Development of an early warning index to predict the fragmentation of bird distributions

2011

Summary: Birds are precious bio-indicator of ecosystems health. Studying their distribution throughout time would help anticipate biodiversity erosion. This study relies on the data collected in order to create the Second Southern Bird Atlas Project (SABAP2). The latter is an initiative that aims to learn about bird distributions and also to raise awareness about biodiversity and particularly about avifauna.

The dataset is substantial and allows statistical analysis. Three different methods are developed to measure the fragmentation of bird distribution. Two of them appear to give similar results: The Pearson's product-moment correlation coefficient and the Moran's I. The data collection has been done by volunteers, therefore the protocol cannot be entirely standardized. It follows that shortcomings exist and that they must be taken into account. This is only partially done in the two selected methods which gives them different advantages and limits. The Moran's I method can probably be improved to get a more comprehensive approach.

The newly created index gives an information regarding the level of fragmentation, it does not directly say whether or not the species is in difficulty. Indeed, we expect different levels of fragmentation between specialist and generalist species. The results regarding Southern African species surprisingly shows that habitat non specialist species have some of the roughest distribution. Moreover, being most of the time abundant and common species, the data collection is fairly correct.

Elsa Bussière
Msc Thesis

ADU: Animal Demography Unit
Zoology Department,
University of Cape Town,
South Africa

ENSAIA
Ecole Nationale Supérieure d'Agronomie et des Industries Alimentaires,
Nancy,
France

Université de Nancy,
France

Table of contents

Introduction 3

The Animal Demography Unit 4

Background 4

Objectives and Assignments 5

1. Material & Method 6

1.1. Brief overview 6

1.2. Data collection 6

1.2.1. Experimental Design 6

1.2.2. Hypothesis 7

1.3. Development of an index 7

1.3.1. Reporting rate 7

1.3.2. The idea of occupancy probability and detection probability 8

1.3.3. Different calculations for different methods 9

1.3.3.1. Distances and correlation coefficients between reporting rates 9

1.3.3.1.1. The distance between reporting rates, our first attempt $[\theta^s]$ 11

1.3.3.1.2. Pearson product-moment correlation coefficient, as a second attempt $[\rho]$ 12

1.3.3.2. The Moran's I, our third and last attempt $[I]^s$ 12

1.3.4. Ornithologists' expertise 14

2. Results and critiques of the methods 15

2.1. Comparison of the three different methods 15

2.1.1. Distance between reporting rates 16

2.1.2. Pearson product-moment correlation coefficient 17

2.1.3. Moran's I 18

2.2. Results for the Pearson product-moment correlation coefficient in Gauteng, its Northern Part and its Southern Part 18

2.2.1. Results in Gauteng 19

2.2.2. Results in the Northern part of Gauteng 20

2.2.3. Results in the Southern part of Gauteng……………………………….22

2.2.4. North vs South…………………………………………………………………...23

2.3. Ornithologists' survey……………………………………………………………..23

3. Discussion…………………………………………………………………………………24

Conclusion………………………………………………………………………………………30

Acknowledgement…………………………………………………………………………….31

References……………………………………………………………………………………...31

Summary………………………………………………………………………………………...32

Appendix………………………………………………………………………………………...33

Introduction

While the biodiversity keeps decreasing, scientists and conservationists try to develop early warning systems to predict any change affecting the ecosystem balance, before the damages become substantial or worse: non reversible. Birds colonized almost every part of the world where they are daily, present and visible. Birds depend highly on environmental conditions where the smallest change would affect their distribution. As a consequence, the avifauna is a precious bio-indicator of biodiversity and therefore, of the ecosystem health.

The Second Southern African Bird Atlas Project (SABAP2) is an updated and refined version of the First Southern African Bird Atlas Project (SABAP1) which ran from 1987 to 1991. This project depends a lot on the work of about 920 passionate citizen scientists who daily collect numerous data sets all over the region. SABAP2 started in 2007 and provides accurate data regarding the avifauna distribution which offers great perspectives for an ecosystem health assessment.

The SABAP2 data are used to give us insights into the continuity of bird distributions which enables us to: 1) define the more or less generalist or specialist nature for each species, 2) analyse distributions at a very small scale and therefore, to assess their texture by developing computational tools to measure their "smoothness". Species that have no particular habitat requirements occur with a rate which remains more or less constant from one place to another, whereas habitat specialists are likely to have fragmented distributions.

