ADU

Biodiversity Observations

http://bo.adu.org.za

biodiversity observations

bo

An electronic journal published by the Animal Demography Unit at the University of Cape Town

THE FUNDAMENTALS OF THE SABAP2 PROTOCOL

Les G Underhill

Recommended citation format:
Underhill LG. 2016. The fundamentals of the SABAP2 protocol. Biodiversity Observations 7.42: 1–12.

URL: http://bo.adu.org.za/content.php?id=235

Published online: 2 August 2016

– ISSN 2219-0341 –

PROJECT REPORT

THE FUNDAMENTALS OF THE SABAP2 PROTOCOL

Les G Underhill

Animal Demography Unit, Department of Biological Sciences,
University of Cape Town, Rondebosch, 7701 South Africa

Email: les.underhill@uct.ac.za

This essay sets out to explain the motivation that underpins the protocol used by SABAP2, the Second Southern African Bird Atlas Project. Used in this sense, the term protocol means a set of instructions to collect data.

In the year before SABAP2 was set up in 2007, a variety of ideas were canvassed about how to go about the new bird atlas project, and there were many brainstorming sessions, both formal and informal. The fact that the 2007 protocol remains unchanged, and with no suggestions for changes being proposed, is evidence of a sensible design right from the start. The final approval to the SABAP2 protocol was made by the project's Steering Committee; this was done before the launch of the project in July 2007. In spite of all this stability, and possibly because of it, the motivations for the SABAP2 protocol have never been articulated comprehensively. That is what this paper aims to do.

The fieldwork component linked to the SABAP2 protocol has gained broad acceptance by the citizen scientists who do the fieldwork. Atlasing, using the SABAP2 protocol, has proved to be an enjoyable activity in its own right. Some would even say it is addictive. Alongside this, the BirdLasser app reduces atlasing fieldwork to bird identification only, the component that atlasers consider to be fun. The non-birding elements of atlasing, i.e. map-reading and data submission, both mostly regarded as drudgery, are taken care of by the app. BirdLasser currently has versions for southern Africa and Kenya, is to launch a single app for Africa and the associated islands. The convergence of these three systems – one analytical, the statistical potential – one sociological, an addictive protocol – and one technical, the cellphone app – creates the opportunity to initiate the African bird atlas. But the main focus of this essay is on the analytical component, the processing of the data collected by atlasers using the protocol.

TRUTH and the OBSERVER PROCESS

Modern statistical analysis considers raw data as having two components: the TRUTH and the OBSERVER PROCESS. In other words, data contains "signal" and "noise" and the task of the statistician is to separate these two components.

The TRUTH is the reality on the ground that is being investigated. In bird atlas terms, the TRUTH might be the answers to questions like "Are Capped Wheatears present in this grid cell or not?", or "What is the relative abundance of Capped Wheatears across grid cells?", or "How does the relative abundance of Capped Wheatears vary through the year?", or "How many Capped Wheatears are there in each grid cell?", or "What is the relative abundance of Capped Wheatears in comparison with Purple Swamphens and Hadedas?" or "Has Capped Wheatear gone extinct in this grid cell?". These questions ought to have time frames attached to them as well.

The OBSERVER PROCESS consists of all factors that lead to data being, to a greater of less extent, a distortion of the TRUTH. If both Capped Wheatears and Purple Swamphens are indeed present in a grid cell, it is more likely that the Capped Wheatear will be on an observer's atlas checklist than the Purple Swamphen, and it is less likely that it will be observed than the Hadeda. In other words, the conspicuousness of the species impacts on whether it is recorded or not. The likelihood of detection varies between species.

The abundance of a species in a grid cell also impacts the likelihood that an observer encounters the species. If there is a single Capped Wheatear in one grid cell, it is far less likely to be detected than if there are a hundred in another, although it is present in both. The ability and personality of the observer also play key roles: in identification skill, in fieldwork skill (such as knowing which habitats are likely to produce additional species, and knowing when to move on to the next habitat), and in persistence to make the list for the grid cell as complete as possible. Weather conditions and time of day also impact the likelihood of observing a species.

Time of year also makes a difference – a July checklist in South Africa almost certainly misses Barn Swallow, although the species is abundant in the grid cell in midsummer, and a Southern Red Bishop is less conspicuous outside of the breeding season than inside it. With a good field work protocol, many (but not necessarily all) aspects of the OBSERVER PROCESS can either be modelled or taken into consideration.

The First Southern African Bird Atlas Project (SABAP1) was one of the first bird atlas projects worldwide to devote intensive attention to the biases due to the OBSERVER PROCESS (Harrison & Underhill 1997, pages li–lviii). Many of the lessons learnt were built into the protocol for SABAP2. It is impossible to eliminate the OBSERVER PROCESS completely; but it is possible to limit the biases it causes, and to design the protocol so that it can be modelled statistically. Most importantly, we now have a good understanding of the biases due to the OBSERVER PROCESS.

