ONLINE SUPPLEMENT for

Estimating Rates and Probabilities of Origination and Extinction Using Taxonomic Occurrence Data: Capture-Mark-Recapture (CMR) Approaches

Lee Hsiang Liow
Centre for Ecological and Evolutionary Synthesis, Department of Biology, University of Oslo, Oslo Norway
(l.h.liow@bio.uio.no)

Introduction

There are multiple programs freely available for CMR analyses, and of course it is also possible to write your own scripts in your programming language of choice for such analyses once you grasp the principles behind the approach. Here we will focus on giving a first introduction to the freely available program MARK (http://welcome.warnercnr.colostate.edu/~gwhite/mark/mark.htm and http://www.phidot.org/software/mark/index.html) which runs on Windows. MARK was developed by Gary White, and he continues to add new models and features to this program. It has a user-friendly interface and can also be called by R. The R package that can be conveniently used in conjunction with MARK is RMark (http://www.phidot.org/software/mark/rmark/). We will not go into the mechanics of RMark here, because our aim is only to familiarize the reader with the first basics of MARK and CMR analyses. Note that there are also many MARK (and RMark) workshops organized in North America and other parts of the world. These cater mainly to population ecologists for now, so one should be a somewhat familiar with ecological jargon for full appreciation at such workshops. This short guide will hopefully help with some of the initial "translations" between ecological and paleontological jargon. Check out the next workshop at the MARK homepage (see links above) and be sure to read the freely available "Program MARK: a gentle introduction" also downloadable from the above links, after reading our short guide. This online MARK book, edited by Evan Cooch and Gary White, which we refer to as Cooch and White (2006) throughout, gives many more details on both MARK and CMR approaches than we are able to do within this space. But again, some familiarity with the ecological jargon used will illuminate how some of the models already available can be applied to paleontological questions.

The Cormack-Jolly-Seber (CJS) model

We will introduce MARK with a CJS model where we can estimate extinction probabilities and sampling probabilities conditioned on "initial encounters," i.e. we use the first appearance of a taxon as a starting point (see text). In other words, we model neither the initial "zeros" (the initial absences) nor the very first detection a taxon. We assume that you have successfully downloaded and installed MARK on your computer. Let's just go through the steps of doing a simple analysis with MARK before going into any details, such that we first get over any initial resistance to using a new, unfamiliar program.

Running 4 simple models on simulated data

i) Click on the MARK icon to open the program. Go to File and click on New (we will use blue when we are referring to tabs, pull-down menus or titles of browsers in MARK).

ii) Under Select Data Type, leave Recaptures only selected. We are going to use a simulated dataset that has 20 time intervals (or 20 sampling occasions , in ecological jargon).

iii) Fill in a title in Title for this set of data (say, CJSexample1); click on Click to Select File and select the file "CJSExample1.inp" from the directory where you have saved the 2 files accompanying this text. Click Open. The Encounter Histories File Name and Results File Name are then automatically filled in. But you can change the file names if that helps you to keep track of your analyses.

iv) Change Encounter occasions to 20 and leave Attribute groups as 1 and Individual covariates as 0.

Before we click OK and go on, let us explain a few things. First, we use "Recaptures only" (essentially specifying the CJS model), because we model neither the zeros that occur before the first appearance (or rather detection) of a taxon nor the first detection itself. In other words, we only use the detections and non-detections following the first appearance (i.e. recapturing after "marking" an animal in population ecology terms) of a taxon to make inferences on extinction and sampling probabilities. Clearly, there are many other kinds of analyses (see under Select Data Type when you start a new file or project), some of which are potentially suitable for paleontological data and questions, and some of which may not be. Second, the results file generated by MARK is automatically put in the same directory as the data file, but that can be of course changed. Third, Encounter occasions are the time intervals to which detected taxa are assigned. The time intervals separating encounter occasions can differ in duration (but see text for assumptions and a later section of this text). We have selected 1 for Attribute groups, because we have not differentiated the taxa in the dataset provided. We assume they are all equivalent, but of course this assumption can be relaxed. For example, if you had a mixture of bivalve and gastropod taxa in your dataset and you had reason to believe that gastropods have lower sampling probabilities than bivalves, you could potentially compare a model in which they had the same sampling probabilities with one in which they had different sampling probabilities. We will demonstrate this in the next example. The individual covariates are set to zero in this example, but one can imagine using say average body size as a trait varying among taxa in the dataset, if one thought that body size might be important in explaining extinction probability, say.

v) Click OK. MARK will indicate that a results file (a .dbf file) was created.

vi) You will immediately see a tab labeled Apparent Survival Parameter (Phi) Group 1 of Live Recaptures (CJS). It looks pretty scary, but ignore it for now and close the tab.

vii) Click on PIM and select Parameter Index chart. The PIM chart (Parameter Index Matrix Chart) should pop up. The numbers on the x-axis are indices for the parameters we are estimating. We'll explain more in a bit.

Let's restate our problem. We have 20 time intervals in our data, and we are interested in asking whether the data are better explained with i) constant survival probability and constant sampling probability through time; ii) constant survival probability and sampling probability varying through time; iii) survival probability varying through time and constant sampling probability; iv) both survival and sampling probabilities varying through time. We are going to use the PIM chart to help us set up these 4 models. Let's first write them down in the "standard notation."