This study will help us assess the smoothness of the habitat-non-specialist species distributions and help us provide an "early warning system" telling us when a species is starting to get into difficulty, and the distribution starts to fragment. Hopefully, this might provide information that could help predict future conservation necessities.

The new index will come into its own when SABAP2 has about 10 years of data, and we can compare the “roughness index” through time.

The Animal Demography Unit

Formerly called the Avian Demography Unit, the ADU is a research unit of the University of Cape Town. It was built on the collaboration between the South African Bird Ringing Unit (SAFRING) and the Southern African Bird Atlas Project (SABAP) as well as on the association with BirdLife South Africa.

Initially part of the Department of Statistical Sciences, the ADU has grown far beyond its starting point and was transferred to the Department of Zoology.

The mission of the Avian Demography Unit is to contribute to the understanding of animal populations, especially population dynamics, and thus provide input to their conservation. The ADU team achieves this through mass participation projects, long-term monitoring, innovative statistical modelling and population-level interpretation of results. The emphasis is on the curation, analysis, publication and dissemination of data.

Background

SABAP1 provided a ‘snapshot’ of the distribution and relative abundance of birds in southern Africa and was an exemplary example of a project which improved the public understanding of science, and which played a key role in science education. SABAP2 plans to build on the results of SABAP1 in order to produce an improved atlas and contribute in a greater way to biodiversity conservation.

SABAP1 showed some weaknesses that SABAP2 attempts to reduce. Indeed, SABAP2 reduced the geographical sampling units from Quarter-Degree Grid Cells (15’*15’) to pentad grid cells (5’*5’), a finer scale to obtain more detailed information on the occurrence of species and that will give us a clearer and better understanding of bird distributions. The quarter degree grid gave us an excellent broad brush picture of bird distributions, but has been demonstrated to be too course for the kind of fine-scale planning decisions which are needed for the conservation of biodiversity.

In order to strengthen the quality of the data collected and give us more possibilities for data analysis, SABAP2 tightened the protocol constraints to reduce the bias created by the observation process:

The observer effort has been standardized as much as possible so that the data can be used in a more comparable way. The minimum observation time period is two hours. The two hour minimum is motivated by the concept that two hours in a pentad with uniform habitat in low-diversity areas is probably enough to locate most species. For pentads with more varied habitats and in high-diversity areas the minimum time could take as long as 10 hours to locate all possible species in a complex grid cell. Atlasers are encouraged to make a special effort to try and cover all the different major habitats in each pentad and to do their initial surveys in the morning (an hour after sunrise) in favorable conditions (no rain, still wind conditions), when birds are the most active. The maximum time period is five days.

SABAP2 had the following primary objectives:

- To measure the impact of environmental change on southern African birds through a scientifically rigorous and repeatable platform which uses standardized data collection on bird distribution and abundance.
- To provide a basis for increasing public participation in biodiversity data collection, and public awareness of birds, through large-scale mobilization of citizen scientists.
- To provide information that can be used to determine changes in the distribution and abundance of birds since SABAP1.

Objectives and Assignments

My personal objectives were: 1) to enhance my skills in the field of statistics, data analysis and computer-modelling, 2) to work on a research project that aims to answer wildlife conservation questions.

My mission was to develop a computational tool to assess the texture of bird distributions on using the SABAP2 data, and to apply this tool to the Southern African bird species. I also carried out another study on bird migration that I will not develop in this report. The reason why I chose to talk about the first study is that I worked on it from the very beginning contrarily to the second one which was the extension of a previously established method in order to apply it to a larger range of species.

We can define sub-objectives that I reached along the internship (Appendix 1 also gives you a simplified calendar of my work):

- Bibliographical work in order to learn more about the SABAP1 and 2 protocols, their strengths and weaknesses
- Development of a computational tool in collaboration with several members of the team, mainly my supervisor, Professor Leslie Gordon Underhill. This step is the one that took most of the time because we did not know where we were going with it. Other researchers' knowledge and experience have been extremely useful because they gave me an idea of where to look at and a real discussion set up between us as and when we got more results. The tool is the achievement of an iterative process relying also on previous studies carried out within the ADU and that takes into account the weaknesses of the database. (Appendix 2 is a scheme of the interactions between the team members).