Naïve protocol

The main challenger to the protocol eventually adopted for SABAP2 is the "naïve protocol". In 2007, at the start of the 21st century, it seemed obvious that the right way to do a bird atlas is simply to generate GPS coordinates for as many individual birds as possible. This provides accurate position fixes, and the bird distributions can subsequently be mapped on any chosen scale. By the second decade of the 21st century, technology has advanced to the extent that the naïve protocol is straightforward to implement as a cell phone app. This is the obvious approach from the information systems, or computer science, perspective. We call this the naïve protocol. It is a bottom-up approach – "data can be collected in this way, therefore we will collect it like this". This seems to be the self-evident and logical protocol from the perspective of "Wow, computers can do this, therefore it must right".

However, there are two fundamental problems with this naïve protocol. First, it generates data with a horrid observer process, with patterns that are dependent on the whims of the observer. And the second follows from this, it is difficult to extract the "signal" from the "noise". The naïve protocol generates hostile data from the statistical analysis perspective. In fact, it generates data which can essentially only be used to indicate where the species is definitely present; it cannot be used to determine whether the species is present at a site, but not recorded, compared with whether it is genuinely absent. In technical terms, it cannot distinguish a "false negative" from a "true negative".

Some of the problems with the observer process of the naïve protocol seem insurmountable. First of all, most birds appear to occur along or close to the routes that humans use. The naïve protocol creates no self-incentive to do fieldwork at, for example, a dam distant from the road that requires special permission to visit, or to visit a rocky hill on private property off the road that involves a hard walk over rough ground. Secondly, it is humanly impossible to GPS the position of every bird encountered. The process is selective, and the intensity of reporting varies enormously between observers, and within observers, through time. For example, an observer might georeference the first 10 Capped Wheatears encountered on a morning in the Swartland, but then start to become gradually more selective after that. For example, observers tend to report what they consider interesting, and they fail to report what they think is common. This results in an unquantified and unquantifiable bias in the observer process. These two factors render the naïve protocol unamenable to elegant and sharp statistical analyses.

The bottom-up approach of the naïve protocol is basically saying: "We will collect data in the simplest possible way, and the analysts can like it or lump it". There are many variations of the naïve protocol.

SABAP2 protocol

Before one embarks on any data collection exercise, careful thought needs to be given to the statistical processes which will be used to analyse the data. It is the analysis specifications which determine the protocol, and not the other way round. Even more important than asking whether the data generated will be analysable is having insight into the main questions that are going to be asked of the data.

One of the amazing things that happened to the SABAP1 data was that it was interrogated in ways that had never been anticipated. The SABAP1 data was the core resource for many research papers and postgraduate theses (Underhill 2016). It became the dataset of choice for the world leaders in macroecology such as Kevin Gaston (then at Sheffield University, now at Exeter University), Steven Chown (then at the University of Pretoria, now at the University of Brisbane), Walter Jetz (then at Oxford University, now at Yale University) and their postgraduate students. They used the SABAP1 dataset to test their ideas on a variety of the "big research issues" in macroecology and biogeography. For example, one of their main research areas was the efficient design of nature reserves in such a way that all species were included in at least one reserve. Their choice of the SABAP1 data over all the other datasets available worldwide, especially in Europe and North America was based on the protocol used for SABAP1. Indeed, these academics and their students used the SABAP1 data to help build their fame and their careers! Their biggest complement was to the SABAP1 protocol; it had to be on the right track (a) for them to use it, and (b) for their papers to survive the rigorous scrutiny of the journal refereeing process! The SABAP1 protocol needed to be tweaked for SABAP2, not discarded.

So our starting point is that the protocol needs to be guided by a paradigm that says (a) think first about the questions we are definitely going to ask of the data collected and (b) think also about the statistical analyses which will be needed to answer these questions. This is a top-down approach to designing the protocol.

In a nutshell, the task of SABAP2 is to generate information that can help us answer the large-scale questions: "How are bird distributions changing in South Africa?", "Are there relationships between the biomes defined by botanists and the distributions of birds?" and "Does the Kruger National Park make a difference to bird conservation?" and "Is the timing of migration to and from South Africa changing through time?". SABAP2 was never intended or designed to answer questions about parcels of land smaller than the "pentad"; these are questions like "Is it OK to build a shopping mall on this property?" or "Does Kirstenbosch National Botanical Gardens make a difference to bird conservation?". However, Harrison (1993) and Harrison et al. (1994) devised a practical method to narrow down the full list of species for a grid cell to a shorter list for a site within the grid cell, from a knowledge of the habitat characteristics of the bird species and the sites.

In brief, the “SABAP2 protocol” has the following features. (1) There is a clearly defined spatial unit which is five minutes north to south, and five minutes east to west. This unit of space is called the pentad. Results can be scaled up to coarser scales, but not to finer scales. (2) Participants in the project aim to make as complete a list as they can of the species that were present in the pentad during the observation period. They do not need to cover the whole pentad, but they need to try to sample as many of the habitats that occur in the pentad as possible (acknowledging that there are often constraints on access).