1)phi(.)p(.)
2)phi(.)p(t)
3)phi(t)p(.)
4)phi(t)p(t)

The dot (.) just means constant for all time intervals, and (t) means varying from time interval to time interval. Phi is survival probability and can be interpreted as the complement of extinction probability whereas p is sampling or detection probability. Note that we will write phi instead of ϕ (as you might encounter in papers and books on CMR approaches) throughout because that is what MARK uses. Note also that $\phi = 1\text{-}\varepsilon$ (see equation (5) in our text).

viii) The PIM chart that shows up automatically is actually Model 4). Phi is allowed to have a different value from time interval to time interval (parameters 1-19 on the x-axis). There are thus 19 survival parameters, one for each period separating adjacent time intervals (usually each period is viewed as extending from the midpoint of one interval to the midpoint of the next interval). Same for p (parameters 20-38). Because the initial detection is not modeled, there are 19 p's corresponding to time intervals 2-20. So we can just click on Run and select Current Model. A new tab labeled Set up Numerical Estimation Run shows up, and we can type in the model name phi(t)p(t) so we know which model it is under Model Name. Under Link

Function click Logit. Now just click OK to Run. A box that specifies Use an identity design matrix since none was specified? shows up. Click YES.

ix) Depending on how quick your machine is, you may or may not see a console box. It closes after the estimation procedure is done and when MARK asks if it should Append this model's output to database (and you check that you put the right name on the model… things quickly get confusing if you have many models to run, and it pays to write model names very clearly and correctly). Click Yes.

x) If you toggle, minimize or close the PIM chart, you should then see a spreadsheet –like-tab labeled Results Browser: Live Recaptures (CJS). You should also see that the model you have run is on the first line, with MARK reporting the AIC$_c$, Model Likelihood and some other values for this model. Congratulations! You have just run your first CMR model in MARK! Get a cup of coffee or some other favorite beverage, and we can celebrate by running the other 3 simpler models listed above.

xi) Click on PIM and then on Parameter Index Chart again. Now we will set up model 2 (phi(.)p(t)). We want all the phi's in all the time intervals to be the same. Right click the mouse on the blue bars of phi group 1. A tab comes up. Click on constant. Immediately, all but one bar disappear for phi. Now MARK knows that it should only estimate one value for phi for all the time intervals. Left click on the blue bars of p and drag and drop them at the number 2 on the x-axis. Now you see that you have 20 parameters to estimate. One for phi and 19 time varying parameters for p. This parameterization describes model 2. Click on Run and then Current model. Change the model name to phi(.)p(t) and check that the Link Function is set to logit. And click OK to Run. Respond yes to using an identity design matrix and yes to appending the models output. Close the PIM chart.

xii) Now you see that the second model we have run appears in the Results Browser. It has a lower AIC$_c$ value than the first model we ran, and its Model Likelihood is 1 whereas the other model has 0 Model Likelihood. Note that Model Likelihood here is the AIC$_c$ weight of the model of interest divided by the AIC$_c$ weight of the best model in the set (see Burnham and Anderson 2002). This is *not* the likelihood of the parameters given the data, i.e. $L(\theta \mid X)$.

xiii) Now go back to step xi) and figure out how to run the two remaining models.

xiv) If you did xiii) correctly, you would see in the Results Browser a table with (approximately) this information.

Model	AICc	Delta AICc	AICc Weight	Model Likelihood	No. Par.	Deviance
{phi(.)p(t)}	2522.1558	0.0000	0.99998	1.0000	20	249.3871
{phi(t)p(t)}	2544.2173	22.0615	0.00002	0.0000	37	236.1014
{phi(t)p(.)}	2566.6114	44.4556	0.00000	0.0000	20	293.8426
{phi(.)p(.)}	2584.7202	62.5644	0.00000	0.0000	2	348.5080

We'll first explain how to interpret this table, then we will go back to a few things you might have wondered about as you clicked through the estimates in steps viii) to xiii).

AIC stands for Akaike's Information Criterion and the extra "c" means corrected for sample size (see Burnham and Anderson 2002, Cooch and White 2006). Delta AIC$_c$ uses the best model as a reference and shows the differences in AIC$_c$ scores between each subsequent model compared with the best model (in this case phi(.)p(t)). AIC$_c$ Weight is often interpreted as the normalized weight of evidence for the given model relative to the other models in the set. No. Par. shows how many parameters MARK estimated for the given model. Deviance shows model fit: the lower this is, the better the model fits the data. Model comparison is based on the Delta AIC$_c$ values and their associated weights (Burnham and Anderson 2002). Looking at the AIC$_c$ weights, it is clear that the model phi(.)p(t) is the most likely of the four models. Note that the most general model (i.e. phi(t)p(t)) will always fit the data best, but we need to strike a balance between model fit and simplicity, i.e., the number of parameters required. Thus, although the deviance of the model phi(t)p(t) is better (lower) than that of the best model, the AIC$_c$ values and weights show clear support for the model phi(.)p(t).

And in fact, in this case, we *know* that the model phi(.)p(t) best approximates the "truth", because that is exactly the model underlying the data that we simulated. In fact we also know the true value of phi and the temporally varying p's (see Appendix). Before we explain how to retrieve the estimated values of phi and p from MARK, we will take a break from looking at the results and very briefly explain a few terms that you might be wondering about.

AIC$_c$

- This is the corrected or adjusted AIC$_c$ and it is written as

$$AIC_c = -2\log(L(\hat{\theta})) + 2K + \left(\frac{2K(K+1)}{n-K-1}\right)$$

where $L(\hat{\theta})$ is the likelihood of the parameters symbolized by $\hat{\theta}$, given the data, *n* is sample size (in our case, usually the total number of taxon detections summed over all taxa and time intervals) and *K* is the number of parameters. The term in large brackets is intended to correct for sample size, and when *n* is very large, AIC$_c$ converges to AIC (Burnham and Anderson 2002).