- Confirmation of the coherence of the results by skilled ornithologists.
- Analysis of the results and discussion about their relevance, trying always to keep thinking critically.

The writing of the research article and of this thesis was done throughout the internship.

1. Material and Method

1.1. Brief overview

The atlas region for SABAP2 includes the countries of South Africa, Lesotho and Swaziland. The project is based on numerous inventories carried out all over this area, thanks to the devotion of hundreds of volunteers.

In this part, we will see in detail how the data are collected and we will emphasize the fact that there is always a difference between the theory and what it is practically occurring in the field. We will also quote and quickly discuss the hypothesis that we made to allow statistical analysis and comparisons. Then, the different methods that have been used to assess bird distributions will be presented.

1.2. Data collection

1.2.1. Experimental Design

The experimental design is the theoretical and therefore, the ideal description of the experimentation.

[a] **Time frame:** Started in July 2007 to July 2011

Studied area: South Africa, Lesotho and Swaziland. (For this specific study, only data collected in the 4°G area were analysed. The 4°G area extends from west of the Pilanesburg to near Greylingstad and includes Gauteng and tracts of Limpopo, Mpumalanga, Free State and North West Province).

Experimental units: The studied area is demarcated by a $5' * 5'$ grid that defines $24 * 24 = 576$ pentads, a square of $8 * 7.5\ \text{km}^2$. Every pentad is an experimental unit in which the survey is carried out.

h_i, is the number of different habitats found in the pentad "i".

[b] **Inventories:** Every pentad is surveyed at least 4 times.

For all the inventories, the h_i habitats are surveyed and for a total period of at least $2 * h_i$ hours and possibly until 5 days. No new inventory starts within 5 days following the start of the previous one.

All the inventories start in the morning, one hour after sunset with no rain and still wind conditions.

Ideally, this should have been repeated at night to get information about nocturnal species. This is hardly done and there is very little information about them, therefore, nocturnal species have not been taken into account in this study.

[d] **Atlasers:** They use binoculars and, if possible, scope to identify birds on sight and according to their calls.

[ec] **Repetitions:** Repetitions are done all along the year, during different seasons.

Data entry: Results are submitted by atlasers to the ADU on this following website: http://sabap2.adu.org.za/ and stored in the SABAP2 database[f]. Once stored, every inventory constitutes a dataset called a card.

1.2.2. Hypothesis

[a] There is no annual effect on bird distributions.
[b] Whatever the type of habitat, all the present bird species are identified after 2 hours of atlasing in an homogenous habitat.
[c] There is no seasonal effect.
[d] Atlasers are all equally skilled.
[f] Data were uploaded without making any mistakes.

1.3. Development of an index

This study will use statistical analysis to assess bird distributions. The more data we use, the stronger the analysis will be. Given that the country has not been equally investigated and covered, this study relies on the abundant data collected in the 4°G area. In this area, atlasers were encouraged to repeat the protocol several times before covering pentads outside the area. The 4°G challenge of getting four lists per pentad in this area was reached recently in May 2011. (We added a row of pentads around this area even if all the pentads did not have four lists because of the substantial quantity of inventories that have been done there.)

The reason that motivated this challenge is that previous studies focusing on occupancy models showed coherent results with a minimum of four checklists per pentad. This number of repetitions became the absolute smallest sample size for all sub projects in order to do relevant statistical analysis. Repeating the protocol several times on the same pentad is really valuable because it reduces the chance of having a “hole” in the distribution when the bird species is actually present.

1.3.1. Reporting rate

The 4°G area is divided into 676 pentads. Each time a pentad is covered, a new checklist, recording all the identified species, is created. For this study, the data were summarized into a database containing, for all the 676 pentads covering the 4°G area, the number of checklists (n) and, for each species, the number of checklists in which the species occurred (x). Which gives us a total of 9948 checklists recording 603 species.

However, only the species that occurred in at least 50% of the pentads were considered for this study, which reduces the total species number to 98.

We define for each pentad$_{(i,j)}$, the following ratio :

$$p^s{}_{i,j} = \frac{\text{number of checklists recording species s} = \ x^s{}_{i,j}}{\text{total number of checklists} \ = n_{i\,i}}$$

It is called reporting rate for species s in the pentad. Reporting rates are a way of extracting quantitative information from presence/absence data; observers did not count birds, they recorded the presence of identified species on checklists. However, reporting rates are not proportional to density (birds per hectare), but provide an index which fluctuates with changes in density.