(3) Atlasers do focused birding for a minimum of two hours in their pentads, but can continue for longer (and are encouraged to do so if they are still regularly adding species). (4) They list the species in the order in which they see the species. (5) They can continue to add “additional species” to their list for five days. (6) Then they can start a new list for the pentad. The five-day gap helps ensure that each successive list submitted by the same observer is not simply a clone of the previous list, which would be the case if observers were allowed to submit lists for the same pentad on successive days. A checklist produced in this way is referred to as a “full-protocol checklist”.

The protocol entered its 10th year of use in July 2016. Underhill & Brooks (2016) reviewed the progress that the project had made in its first nine years. On the coverage map dated 30 July 2016 (Figure 1) 75% of the pentads in South Africa, Lesotho and Swaziland had at least one full-protocol checklist made for them.

SABAP2 grid system

Having a consistent predefined spatial unit simplifies statistical analyses massively. The spatial unit, the pentad, is objectively defined, and all observers operate to the same spatial scale. In technical terms, pentads are mutually exclusive and exhaustive. Mutually exclusive means that they do not overlap with each other; exhaustive means that they cover the entire country (they actually cover the continent of Africa and, in fact, the planet).

Why was the five-minute grid that generates pentads finally chosen as the spatial unit for SABAP2? There was one lesson that was abundantly clear from SABAP1 – that the grid system which had been used was too coarse. This was the 15-minute grid that generates “quarter-degree grid cells” (a misnomer because there are 16 quarter-degree grid cells in a degree square). At the time of SABAP1, this was the finest manageable grid. There are about 2000 quarter-degree grid cells in South Africa, each a single “1:50,000 map sheet”, a cumbersome thick paper document, 75 cm×53 cm in size. Most SABAP1 atlasers used 1:250,000 maps; each map sheet covered two degrees, and there were 70 for South Africa as a whole. Atlasers purchased these paper maps, ruled lines across them to show the 32 quarter-degree grid cells on the sheet. On these maps, a quarter-degree grid cell measured 115 mm north-south and 100 mm east-west. At this scale, and working without a GPS, the exact boundaries of grid cells on the ground are uncertain, and it is easy to make map-reading errors of 1 km or more (1 km on the ground = 4 mm on the map).

Once a decision to use a finer grid system for SABAP2 had been taken, the alternatives were (1) an eighth-degree grid, generating 64 cells per degree, (2) a 16th-degree grid, generating 128 cells per degree, (3) a five-minute grid, generating 144 cells per degree (pentads), (4) a three-minute grid, generating 400 cells per degree (triads), and (5) a one-minute grid, generating 3,600 cells per degree (monads). Alternatives such as a four-minute grid were not considered, because the resulting data could not be scaled up to the quarter-degree grid, so that comparisons with SABAP1 would not be feasible.

The concept of splitting the one-degree cells successively into four is an attractive one, and promoted by Larsen et al. (2009); although this paper was published in 2009, the concept had been presented at a conference several years earlier (see "Ragnveld" (2005–2015), an article called "QDGC" in Wikipedia). The eighth degree grid had been used by Vincent Parker for his bird atlas of Swaziland (Parker 1994). It was also used for the statistical analyses for his MSc dissertation, and the papers that flowed out it (Parker 1995, 1996, 1999). However, the eighth degree grid cell (64 per degree), was not considered to be of sufficiently fine resolution for SABAP2. Dividing these grid cells in four generates the 16th-degree grid, but this has the boundaries of the grid cells at multiple of 3.75 minutes (0 minutes, 3.75 minutes, 7.50 minutes, 11.25 minutes, …, 56.25 minutes), and is totally unworkable. That left a choice between pentads, triads and monads. Given that there are roughly 140 degree cells in South Africa, Lesotho and Swaziland, the back of envelope calculations were 20,000 pentads, 56,000 triads, and 500,000 monads.

On that basis, it was pretty easy to decide that the largest challenge we dared put before the citizen scientists was to operate at the scale of pentads! Because parts of many of the 140 degree cells which touch South Africa, Lesotho and Swaziland are in the oceans, it turned out that there were 17,000 terrestrial pentads in these three countries, the initial SABAP2 region in which the project was launched in 2007.

Pentads have turned out to be an excellent choice. Generally speaking, it is a sensible amount of territory to explore in 2–5 hours. Even at the latitude of South Africa, which stretches to 34°S, they are almost square, with sides of 9.2 km north to south and 8.3 km east to west (but this depends on how far south you are). It takes five minutes to drive across one at 120 km/hour. Usually, most habitats within a pentad can be accessed within a few hours, where "habitat" is defined as places to visit to add a few more species to the bird list. In many pentads, especially in rural, arid areas, species are added quite slowly towards the end of two hours minimum period of intense observation. Even so, the discipline of spending a minimum of two hours in a pentad works.