PIM

- The Parameter Identification Matrix (PIM) specifies model structure by assigning an identification number to each mode parameter. Let us first clarify why there were 38 parameters in the model {phi(t)p(t)} although there were 20 time intervals. What we can estimate are the survival probabilities, phi, from one time interval to the next, i.e. phi$_1$ is the survival from time interval $1 \rightarrow 2$ (indexed as parameter 1 on the PIM chart in the model phi(t)p(t)) and phi$_2$ is the survival from time interval $2 \rightarrow 3$ (indexed as parameter 2) and so on, such that we have 19 phi's to potentially estimate. We also assume that the survival probabilities estimated are approximately from the mid-point of time interval *t* to the mid-point of time interval *t+1* (time intervals are encounter occasions using MARK terminology). And since we really only start looking at sampling probabilities from the second time interval (the CJS model conditions on the initial detections and does not model them), we start estimating the sampling probabilities from time interval 2, i.e. p$_2$ is the sampling probability in time interval 2 (indexed as parameter 20 in phi(t)p(t)) and p$_3$ is the sampling probability in time interval 3 and so on, giving us as total of 19 parameters. But wait! You might have noticed that in the Results Browser, MARK indicated that it only estimated 37 parameters! We'll see why this is a bit later when we look at the output of the estimates.
- We have used the option of Parameter Index charts for our first example because they provide a good visual view of how we are structuring our models. But another way to change our models is to use the Parameter Index Matrices (PIMs) directly. You can see this in the same pull-down menu (PIM). If you still have Results Brower from the last exercise open, highlight the model {phi(t)p(t)}, click Retrieve and click on Current model. Now click on PIM and then Parameter index chart. You see what you have seen before, 19 phi parameters and 19 p parameters. Now if you click on PIM and then Open Parameter Index Matrix, you should see two lines. One line is labeled Apparent Survival Parameter (Phi) Group 1 while the other is labeled Recapture Parameter (p) Group 1. Select the first line and click OK (you might need to reduce the PIM chart or close it to see the PIM). The numbers you see are the

indexing for the parameters we are interested in. 1 refers to phi_1 or survival from the mid-point of time interval 1 to the mid-point of time interval 2 (we'll save saying this every time after this, you get the point that it's the mid-points we are talking about) while 19 refers to phi_{19} or survival from time interval 19 to 20. These indices are equivalent to the numbering on the PIM chart. The PIM for *p* functions on the same principle. If we wanted to ask MARK to run the model phi(.)p(t), and we used the PIM to do it, then we would have to change all the numbers in the phi cells to 1. This will tell MARK to estimate only one value for all the phi's. The PIM is a more flexible way than the Parameter Index Chart to tell MARK about the models we want to run, as each individual cell can take different values (e.g. when you want to do cohort-based analyses where you model the survival as a function of the time interval of first appearance or detection).

- You might feel now that MARK seems a little cumbersome. But once you understand MARK, RMark is a great way to go because it gives you all the conveniences of working with R, but saves you the trouble of writing complex scripts to run models that MARK is very efficient at. So it really pays to learn MARK well first.

Link Functions

- We have briefly explained in our main text that link functions transform the probabilities we are interested in (say survival or extinction or sampling) such that they map on a scale that is suitable for the models we are building. More concretely, sampling probabilities, for instance, range from 0 to 1 by definition, but if we were interested in modeling those as a linear function of say facies and geographic range, then we would likely use the logit transform for sampling probabilities that yields values between 0 and 1, even when covariates may attain very large or small values. For identity design matrices (see later sections) the sin link function is a good choice for reasons of the optimization procedure in MARK, but for design matrices developed for covariates modeling, the logit link function is usually the appropriate choice. Detailed discussion of the choice of link functions to use is beyond the scope of this short guide (also see Lebreton et al. 1992 for an introduction).

Goodness of fit testing and c-hat

- We were not totally "kosher" in this short guide, because the first thing we would have done in a real analysis is to assess the fit of most general model (in Case Study 1, it is {phi[t)p(t)}) to the data. We cannot draw good inferences from a model set if its most general model does not fit the data adequately. But we didn't do this firstly because we generated the data set and hence we already know the answer to the model fit question, and secondly because we wanted to show you that MARK does not need to be intimidating although there might be many details you should take time to learn about.

- Although GOF sounds like a nuisance, it can tell you a lot about your data and how they should be modeled. For instance, if model {phi(t)p(t)}) didn't fit our first example well, it might actually be telling us that we cannot assume that all the taxa have similar survival probabilities. Perhaps we have a mixture of extinction resistant and extinction prone taxa in the data? The lack of fit may also be telling us that the cohorts of taxa (based on first appearances) may not be equivalent with respect to subsequent extinction and encounter probabilities . Perhaps those first-appearing later (say in the late Miocene), rather than earlier (mid-Miocene), have a greater survival probability?

Design matrix

- Design matrices are necessary when we want to ask MARK to run models that include covariate relationships and include certain kinds of constraints (e.g., certain kinds of additive models). We will explain design matrices in a bit more detail in the next example since they are easier to understand with concrete examples.

Looking at the estimated values of phi and p from the simulated model

Now that we have concluded that the model phi(.)p(t) is the best model of the 4 we came up with for explaining our data, it's time to retrieve the estimates of the parameters. Back to the Results Browser window. Highlight the model phi(t)p(t), right click and select Real Estimates. In the results file that is called up (it is a .tmp file which is really just a simple text file), you can see that there are 19 parameters for phi and 19 for p as we explained earlier. But remember MARK said that it only estimated 37 parameters. Look at parameters 19 and 38, i.e. ϕ_{19} and p_{20}. They have really huge standard errors and ridiculously large 95% confidence intervals, and their values are either very similar or identical. This is because the time series ends at time interval 20 and there is no further information with which to estimate p_{20}, so we cannot separate the term ($phi_{19}p_{20}$). Instead we can only estimate the product. So that was the mystery of the reduced number of parameters. There may be other parameters estimated with large confidence intervals, and you should be cautious in interpreting those (see Cooch and White 2006). If you right clicked again on phi(t)p(t) and selected Beta Estimates, you will see a similar results table showing the β estimates (more on these in Example 2). These are transformed back to the "real parameter values" as you have seen in the True Estimates. You can export these estimates as data to R or any other of your favorite statistical packages to plot the results, but MARK also provides a quick visual look at the estimates. Click on the icon showing line graphs within the Results Browser. Type in the Legend text say, phi(t). In the next box, type 1 to 18 (i.e. we are selecting all the estimated phi's, alternatively you can click the boxes beside the phi's to select them individually) and click ok. A plot of the estimates pops up. Not the prettiest, but it lets you visualize your results quickly. Now click on Add Data Series in the Graphics Window. In legend text type p(t) and in the next box type in 20 to 37 and click OK. Now you have added the estimates of p to the plot. You can save this file as a jpeg (click on FILE and you will see the options). Try plotting for the estimates for other models to get a feel for the results.