We can plot for every species a colored map showing their distribution over the 4°G area. The SABAP2 manual gives you those maps for all different studied species. Thanks to those maps we can already have a fairly good estimate of the bird distribution continuity.

1.3.2. The idea of occupancy probability and detection Probability

As we mentioned earlier, the observed reporting rates depend on the biological and observation processes of the study. The effect of each process can be represented by a probability: the occupancy probability and the detection probability.

- The occupancy probability, $P_o{}^s$, is related to a species "s". Its value depends on the place, on the time of the year, and on the year. We assumed that there was no seasonal and annual effects which enable us to define $p_o{}^s$ as the probability for a species "s" to be present in a specific area, in our case, a pentad. $(p_o{}^s{}_{i,j})_{i,j\,\in(1:26)}$ are the occupancy probabilities in the 676 pentads for the species "s". This probability is not affected in any way by the protocol of the study. It is a process totally independent from the experimentation and the probabilities values are not assessable. It is easy to understand the link between the occupancy probability and the reporting rates. For example, let's say that for a species "s" in a pentad$_{(i,j)}$, $p_o{}^s{}_{i,j} = 0$, then $p^s{}_{i,j}$ cannot take any other value than "0". If $p_o{}^s{}_{i,j} > 0$, we cannot say anything about $p^s{}_{i,j}$, except that its value can possibly increase. Indeed, if a species is present, it still needs to be encountered and identified before being recorded in a card. This is part of the observation process and therefore, taken into account in the detection probability.

- The detection probability, $P_d{}^s$, is related to a species "s" and depends on many factors that will be detailed further in the discussion. For every card, $P_d{}^s$, is the probability to detect the present species "s" in a specific area, which is here again, a pentad. To be detected, a species needs to be encountered and then identified. In consequences, the detection probability is here again, the product of two probabilities:

- ${P_e}^{s}{}_{i,j}$, the probability to encounter the specie “s” in the pentad$_{(i,j)}$.
- ${P_{id}}^{s}{}_{i,j}$, the probability to identify the species “s” in the pentad$_{(i,j)}$.

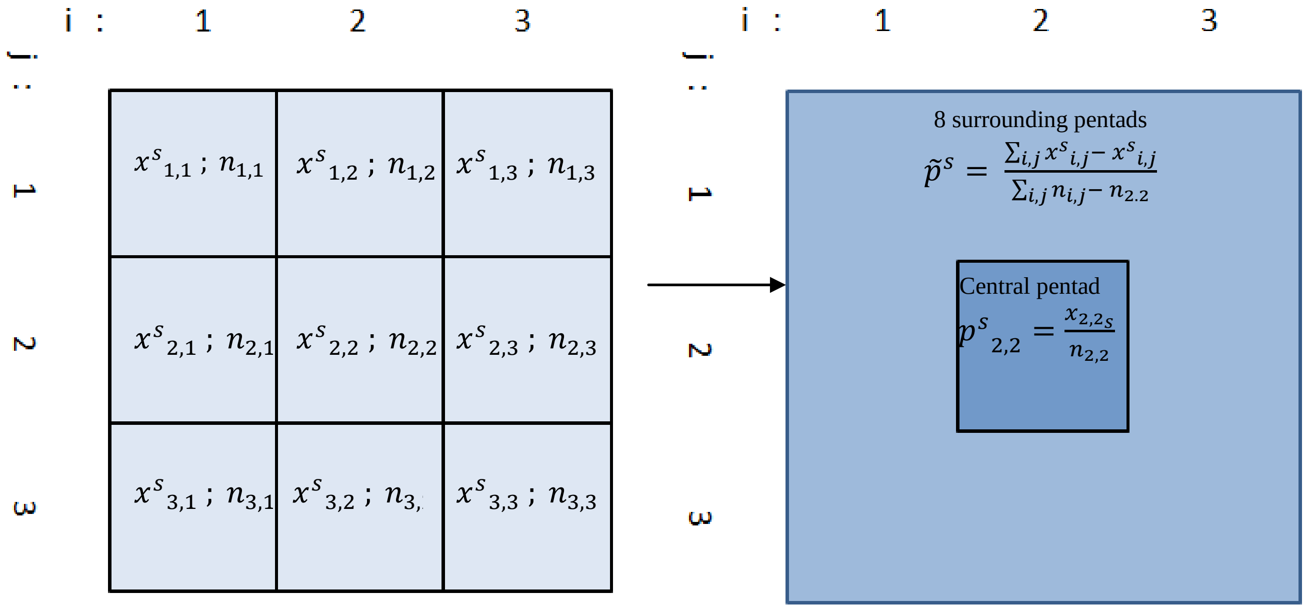
If each card was done in exactly the same conditions, we could define ${P_d}^{s}{}_{i,j}$ as the probability to record the present species “s” in the pentad$_{(i,j)}$. And $({p_d}^{s}{}_{i,j})_{i,j\,\in(1:26)}$ would be the detection probabilities in the 676 pentads. However, there is an important bias that needs to be taken into account, which is that cards are done by different atlasers with different ${P_{id}}^{s}{}_{i,j}$.