SABAP2 protocol psychology

The basic psychology of atlasing is vastly different between the naïve protocol and the SABAP2 protocol. With the naïve protocol there is no sense of a predetermined start and finish. With the SABAP2 protocol there is a precise sense of entering the pentad, and starting fieldwork. The atlaser remains within the boundaries of the pentad and strategizes to record as many different species within the pentad as feasible. It is this strategizing which introduces the concept of gamification (persuasive motivation) into the fieldwork. There is a sense of working against the clock which has the positive effect of sharpening the senses. There is a precise moment of leaving the pentad and completing the fieldwork. This generates a sense of accomplishment which the naïve protocol cannot provide.

The SABAP2 protocol provides a precise mechanism for evaluating how much fieldwork has been done in a region: measured in either number of checklists submitted for the pentad or the region, or numbers of hours spent doing intensive fieldwork. The pentad system spawns the crucial concept of the "coverage map", clearly showing the areas for which no data exists, shown as pentads with no data on the coverage map. It is only the SABAP2 protocol which could generate the vulgar slang which is used to describe the process of doing the first checklist for a pentad: "I did a virgin pentad." However unsatisfactory the metaphor, it is profoundly motivating.

The coverage map is the ever-visible tool for planning atlasing expeditions to poorly covered regions (Figure 1). It has even been used as the planning guide for family holidays. An unvisited pentad surrounded by pentads with data is perceived by atlasers as a blot on the landscape, and it becomes a personal challenge to work out the access strategy to reach it.

The protocol is fundamental to enabling initiatives to reach successive levels of fieldwork coverage in a region. The leading example of this has been the series of challenges in the "Four Degrees" region centred on Gauteng, and stretching into Mpumalanga, Limpopo, North-West and Free State; these 576 pentads have had challenges to various levels of coverage. The "Four Degrees Blue" challenge ended in June 2016, when every pentad had had 11 full-protocol checklists made in it, and it turned "light blue" on the coverage map (Ainsley 2016).

Statistical analyses enabled by the SABAP2 protocol

The SABAP2 protocol is designed with statistical analyses in mind. In particular, five standard analysis methods are applicable. All have well-understood properties:

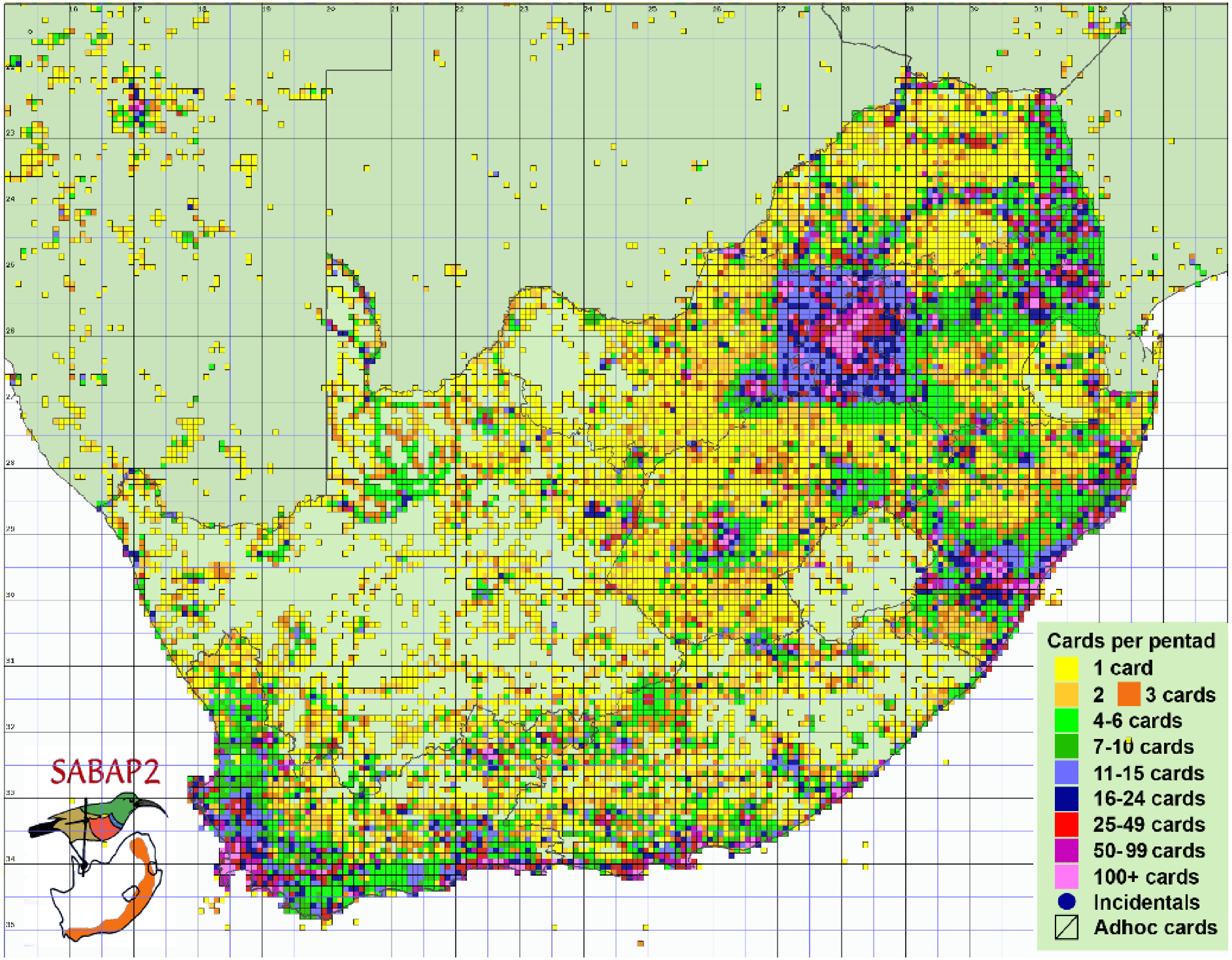

Figure 1. The SABAP2 coverage map, as at 30 July 2016, when 75.8% of the original area had at least one checklist, and 33.0% had foundational coverage of four checklists.

(1) generalized linear model; (2) occupancy modelling; (3) survival analysis; (4) the Griffioen transformation from “reporting rate” to relative abundance; (5) the modified species diversity index. All of these analyses can be performed using “off-the-shelf” statistical software systems for data analysis. None of these can be applied to naïve protocol data in the same direct way with which they can be applied to data collected with the SABAP2 protocol. We will consider statistical method each in turn.

(1) Generalized linear model

Arguably, the most important development in statistics in the last few decades of the 20th century was the generalized linear model, which is a family of models that can be used to relate a response variable to explanatory variables in a variety of contexts (McCullagh & Nelder 1989, Crawley 1993). In the bird atlas context, the response is the number of times Capped Wheatear was recorded on the 25, say, checklists made for a pentad, and the explanatory variables are factors which possibly “explain” its occurrence there: annual rainfall, average temperature, percentage of the pentad which is agricultural, the human population of the pentad, etc. Each member of the generalized linear model has a standard statistical distribution for the response variable, and a “link function” which describes how the explanatory variables are incorporated into the statistical model to explore the way they are related to the response variable.

There is one member of this family of models which is pre-adapted for the SABAP2 protocol. This is the model for the binomial distribution, which is used for modelling the number of successes after a “process” has been repeated a fixed number of times. An example is the process of having a baby. If a family consists of four children, and “success” is defined as “the baby is a girl” then the binomial distribution enables the probabilities of zero, one, …, four girls in the family. If you have enough families, you can use the information to estimate the probability of success (“a girl”) in the process of having a baby. Intuitively, we believe that this probability is one half (or “50%”). But the theory of mathematical statistics say that the correct way to make this estimate from the actual data is to add up the total number of times the process has been repeated, to count the number of successes, defined as “the number of girls”, and to calculate the proportion of girls. In the way that the binomial distribution is applied to bird atlas data, all the families only have one child. So the estimate of the probability of a girl simplifies to (number of girls/number of families).

For the atlas, the “process” is doing the fieldwork for an atlas checklist and submitting it. Success might be defined as “the checklist contains Capped Wheatear”. The only options are success or failure. In any given pentad, the estimate of the probability of a Capped Wheatear being observed in the pentad is ((number of checklists with Capped Wheatear recorded)/(number of checklists)). This is precisely the definition of “reporting rate”.

When the “process” is having a baby, we believe that the probability of having a girl is fixed. It does not vary with latitude or longitude, nor with rainfall or temperature. It is the same in every pentad! But we do not believe this for the reporting rates of Capped Wheatears. We intuitively believe that this reporting rate depends on all sorts of factors, which statisticians call “explanatory variables”. The genius of the generalized linear model is that it provides the strategy for deciding which of the available explanatory variables are in fact related to the reporting rate for Capped Wheatears and describes the relationship in mathematical formulae.

This analysis approach can be applied to data collected using both the SABAP1 and SABAP2 protocols. The pioneer application of the generalized linear model to bird atlas data was a conference paper presented by Underhill et al. (1995) and the ideas were further developed by Vincent Parker for his MSc (Parker 1995). He related species occurrence in an eighth degree grid cell of the Swaziland

Bird Atlas (Parker 1994) to environmental variables (Parker 1996, 1999), and he paved the way for the subsequent analyses that have been performed on SABAP1 and SABAP2 using the generalized linear model.

