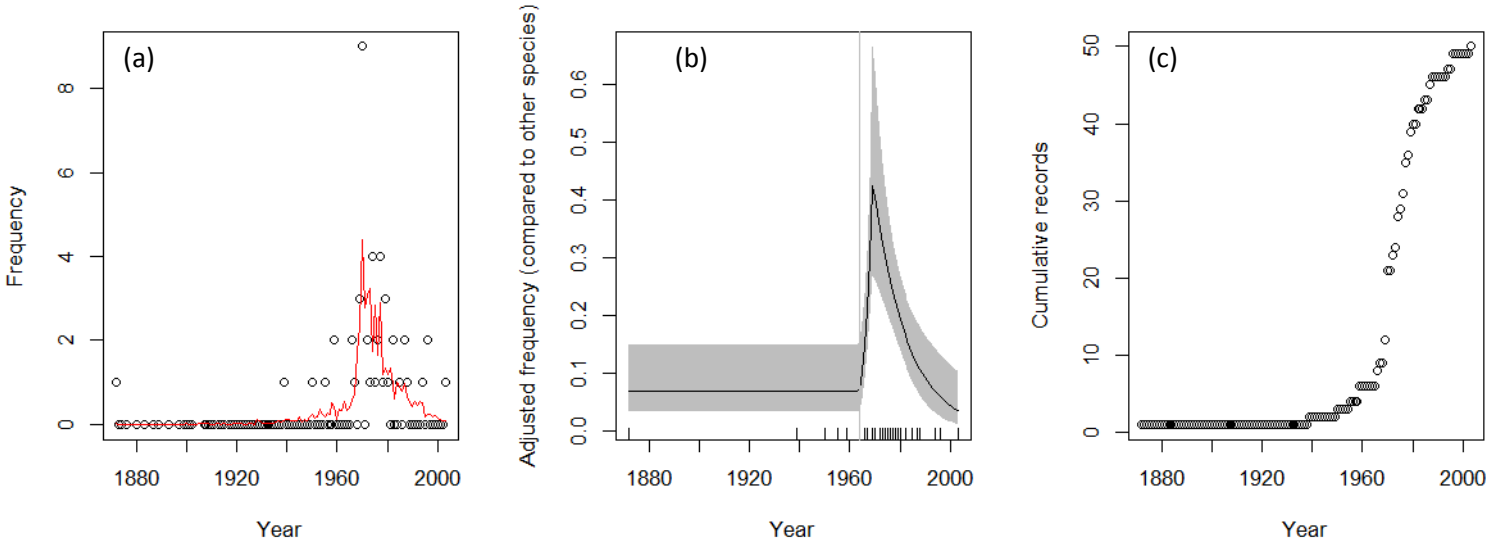
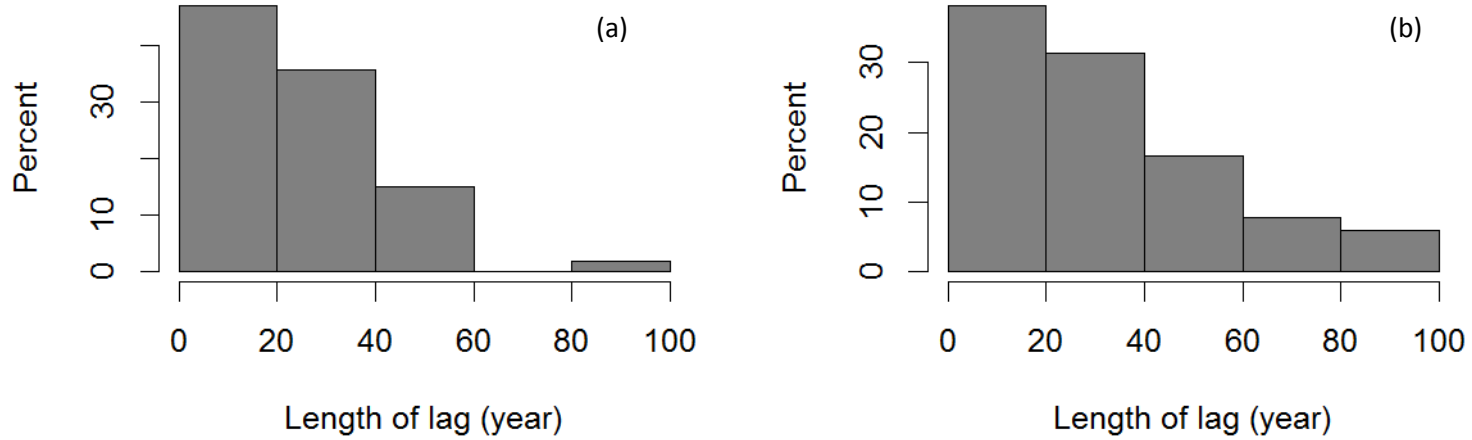

(a)
Frequency
0
2
4
6
8
1880
1920
1960
2000
Year
(b)
Adjusted frequency (compared to other species)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1880
1920
1960
2000
Year
(c)
Cumulative records
0
10
20
30
40
50
1880
1920
1960
2000
Year

Fig 5.


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