1.3.3. Different calculations for different methods

The development of a computational tool is an iterative process that led to different methods, more or less relevant according to the biological question that we aim to answer. As more results come, some leads are abandoned, new ones are explored and new questions emerge. The non anticipated questions sometimes remain unanswered cause some information has not been collected for this purpose while the data collection occurred.

1.3.3.1. Distances and correlation coefficients between reporting rates

The studied area is a square made up of $26 * 26$ pentads. We define 576 units called blocks that are the combination of 9 adjacent pentads, a central pentad surrounded by 8 others. Let’s consider one block. $p^{s}{}_{2,2}$ and $\tilde{p}^{s}$ are respectively the reporting rates of a the species “s” in the central pentad and the 8 surrounding ones.



j : \ i :	1	2	3
1	$x^{s}{}_{1,1}$; $n_{1,1}$	$x^{s}{}_{1,2}$; $n_{1,2}$	$x^{s}{}_{1,3}$; $n_{1,3}$
2	$x^{s}{}_{2,1}$; $n_{2,1}$	$x^{s}{}_{2,2}$; $n_{2,2}$	$x^{s}{}_{2,3}$; $n_{2,3}$
3	$x^{s}{}_{3,1}$; $n_{3,1}$	$x^{s}{}_{3,2}$; $n_{3,}$	$x^{s}{}_{3,3}$; $n_{3,3}$

8 surrounding pentads

$$\tilde{p}^{s} = \frac{\sum_{i,j} x^{s}{}_{i,j} - x^{s}{}_{i,j}}{\sum_{i,j} n_{i,j} - n_{2.2}}$$

Central pentad

$$p^{s}{}_{2,2} = \frac{x_{2,2s}}{n_{2,2}}$$

Figure 1 : A block and its reporting rates : $p^{s}{}_{2,2}$ and $\tilde{p}^{s}$. The 4°G area is made up of $24 * 24 = 576$ blocks

The smoother the bird distribution is, the closer $p^{s}{}_{2,2}$ and $\tilde{p}^{s}$ should be.

The reporting rate of a specific species in a specific pentad$_{(i,j)}$, $p^{s}{}_{i,j}$is a random variable with a probability distribution. To define the latter, we first need to consider another random variable, ${X^s}_{i,j}$, the number of times the species is recorded among all the cards done in the pentad.

${X^s}_{2,2}$ and $\tilde{X}^s$ are discreet random variables with binomial distributions, as defined below.

Let's consider a block again. $n_{2,2}$ is the number of checklists in the central pentad, and $\tilde{n}$, the total number of checklists in the 8 surroundings pentads: $\tilde{n} = \sum_{i=1}^{3} \sum_{j=1}^{3} n_{i,j} - n_{2,2}$.

Let ${\pi^s}_{i,j}$ be the true probability that the species "s" is recorded in each card made in the pentad$_{(i,j)}$. It takes into account the occupancy probability of the species "s" and the detection probability of the observer. We assume it is constant, from one card to another.

In the central pentad: ${X^s}_{2,2} \sim B\left(n_{2,2}, {\pi^s}_{2,2}\right)$ & in the related block: $\tilde{X}^s \sim B(\tilde{n}, \tilde{\pi}^s)$ with ${\pi^s}_{2,2}$& $\tilde{\pi}^s$ respectively, the true probability to record the species "s" in the central pentad and in the 8 surrounding ones.