(2) Occupancy modelling

Occupancy modelling, as its name suggests, is concerned with the probability that a cell is occupied by at least one individual of a species at a point in time (MacKenzie et al. 2006). Occupancy modelling thrives on the concept of a fixed spatial area in which repeated surveys are performed, with fixed (or at least measured) amount of fieldwork. The SABAP2 protocol serves the data needs of occupancy modelling well. Occupancy modelling is made vastly more difficult, if not impossible, by data generated by the naïve protocol.

In applications to the SABAP1 and SABAP2 bird atlas the fixed spatial area is either the quarter degree grid cell or the pentad. Occupancy modelling approaches to the SABAP1 and SABAP2 data are being developed by the analysis team at SEEC (Centre for Statistics in Ecology, Environment and Conservation) at UCT. A selection of publications using occupancy modelling suggests that this is a rapidly growing approach to the analysis of bird atlas data collected with the SABAP2 protocol (Broms 2013, Broms et al. 2014, 2016, Peron & Altwegg 2015, Peron et al. 2016).

Occupancy modelling, as applied to the bird atlas data, explicitly models it as two components: the detection process, and the real biological process.

(3) Survival analysis

The SABAP2 protocol uses ordered checklists, i.e. checklists with the species recorded in the order in which they were seen (and the cumulative number of species that were recorded at the end of each hour of intensive fieldwork). It is intuitively obvious that there is more information in the ordered checklist than in a simple list of species seen, but it is less than obvious how to extract that information.

Soon after SABAP2 started, Elizabeth Kleynhans did an internship which provided enough insight to assure us that we know how to extract the information in an ordered checklist. The results were not publishable at the time, because the project had not been running long enough to produce a paper suitable for journal submission. The method has not subsequently been revisited, and we are in the process of rectifying this.

An important analysis in medical statistics is to decide which of several treatments for a disease, e.g. cancer, offers the best long-term survival rates for patients. The statistical method developed for this scenario is therefore called “survival analysis” although it is applicable in contexts in which there is no sense of “survival.” In the simplest application, there is just one treatment and the question is “What is the survival rate for this treatment?” The survival rate is most easily grasped as the “average time till death.” The statistician receives from the clinic a sample of data values which consist of the times till death of a number of patients treated for a disease. There is also always a puzzling and frustrating set of additional data. These are the patients who were alive when they last visited the clinic; but then they failed to keep their next appointment. The technical expression to describe them is “lost to follow-up”. They might have moved cities, sought further treatment elsewhere, or they might have died. The clinic cannot find out what happened to them, hence the term “lost to follow up”. The question that survival analysis answers is how best to incorporate these two classes of data into the value we are searching for: “average time to death”: (1) time to death, and (2) the known survival time until “lost to follow up”.

With the SABAP2 protocol, we can estimate for each species recorded, the time when the record was made. For example, if 20 species were seen in the first hour, we could assume that the records were made at five-minute intervals – we can do better than that (and with BirdLasser we can get the actual times), but the method does not rest on this value being completely accurate. So we can work out, at least to a good approximation, how long it took to see the first Capped Wheatear, the one recorded on the checklist for the pentad. To connect the statistical method to the medical analysis described in the previous paragraph, the question being asked is: "How long after I start atlasing can I survive until I see my first Capped Wheatear?" Effectively, you die when you see your first Capped Wheatear. But even in a pentad in which Capped Wheatears are present, there are observers who do not see a Capped Wheatear during the hours while they were atlasing. They ought to have atlased for a few more hours! These checklists are analogous to the patients who were "lost to follow up". Exactly the same method can be used to estimate, for a pentad, how many hours you can survive on average until you see a Capped Wheatear.

In 2008, Elizabeth Kleynhans and I called this number the "time-to-see" index. We did enough analyses to show that it was more closely related to the concept of abundance than reporting rate is. Re-starting the analyses for the time-to-see index is an important priority for the Animal Demography Unit.

(4) The Griffioen transformation

Peter Griffioen, in a 2001 PhD from La Trobe University, developed a method for transforming reporting rates into relative abundance (Griffioen 2001). He had, at his disposal, a data set from a project called the Australian Bird Counts (Ambrose 1991). Citizen scientists in Australia had undertaken a project in which they counted the number of birds of each species in plots. The fieldwork had been done in each plot, a fixed spatial area, multiple times. He simplified this data into reporting rates, i.e. the proportion of counts on which the species had been recorded (regardless of the number of individuals seen), so he had available to him both counts and reporting rates. Two decades earlier, a German statistician had developed precisely the mathematical theory which was needed, the relationship between population density in a grid and the probability that the species was not recorded in the grid cell (which is one minus the reporting rate) (Nachman 1981), and it was Peter Griffioen who spotted that it applied to bird atlas data.

Reporting rates increase with relative abundance, but not on an arithmetic scale. Doing subtraction with reporting rates is a meaningless operation. What the Griffioen (2001) transformation enables us to do is to grasp, in terms of relative bird abundance, what a change in reporting rate from 10% to 20% means compared with a change in reporting rate from 80% to 90%.