Now you know how to start a new project in MARK, specify with PIM charts the different models you want to compare, interpret the results of model comparisons using AIC, and plot estimates from the various models within MARK. You also know where to find the numerical estimates if you want to plot them in another computer package. If you go to the Appendix, you can check what the true values of phi and p were (the values used to generate the simulated data) and compare those to the estimates you got running MARK. These should not be exactly the same but should correspond closely.

In the next example, we'll add some covariates so you see how covariate analyses can be done in MARK.

CJS model with covariates

Go to File and then select New. We are still with Recaptures only. Type in "CJS example 2 with covariates" in the Title field. Select the file "CJSExample2.inp". You may also like to click on View File to see what the input file looks like (this will help you format your own data). Notice that in the data file (i.e. the CJSExample2.inp file), we have different encounter or detection histories and each of these is followed by two other numbers. The first number indicates how many bivalve taxa there are with the detection history in the given row, while the second indicates how many gastropod taxa there are. (Note that you can also simply list the detection history for each taxon and then index each row with 1 0 for a bivalve or 0 1 for a gastropod. This will just result in more lines of data to input to MARK, but of course MARK doesn't mind). These will be our group covariates. Notice that there is no space between the 1's and 0's in each detection history, but there is a tab to the group covariates and each line ends with a semi-colon. The semi-colon tells MARK that we have finished telling it about that particular detection history. Close the data file. You noticed that in the detection history, there was a string of 10 ones and zeros, so Encounter Occasions should be set to 10. Attribute groups can be set to 2.Click on Enter Group

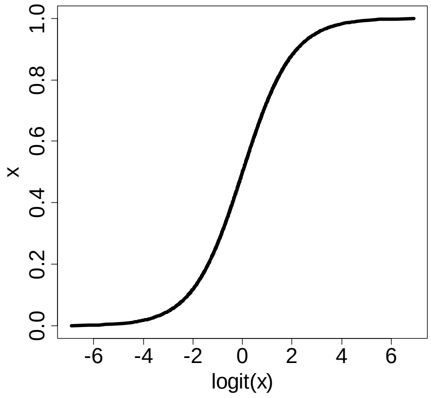
Labels and type Bivalves for the first and Gastropod for the second. We still have no individual covariates, but we could have used, for example, average body size of each taxon. Click OK (twice) (we'll save saying this every time after this, you get the point). You will see the PIM of Apparent survival Parameter (Phi)= Bivalves of Live Capture (CJS) show up. Let's just close that for now.

First, let's think about the models we want to compare. We know from previous work that it is likely that bivalves may be more persistent (lower extinction probabilities) than gastropods which have higher turnover rates. And many gastropods seem more fragile as well, so we might not be able to sample them as well as we do bivalves (smaller encounter probabilities). Let's also say that in the 10 time intervals for which we have data, we think that there was a period of exceptionally high extinction rates, say in intervals 4 to 6 (let's call this ME for mass extinction). Armed with these gut feelings, let's write down *some* of the models that might be reasonable to compare. (These are approximately in order of decreasing complexity).

1. Phi(group*t)p(group*t)
2. Phi(group+t)p(group+t)
3. Phi(group)p(group+t)
4. Phi(group+t)p(group)
5. Phi(group+ME)p(group+ME)
6. Phi(t)p(t)
7. Phi(group)p(group)
8. Phi(group)p(.)
9. Phi(.)p(group)

There are many more models we can write, but for economy of time and the efficiency simply of learning the ropes, let's just consider these. We recommend reading Burnham and Anderson (2002) for learning about constructing a list of models to compare. Among models 1-9 listed above, the "true model" (the one with which we simulated the data for this example) is included. Let's see what the models mean. Model 1 is the most general model including all the things we are concerned about: survival (Phi, the complement of extinction probability) is influenced by whether the taxon under consideration is a bivalve or a gastropod (group), which time interval we are taking about (t) and the interaction between these factors (group*t). Thus, under this model there is a separate parameter for each group in each time interval. Sampling probabilities are influenced by the same factors. All the other models are simplifications of Model 1. Model 1 is also the one for which we need to check the GOF. In Model 2, the interaction terms (group*t) are removed. Under this model, phi is modeled as dependent on taxon, but temporal variation is parallel on the logit scale, such that members of both taxa experience the same periods of relatively high and low extinction. The logit scale is such that $\text{logit}(x) = \log(\frac{x}{1-x})$, where *x* refers to either survival probability or detection probability in our examples (Fig. 1).

Fig.1



In Model 3, we are saying that survival depends on whether the taxon is a bivalve or a gastropod, but that the survival probabilities during all time intervals are the same. However, the sampling probabilities differ from one time interval to another and also differ between gastropods and bivalves. Model 4 is the opposite situation where true survival differs from time interval to time interval and also depends on whether the taxon in question is a gastropod or a bivalve; however, sampling probabilities are not time-dependent and only dictated by the group identity of the taxon involved. Model 5 is the only one in the set of models that explicitly formulates the idea of the extinction period. It calls for the estimation of separate survival/sampling probabilities for gastropods and bivalves and separate survival/sampling probabilities for time intervals during the “extinction period” and outside of that period. In practice we will tell MARK to make one estimate for time intervals 1-3 and 7-10, and a separate one for 4-6, for both survival and detection parameters. If we were right about the extinction period, survivorship and sampling might both be lower for time intervals 4-6 compared with 1-3 and 7-10. Of course, you might also see evidence of this in the time-varying estimates of, for example Models 1, 2, 4 and 6. You have seen Model 6 in the first example. Model 7 is really simple and implies that nothing matters but the group identity of the taxa involved, and that a single extinction and detection probability can be estimated for gastropods and then for bivalves regardless of time period. Models 8 and 9 are further simplifications of Model 7 where only one sampling probability is estimated and only one survival probability is estimated for both gastropod and bivalve data, respectively. As already mentioned, you can write down many more models. Do this as an exercise and think about what each of the models implies.