We use normal distributions to approximate the binomial distributions and get continuous random variables. (Proschan, M.A.,2008).

$$X_{2,2_s} \sim N\left(n_{2,2}{\pi^s}_{2,2}\,, n_{2,2}{\pi^s}_{2,2}\left(1-{\pi^s}_{2,2}\right)\right) \qquad \& \qquad \tilde{X}_s \sim N\left(\tilde{n}\tilde{\pi}^s, \tilde{n}\tilde{\pi}^s\ (1-\ \tilde{\pi}^s\ )\right))$$

We assume that within the same block, for a specific species, the occupancy and detection probabilities are the same in every place, which means that the probability to record the species in the central pentad, ${\pi^s}_{2,2}$, and the one in the 8 surrounding ones, $\tilde{\pi}^s$ are equal.

On dividing ${X^s}_{2,2}$ and $\tilde{X}^s$ by the number of cards in each area, that are respectively $n_{2,2}$ and $\tilde{n}$, we obtain the distributions of the observed reporting rate of the central pentad, ${p^s}_{2,2}$, and the observed reporting rate of the 8 surrounding ones, $\tilde{p}^s$.

$${p^s}_{2,2} \sim N\left({\pi^s}_{2,2},\ \frac{{\pi^s}_{2,2}\left(1-{\pi^s}_{2,2}\right)}{n_{2,2}}\right) \qquad \& \qquad \tilde{p}^s \sim N(\tilde{\pi}^s,\ \frac{\tilde{\pi}^s(1-\ \tilde{\pi}^s)}{\tilde{n}})$$

Those reporting rates are not the true reporting rates, because the variables X defined above, took the detection probability into account. The true reporting rates only depend on the occupancy probability of the species.

The idea to assess the continuity of the distribution of the species "s" is to look at how close ${p^s}_{2,2}$ & $\tilde{p}^s$are in average over the 4°G area. To do so, we will first experiment two different methods. The first one depends on the sum of distances between ${p^s}_{2,2}$& $\tilde{p}^s$ in all the 576 blocks and a second one relying on the correlation coefficient of the two following datasets $({p^s}_{2,2})_{b=1:576}$ & $(\tilde{p}^s)_{b=1:576}$. With b, the 576 different blocks.

To be able to identify any variation in reporting rates from place to place, we first assume that within the same block, for a specific species, the occupancy and detection probabilities are the same in every place, which means that the probability to record the species in the central pentad, $\pi^{s}{}_{2,2}$, and the one in the 8 surrounding ones, $\tilde{\pi}^{s}$ are equal. In consequences, the probability distribution of the continuous random variable $p^{s}{}_{2,2} - \tilde{p}^{s}$ in each block, is:

$$p^{s}{}_{2,2} - \tilde{p}^{s} \sim N\left(0, \dot{\pi}^{s}(1 - \dot{\pi}^{s}{}_{b})\left(\frac{1}{n_{2,2}} + \frac{1}{\tilde{n}}\right)\right) = N(\mu, \sigma^{2})$$

Under this assumption, our best estimate of $\dot{\pi}^{s}$, is $\dot{p}^{s}$ the reporting rate of the species "s" in the block as a whole. $\dot{p}^{s} = \frac{\sum_{i,j} x^{s}{}_{i,j}}{\sum_{i,j} n_{i,j}}$ (Cf. Figure 1).

Thus, we can write the variance formula as follows: $V^{s}{}_{b\in[\![1,576]\!]} = \dot{p}^{s}(1 - \dot{p}^{s})\left(\frac{1}{n_{2,2}} + \frac{1}{\tilde{n}}\right)$

Whatever the chosen method, we must keep in mind that all the blocks do not have the same number of checklists and thus, do not provide us with the same reliable reporting rates. To take this bias into account, all the 576 couples of data points will be weighted by a coefficient, $\omega^{s}{}_{b\in[\![1,576]\!]}$, depending on the variance: $\omega^{s}{}_{b} = \frac{{V^{s}{}_{b}}^{-1}}{\sum_{b=0}^{576} {V^{s}{}_{b}}^{-1}}$.