(5) Modified species diversity index

In all bird surveys, the number of species in an area never reaches an asymptote (Harrison & Martinez 1995). For example, in southern Africa as a whole, the number of species that needs to be included in each successive edition of *Roberts' Birds of Southern Africa* increases steadily. Almost all of the additional species are not regular members of the southern African bird community. The total number of species recorded in a region is referred to as "species richness". This is a particularly crude measure, because every species counts equally, regardless of whether it is abundant, rare or a vagrant in the region. Clearly, what is needed is a measure that takes relative abundance into account in a way that vagrants have minimal impact on the value of the measure. Information theory provides a series of solutions to this problem.

The informal approach to understanding the solution, in the context of the full-protocol bird atlas data, runs like this. Take the species list from each checklist made for a pentad and “concatenate” them into one long string of species. Number the list from 1 to N, where N is the total number of records on all the checklists. The common species are in this list multiple times, and the vagrants only a few times. Now choose two numbers at random between 1 and N. Calculate the probability that you have hit the same species. If the pentad is in the Knersvlakte, where individual checklists are short and a core set of species occurs on almost every checklist, then the probability that you have hit the same species is relatively large. In contrast, if you are in the northern Kruger National Park, where there are many species, the probability of choosing the same species is relatively tiny. The mathematicians have developed the exact formulae to calculate these probabilities, and closely related ideas. In the biological literature, these formulae are called diversity indices.

Harrison & Martinez (1995) applied one particular index, the Shannon diversity index, to SABAP1 data, with promising results. In spite of the fact that the species richness for a grid cell kept climbing with more and more checklists, the diversity index reached an asymptote, and remained stable. Even more stunning was the fact that the diversity index reached 93% of its final value after five checklists, 96% after 10, and 97% after 15 checklists.

This approach can only be applied if there is a strict spatial system for the checklists, and works best if each checklist is as complete as the observer can achieve. It does not work with the naïve protocol, because there is no concept of making comprehensive lists for a predefined area, and there is a tendency to report the rarer species at the expense of the more common species.

The approach developed by Harrison & Martinez (1995) has not been used to the extent which it deserves. The only other application has been by Underhill et al. (1998) who showed that, as a community, the insectivorous Palearctic migrant passerines concentrated in the arid thorn-savannas of the Limpopo valley and westwards into the Kalahari basin, whereas the resident insectivorous passerines were concentrated in the more mesic savannas of eastern South Africa. The amount of rainfall in the areas occupied by the migrants is more unpredictable and more patchy than in the areas occupied by residents, and it is an advantage not to be breeding and confined to a territory, but instead to be mobile to exploit patchy resources as they become available.

Scoping the African bird atlas

By mid-2016, the protocol developed for SABAP2 was in use in seven southern African countries: Botswana, Lesotho, Mozambique, Namibia, South Africa, Swaziland and Zimbabwe. Farther north, it was adopted in Kenya and in Nigeria. By mid-2016, Kenya was closing in on 10% coverage of its 6,817 pentads, and Nigeria had passed the 1% coverage mark of its 11,141 pentads within four months. In addition, full protocol checklists had been submitted from many other countries, and the Animal Demography Unit had received requests from many other African countries which were keen to make a start with the SABAP2 protocol.

There are approximately 400,000 pentads in Africa. By mid-2016, the ADU had checklists for nearly 16,000 of them. Thus, about 4% coverage had already been achieved for the continent of Africa. The 400,000 African pentads in total seems an overwhelming number, but when it is decomposed into national totals, each becomes a manageable task. Many countries have about 10,000 pentads, so that each 100 pentads visited adds 1% coverage to the national coverage statistic.

A target of 2,000 additional pentads added to coverage per year for the first five years of an African bird atlas project seems achievable.

With solid funding the target could be moved upwards to 4,000 per year. Coverage could be in the range 25% to 50% within this five-year period. It is a big challenge, but it is feasible.

This essay ends close to where it started. Its aim was to describe the analytical power enabled through having data collected using the SABAP2 protocol. This strength of analysis is reinforced first by the sociological reality that the fieldwork to collect data according to this protocol is fun, and then by the technological development of the cellphone app, BirdLasser, that removes the two difficult components of the SABAP2 protocol, map-reading and data entry.

In 2010 we said it for the World Cup. Now we say it for the bird atlas. This time for Africa.

Acknowledgements

Several people made invaluable suggestions which have been incorporated into this essay: Colin Jackson, Dieter Oschadleus, Etienne Marais. Jerome Ainsley, Megan Loftie-Eaton and Michael Brooks. SABAP2 is a partnership of SANBI, the main funder of the coordination hub of the project, BirdLife South Africa, and the Animal Demography Unit, which undertakes the project execution and management.