Let’s now go to MARK again. First, as promised, let’s look quickly at the GOF of the most general model in the set of nine. GOF testing is a very involved subject so please read more about it in Cooch and White (2006). Our treatment here does not do it justice at all. We just want to say that if your data do not fit the most general model in your set of models, you can ask MARK to estimate a correction factor which you can use in adjusting the AIC values of your models and variances of model parameter estimates. But you might also want to rethink both your data and what you can ask of it, and whether your general model is general enough. It might also be time to collect more or better data or to restructure your research questions if more data collection or rearrangement of your data is not possible. If you had thought through all this, then the correction factor is a good way to go.

First, let's run the general model. Click on RUN and then Pre-defined model(s). Click on Select Models. On the Phi tab, click on (g*t) and do the same for the p tab. At the bottom of the browser, you should see Number of Models to run is 1. Select the Logit Link Function. Click OK and then in the Set Up Numerical Estimation Run browser, click OK to Run. After this model runs, you will see {phi(group*t)p(group*t) PIM} show up in the Results Browser. This model is equivalent to having both additive main effects and interaction effects for group and time. We'll also use this opportunity to talk about the design matrix.

We want to model phi and p as functions of both time interval and group. One way to do this is via the logit link function, as we described above. By doing so, we can model phi and p using covariates (remember that p and phi are probabilities and hence range from 0 to 1 by definition but the logit of p and phi are not constrained to 0 to 1, see Fig. 1). The β's are what we want to estimate, and we need X, the design matrix (consisting of dummy variables, of ones and zeros) to track particular phi or p with their specific combination of β's. For instance, if we are interested in the phi for bivalves at t1, we would write: $\text{logit}(phi_{t1(bivalves)}) = \beta_0(1) + \beta_{group}(1) + \beta_{t1}(1) .... + \beta_{t8}(0) + \beta_{t1*group}(1) .... + \beta_{t8*group}(0)$. The design matrix is a convenient and economical way to write, in a matrix, the ones and zeros in the parentheses next to the β's for all the time-varying and group varying phi combinations.

Click on Design and Full to show what the Design Matrix for this model looks like. There are 36 columns in this full design matrix. The first column (labeled as B1: Phi Int in MARK) is the intercept term for survival (β_0 for Phi). The second column describes the group effect for survival (β_{group} for Phi: the cells in this column are 1's if the taxon in question is a bivalve and 0 if it is a gastropod). Columns 3-10 (labeled B3 Phi t1 through t8 in MARK) are for the time intervals (β_{t1} throughβ_{t8} for Phi) Why are there only 8? Well, remember that although we have 10 time intervals in our data, we can estimate survival probability from one time interval to the next (that makes 9). But because of the way the design matrix is written, we can specify phi for the last period by inserting 0's in all of the beta parameters for periods t1 through t8 (it is represented by the sum of the first beta and the second beta multiplied by the taxon), i.e.

$\text{logit}(phi_{t9}) = \beta_0(1) + \beta_{group}(1)$. Now look at columns 11- 18. Those are really just the dummy variables (i.e. the ones and zeros) in the group column (column 2) multiplied by each of the time intervals (columns 3-10), and they give the interactive effect between the taxonomic identity and time interval. In the next 18 columns, you see the equivalent β's for the *p*'s. A good thing to remember is that the columns represent all the β's you want MARK to estimate, and the rows correspond to the real parameters in the PIM. Let's go back to the Results Browser. Look at the column No. Par. MARK notes there are 34 parameters. But wasn't it supposed to be 18+18=36? MARK can only estimate 35 parameters because remember that the last parameters cannot be separated. MARK is not very good at counting the number of parameters it has estimated, although you automatically get a guess from MARK. If you highlight the model and left click and call up the Real Estimates, you will see that 4 of the estimates have super wide 95% confidence intervals, two of those we expected (the last parameters for phi and p) and two we didn't *a priori* know about. Before we continue with specifying and running the other models, let's look very briefly at GOF and c-hats.

c-hat is a variance inflation factor (see Lebreton et al. 1992) which quantifies over-dispersion. Over-dispersion means that the data we have contain more variation than our model manages to account for. Under-dispersion is the opposite, i.e. the data contain less variation than we have incorporated in our model. Over-dispersion is frequently the case with real data. What we can do is to use c-hat to get an adjusted AIC$_c$ value in our model comparison using:

$$QAIC_c = \frac{-2\log(L(\hat{\theta}))}{\hat{c}} + 2K + \left(\frac{2K(K+1)}{n-K-1}\right).$$

when c-hat =1, there is no over-dispersion and QAIC$_c$= AIC$_c$.

There are many ways to properly estimate c-hat (Cooch and White 2006) so that we can adjust both the AICs and the variances of the model parameter estimates. Basically, we have to do a GOF test, usually via bootstrapping or simulation techniques, and then estimate a c-hat for use . If the estimated c-hat is larger than 1, then we have to go to Adjustments to modify the c-hat used for comparing the models. In the case of our simulated dataset, the general model fits the data well and estimated c-hat is slightly less than one, so we don't have to do anything. Please refer to Cooch and White (2006), Chapter 6 for more on GOF and c-hats because this is a crucial subject in CMR approaches.

Now let's continue with specifying the other models. Models can be specified in MARK in different ways, via pre-defined models as we have just done, using the design matrix and using the PIM or the PIM charts. All models can potentially be expressed using the design matrix approach, but not all can be defined using the other approaches. For didactic purposes, we won't specify models in the order we listed them.

The Model 7, phi(group) p(group), implies that the survival and sampling probabilities of bivalves and gastropods are different, but that these parameters do not vary by time interval. Let's create the model the way we have already discussed in CJS example1. Click on PIM and then Parameter Index Chart. Highlight each one of the blocks you see and right click the select Constant. Click RUN and then Current model. Name it phi(group)p(group)PIChart or something that as this will help you distinguish this run from the next two runs you will do for the same model. Agree to using an identity design matrix. The model runs quickly and the results should appear on the Results Browser. We'll run the same model again in a different way, but let's move on to another model first.

Let's now look at Model 5, Phi(group+ME)p(group+ME). Here, we want bivalves and gastropods to have their own survival probabilities and sampling probabilities. And these should be allowed to take different values during the special extinction period. So the phi's from time interval 4 to 5 and 5 to 6 (phi_{t4} and phi_{t5} respectively) should take on the same values while those outside of those time intervals can take on another value, for bivalves and gastropods separately. Similarly, the p's for time intervals 4-6 should take one value and those outside of these time intervals another, for bivalves and gastropods separately. Click on PIM and Open Parameter Index Matrix and select the Phi matrix for Bivalves. Change all the numbers in the first 3 columns to 1, the next two columns to 2 and the last four to 1 such that it looks like this:

1	1	1	2	2	1	1	1	1
	1	1	2	2	1	1	1	1
		1	2	2	1	1	1	1
			2	2	1	1	1	1
				2	1	1	1	1
					1	1	1	1
						1	1	1
							1	1
								1

(Actually if you ran Model 7 previous to this, the PIM should already be all 1's, so even though it was cumbersome, it wasn't that bad). But we don't have to do it this way. Close this matrix and open the phi matrix for gastropods. Open an Excel Spreadsheet and using the structure of the phi matrix for gastropods, type in 3 for the first 3 columns, 4 for the next two and 3 again for the last two. Like this:

3	3	3	4	4	3	3	3	3
	3	3	4	4	3	3	3	3
		3	4	4	3	3	3	3
			4	4	3	3	3	3
				4	3	3	3	3
					3	3	3	3
						3	3	3
							3	3
								3

Copy these excel data, right click on the grey space on the phi matrix of gastropods and click on Paste from Clipboard. The numbers that we want will appear and now you can also have the PIM saved else where for easier manipulation. Let's just finish up with the p PIM for bivalves:

5	5	6	6	6	5	5	5	5
	5	6	6	6	5	5	5	5
		6	6	6	5	5	5	5
			6	6	5	5	5	5
				6	5	5	5	5
					5	5	5	5
						5	5	5
							5	5
								5

and the p PIM for gastropods which should end up looking like this:

7	7	8	8	8	7	7	7	7
	7	8	8	8	7	7	7	7
		8	8	8	7	7	7	7
			8	8	7	7	7	7
				8	7	7	7	7
					7	7	7	7
						7	7	7
							7	7
								7

Run this model. Remember to write the name of the model in the title field. It is easy to get names wrong and screw up interpretations and analyses especially with a greater number of models. Agree to using an identity matrix when asked. You should get results saying that the general model is still better (great AIC$_c$ weight). Now that you have run this model, let's examine what we told MARK. In the phi matrix for bivalves, we told MARK to have the survival probabilities from time intervals 1 to 2, 2 to 3 and 3 to 4 be the same as those from after the extinction period (time intervals 6 to 7, 7 to 8, 8 to 9 and 9 to 10). The mass extinction time intervals get their own value (parameter 2) in the PIM (i.e. survival from time interval 4 to 5 and 5 to 6). Similarly, for the phi's for gastropods, parameter 3 is for survival probabilities from time intervals 1 to 2, 2 to 3, 3 to 4, 6 to 7, 7 to 8, 8 to 9 and 9 to 10 and parameter 4 for time intervals

4 to 5 and 5 to 6. There is no typo in the sampling matrices; why do they have different parameterizations from the phi matrices (i.e. 3 extinction time intervals instead of the 2 as for the phi matrices)? If the mass extinction period is "real," then we might expect to see that the phi's for both gastropods and bivalves are lower for time interval 4 to 5 and 5 to 6, whether or not p are higher or lower for time intervals 4-6.

Let's go back to Model 7 {phi(group) p(group)}and use what we just learned in the last model to tell MARK how to run Model 7 via PIM. Open the 4 PIMs. How would we specify Model 7 ? Basically we want each of the PIMs to contain only a single value. So change all the cells in the phi parameter for bivalves to 1, all in the phi parameter for gastropods to 2, all the cells in the p parameter for bivalves to 3 and lastly the p parameter in gastropods to 4. Run this model and name it {phi(group)p(group) PIM} or something else to distinguish it from our previous run of Model 7. The results of this run should be exactly the same as our previous run of Model 7 using the PIM chart. But you get an extra entry in the Results Browser because you have named this differently to distinguish it from the first Model 7 run you did. For the sake of being didactic again, let's run Model 7 for the last time, but this time via the design matrix. Click Design and then Reduced or Identity Matrix. Specify that you want 4 covariate columns in the design matrix (if you tried to select Full matrix, MARK will tell you the full matrix should have 36 parameters in the PIM). How would we write the design matrix? Let's look at the survival parameter phi. We want to write

$$\text{logit}(Phi_i) = \beta_0 + \beta_{group} x_i$$

such that there is an intercept β_0 and a slope β_{group} describing the difference between bivalves and gastropod survival. The x's are dummy variables while the index *i* is for Bivalves or Gastropods. And we want the same for the sampling parameter. Label the 4 columns perhaps something like phi int, phi group, p int, p group. (You do this by right clicking and selecting label columns for each column). Enter these values:

Phi int	Phi Group	P int	P group
1	1	0	0
1	0	0	0
0	0	1	1
0	0	1	

Basically we have told MARK that we want

$$\text{logit}(Phi_{bivalves}) = \beta_0 \times 1 + \beta_{group} \times 1 = \beta_0 + \beta_{group}$$

$$\text{logit}(Phi_{gastropods}) = \beta_0 \times 1 + \beta_{group} \times 0 = \beta_0$$

where the ones and zero in parentheses refer to those in our table. The beta estimates from MARK are those on the right hand side of the equations while the real estimates are the logit-transformed phi's on the left hand side. β_0 is the logit transform for gastropod survival while β_{group} is the difference (on the logit scale) between bivalve and gastropod survival. The p's are identical in their form.