References

Ainsley J 2016. The SABAP2 "Four Degrees Blue" project: the challenge to obtain at least 11 checklists in 576 pentads. Biodiversity Observations 7.36: 1–7. Available online at http://bo.adu.org.za/content.php?id=232.

Ambrose S 1991. Australian Bird Count instruction sheet. Royal Australian Ornithologists' Union, Melbourne.

Broms KM 2013. Using presence-absence data on areal units to model the ranges and range shifts of select South African bird species. Unpubl. PhD thesis. University of Washington, Seattle, Washington, USA. Available online at

https://digital.lib.washington.edu/researchworks/bitstream/handle/1773/24102/Broms_washington_0250E_12315.pdf

Broms KM, Hooten MB, Johnson DS, Altwegg R, Conquest LL 2016. Dynamic occupancy models for explicit colonization processes. Ecology 97: 194–204.

Broms KM, Johnson DS, Altwegg R, Conquest LL 2014. Spatial occupancy models applied to atlas data show Southern Ground Hornbills strongly depend on protected areas. Ecological Applications 24: 363–374.

Crawley MJ 1993. GLIM for Ecologists. Blackwell Scientific Publications, Oxford.

Griffioen P 2001. Temporal changes in the distributions of bird species in eastern Australia. Unpubl. PhD thesis, La Trobe University, Australia.

Harrison JA 1993 Southern African Bird Atlas Project and its relevance to nature conservation. Unpubl. MSc thesis, University of Stellenbosch.

Harrison JA, Allan DG, van Hensbergen HJ 1994. Automated habitat annotation of bird species lists – an aid to environmental consultancy. Ostrich 65: 316–328.

Harrison JA, Martinez P 1995. Measurement and mapping of avian diversity in southern Africa: implications for conservation planning. Ibis 137: 410–417.

Harrison JA, Underhill LG 1997. Introduction and methods. In Harrison JA, Allan DG, Underhill LG, Herremans M, Tree AJ, Parker V, Brown CJ (eds) The atlas of southern African birds. Vol. 1: Non-passerines. Johannesburg: BirdLife South Africa: xliii–lxiv.

Larsen R, Holmern T, Prager SD, Maliti H, Røskaft E. 2009. Using the extended quarter degree grid cell system to unify mapping and sharing of biodiversity data. African Journal of Ecology 47: 382–392

MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey LL, Hines JE 2006. Occupancy estimation and modelling: inferring patterns and dynamics of species occurrence. Academic Press, San Diego.

McCullagh PA, Nelder J 1989. Generalized linear models. 2nd edition. Chapman & Hall, London.

Nachman G 1984. A mathematical model of the functional relationship between density and the spatial distribution of a population. Journal of Animal Ecology 50: 453–460.

Parker V 1994. Swaziland bird atlas 1985–1991. Websters. Mbabane, Swaziland.

Parker V 1995. Statistical analysis of bird atlas data from Swaziland. MSc, University of Cape Town. Available online at

http://open.uct.ac.za/bitstream/handle/11427/20195/thesis_sci_1995_parker_vincent.pdf

Parker V 1996. Modelling the distribution of bird species in Swaziland in relation to environmental variables. Ostrich 67: 105–110.

Parker V 1999. The use of logistic regression in modelling the distribution of bird species in Swaziland. South African Journal of Zoology 34:39–47.

Peron G, Altwegg R 2015.Twenty-five years of change in southern African passerine diversity: non-climatic factors of change. Global Change Biology 21: 3347–3355.

Peron G, Altwegg R Jamie GA, Spottiswoode CN 2016 Coupled range dynamics of brood parasites and their hosts responding to climate and .vegetation changes. Journal of Applied Ecology. doi: 10.1111/1365-2656.12546.

"Ragnvald" 2005–2015. QDGC. Wikipedia: the Free Encyclopedia. Wikimedia Foundation. https://en.wikipedia.org/wiki/QDGC, consulted on 26 June 2016.

Underhill LG 2016. Research publications and postgraduate theses which have been largely dependent on data from the Southern African Bird Atlas Projects. Biodiversity Observations in press.

Underhill LG, Brooks M 2016. SABAP2 after nine years, mid 2007–mid 2016: coverage progress and priorities for the Second Southern African Bird Atlas Project. Biodiversity Observations 7.35: 1–18. Available online at http://oo.adu.org.za/content.php?id=230.

Underhill LG, Herremans M, Navarro RA, Harrison JA 1998. Where do Palearctic migrant passerines concentrate in southern Africa during the austral summer? In Spina F, Grattarola A (eds) Proceedings of the first meeting of the European Ornithologists' Union. Biologia e Conservazione della Fauna 102: 168–174.

Underhill LG, Prys-Jones RP, Harrison JA, Martinez P 1992. Seasonal patterns of occurrence of Palaearctic migrants in southern Africa using atlas data. Ibis 134 Supplement 1: 99–108.