Name this model {phi(group)p(group) design matrix} and run it. Open the text files for both Real and Beta Estimates (by right clicking on the model you just ran). You see there are 4 real estimates and you should check that these are the same as the real estimates for Model 7 you have run using either the PIM or PIM chart. But note that the beta estimates are not the same! First, let us use the estimates from this design matrix run to make sure we understand how to convert the beta estimates to real estimates. The beta estimates for the phis are 1.492 and -1.378 for phi Int and phi group respectively. The real estimates are 0.528 and 0.816. We have assigned bivalves the dummy variable of 1. Using the bivalve estimates,

$$\operatorname{logit}(Phi_{bivalves}) = \beta_0 + \beta_{group} = 1.492 - 1.378 = 0.114$$

hence

$$\operatorname{logit}(Phi_{bivalves}) = \log(\frac{Phi_{bivalves}}{1 - Phi_{bivalves}}) = 0.114$$

And rearranging, we have

$Phi_{bivalves} = \dfrac{1}{1+e^{-0.114}} = \; 0.528$, the real estimate for bivalves! Note that β_{group} is negative, telling us the survival probability of bivalves is actually lower than that of gastropods.

Similarly for the gastropod estimates

$$\operatorname{logit}(Phi_{gastropods}) = \beta_0 = 1.492$$

Hence

$$\operatorname{logit}(Phi_{gastropods}) = \log(\frac{Phi_{gastropods}}{1 + Phi_{gastropods}}) = 1.492$$

Rearranging

$Phi_{gastropods} = \dfrac{1}{1+e^{-1.492}} = \; 0.816$, the real estimate for gastropods.

In the PIM and PIM chart runs of Model 7, we specified an identity matrix, which looks like this:

Phi	Phi	P	P
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

In such a case, there is no intercept and the real and beta estimates are related such that

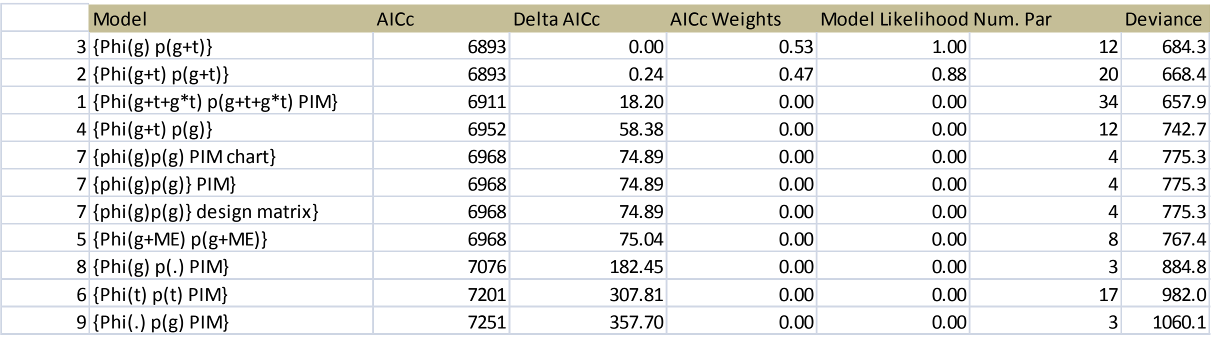
$$\text{logit}(Phi_{bivalves}) = \beta_1$$

$$\text{logit}(Phi_{gastropods}) = \beta_2$$

(Look at the estimates from MARK and try to convert between the real and beta estimates yourself).

Let’s write Model 2 {Phi(group+t)p(group+ t)} using the design matrix to get more practice. Let’s go via the full identity matrix this time. Retrieve the general model by right clicking on the highlighted Model 1 in the Results Browser. Now you can select Design and chose Full Matrix. To get Model 2, all we really have to do is to delete the interaction columns. Right click on the design matrix and select Delete multiple columns. Enter columns 11 to 18 since that corresponds to the interaction terms for the phi parameter. Make sure the columns are labelled (otherwise it is harder to keep track of the parameters). Delete also the interaction terms for the p parameter (columns 21-28). We now have Model 2. It is the best model so far. For Model 3, delete columns 3-10 corresponding to the time parameters for phi. It did even better than Model 2, although the model likelihoods of Models 2 and 3 are quite close. Now run Model 4. The easiest way to do it is to retrieve Model 2 and then delete the time betas for p. Let’s run and save this. It is just a matter of deleting the “time” columns for phi and p respectively. Model 4 didn’t do so well.

Models 6, 8, 9 are straightforward and you can either use pre-defined models or any other way to run them. Perhaps try different ways of specifying these in MARK to make sure you have understood how to communicate with MARK. When you are all done, these are the (approximate) results you should be looking at:



	Model	AICc	Delta AICc	AICc Weights	Model Likelihood	Num. Par	Deviance
3	{Phi(g) p(g+t)}	6893	0.00	0.53	1.00	12	684.3
2	{Phi(g+t) p(g+t)}	6893	0.24	0.47	0.88	20	668.4
1	{Phi(g+t+g*t) p(g+t+g*t) PIM}	6911	18.20	0.00	0.00	34	657.9
4	{Phi(g+t) p(g)}	6952	58.38	0.00	0.00	12	742.7
7	{phi(g)p(g) PIM chart}	6968	74.89	0.00	0.00	4	775.3
7	{phi(g)p(g)} PIM}	6968	74.89	0.00	0.00	4	775.3
7	{phi(g)p(g)} design matrix}	6968	74.89	0.00	0.00	4	775.3
5	{Phi(g+ME) p(g+ME)}	6968	75.04	0.00	0.00	8	767.4
8	{Phi(g) p(.) PIM}	7076	182.45	0.00	0.00	3	884.8
6	{Phi(t) p(t) PIM}	7201	307.81	0.00	0.00	17	982.0
9	{Phi(.) p(g) PIM}	7251	357.70	0.00	0.00	3	1060.1

We have added the model number (1st column in the table above) as listed earlier for easier reference. The two top models {phi(group)p(group+t)} and {phi(group+t)p(group+t)} and the other ones we have run are not really comparable (look at the AIC weights and Model likelihood). This tells us that our data strongly support the additive model of taxon and time (p(group+t)) for detection probabilities. Thus, detection probabilities differ by taxon but vary over time in parallel on the logit scale. Model selection results are ambiguous about whether survival is better described as constant over time or time-varying, but parallel for the two taxa. Make sure you look at the results files and ask MARK to help plot the data to make sure you understand the model structures since it was almost too easy to just ask MARK to do things in its black box. The top model {phi(group)p(group+t)} is in fact the model from which we have simulated our data (see Appendix to compare the estimates with the input parameters), but the second model is quite close. In practice, we might want to average our estimates (of survival and detection probabilities, especially if we want to use these estimates for further analyses). This can be done via model averaging, where an average value of an estimated parameter can be calculated using the weighted (AIC weights) averages of 2 or more probable models (Burnham and Anderson 2002 and Mark book).

Varying durations of time intervals

The time intervals in paleontological data often vary in length. One solution to this issue is to include the length of the period between successive time intervals in the modeling. MARK permits addition of such data (Cooch and White 2006).

Coda

We hope this short D-I-Y exercise has given a good introduction to both the program MARK and CMR. Clearly, this should only serve as a springboard to exploring and understanding both MARK and the details of CMR approaches. We stress that the process of thinking about your data, both those you will collect and those you already have in hand is really important to making trustworthy inferences. If a good approach, like CMR does not “give” you results you seek, for instance, the GOF is very poor, or the parameters have very poorly constrained estimates, you should rethink whether you have enough data to answer the questions you are interested in. Both research design and data volume are important factors in contributing to sound inference.

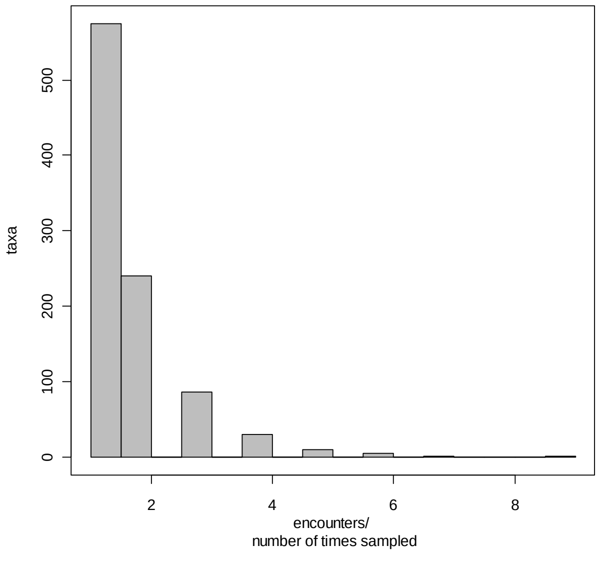
Appendix

There is a simulation module in MARK which can be used for many purposes, among which to simply simulate datasets for which we know the underlying model structure and input parameters. It is often very instructive to test out various approaches with simulated data where you have full control of underlying parameters. Because there is stochasticity in simulations, the output from single simulations will not be the same. Our two simulated datasets are generated with the input as described below, but if you ran the same simulations, you will not get the same data as the attached files, but the process that generated those data will be the same.

CJS example 1

Specifications: Recaptures only; 20 occasions, 1 group, true model is phi(.)p(t)

We set the true value of the constant phi to 0.5 (survival probability is 0.5 for all taxa for all time intervals). The true values for p are set to alternating between 0.5 and 0.8 from one time interval to the next, i.e. p_2 = 0.5, p_3= 0.8 etc. We “released” 50 taxa during each time interval, that is, 50 new taxa have first appearances at each time interval, which are also “sampled” during that particular time interval. Hence there are 50x19 = 950 taxa in the data, 134 types of encounter histories and most taxa are only sampled once, i.e. during the time interval they are “released” (see histogram).



CJS example 2

Specifications: Recaptures only; 10 occasions, 2 groups (Bivalves = group 1; Gastropods = group 2), true model is phi(group)p(group,t).

We set the true value of phi for bivalves to 0.5 and that for gastropods to 0.8. We varied the detection probability of both gastropods and bivalves such they increased through time. But bivalves are given higher p's though out. Specifically, p_2=p_3=0.6, p_4=p_5=0.7, p_6=p_7=0.8, p_8=p_9=p_{10}=0.9 for bivalves and p_2=p_3=0.3, p_4=p_5=0.4, p_6=p_7=0.5, p_8=p_9=p_{10}=0.6 for bivalves. We "released" 100 taxa during each time interval.

References and further reading

Burnham, K. P. & D. R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd edition. Springer, New York.

Cooch, E. & G. White. 2006. Program Mark: A Gentle Introduction (http://www.phidot.org/software/mark/docs/book/).

Lebreton, J. D., K. P. Burnham, J. Clobert, et al. 1992. Modeling survival and testing biological hypotheses using marked animals - a unified approach with case-studies. Ecological Monographs **62**:67-118.

Acknowledgements

My deepest gratitude to Jim D. Nichols for helping me improve this supplementary text. All remaining errors are my own.