The formula is not adapted to every situation, for instance when $\dot{p}_{s} = 0$. "Zero" or "one" reporting rates are often especially observed in areas where few checklists have been collected. This problem is resolved by using the empirical logistic transform (Cox & Snell 1989):

We approximate the term $\dot{p}^{s}(1 - \dot{p}^{s})$ by $\frac{(\dot{x}^{s}+1)(\dot{n}-\dot{x}^{s}+1)}{(\dot{n}+1)(\dot{n}+2)}$, with $\dot{x}^{s}$ & $\dot{n}$ respectively, the number of records of the species "s" and the number of cards in the block. Calculate the two and see how different it is.

1.3.3.1.1. The distance between reporting rates, our first attempt [θ^{s}]

Our first attempt was to define for each species an index, θ^{s}, depending on the distances, in every blocks, between the reporting rate of the central pentad, $p^{s}{}_{2,2}$, and the reporting rate of the surrounding area, defined by the 8 surrounding pentads, $\tilde{p}^{s}$.

$$\theta^{s} = \sum_{b=1}^{576} \omega^{s}{}_{b} * [\left(p^{s}{}_{2,2} - \tilde{p}^{s}\right)^{2}]_{b}$$

Given that this distance can be either positive or negative, it is squared to enable us to sum the distances for every block and get a comparable value for each species.

The results are given on page 16, figure 2.

1.3.3.1.2. Pearson product-moment correlation coefficient, as a second attempt $[\rho]$

Our second attempt was to define for each species the Bravais-Pearson linear correlation coefficient, ρ^s, between $(p^s{}_{2,2})_{b\in[\![1,576]\!]}$ and $(\tilde{p}^s)_{b\in[\![1,576]\!]}$. Here again, every couple of data points is weighted with the $\omega^s{}_{b\in[\![1,576]\!]}$ as defined above.

Some species only have high (or low) reporting rates, which means that when we draw the scatter $(p^s{}_{2,2})_{b\in[\![1,576]\!]} \sim (\tilde{p}^s)_{b\in[\![1,576]\!]}$, all the points are concentrated in the right upper (or left lower) corner. This trend impacts the value of the weighted correlation coefficient and therefore, biases our assessment.

Eventually to get around this difficulty, the data are transformed with the logit function as follows:

${p_2}^s{}_{2,2} = log\frac{a}{1-a}$ (idem for $\tilde{p}^s$that becomes $\tilde{p}_2{}^s$). Given that this log transformation does not allow certain values like 0 and 1, we first slightly change $p^s{}_{2,2}\ and\ \tilde{p}^s$ as follows: $a = \frac{p^s{}_{2,2}-0.5}{1.1} + 0.5$ (idem for $\tilde{p}^s$). Whereas $p^s{}_{2,2}$ and $p^s{}_{2,2}$ vary from 0 to 1, ${p_2}^s{}_{2,2}$ and $\tilde{p}_2{}^s$ vary from -3 to 3.

We noticed that several species are absent in large areas covering some of the blocks entirely, where $p^s{}_{2,2}\ and\ \tilde{p}^s$ are equal to 0. This equality between $p^s{}_{2,2}\ and\ \tilde{p}^s$ makes the correlation coefficient high overall, and the distribution in the 4°G area smoother than it should be. To solve this bias, we excluded all the blocks where the reporting rate is nought from the database that was used to calculate the correlation coefficient. Therefore, for every species we have a certain amount of data, according to the number of blocks in which they occur. Let B_s be the set of blocks in which the species s occur.

$({p_2}^s{}_{2,2})_{b\in B_s}$ and $(\tilde{p}_2{}^s)_{b\in B_s}$can be considered to be two vectors in a B_s- dimensional space. $\overline{{p_2}^s{}_{2,2}}$ and $\overline{\tilde{p}_2{}^s}$ are respectively their means. We subtract the mean to every data point and we get two new vectors : $({p_2}^s{}_{2,2} - \overline{{p_2}^s{}_{2,2}})_{b\in B_s}$ and $(\tilde{p}_2{}^s - \overline{\tilde{p}_2{}^s})_{b\in B_s}$.

Eventually, we can determine the cosine of the angle, α, defined by those two vectors, which is the correlation coefficient that we are looking for.

$$\rho^s = \cos\alpha = \frac{\sum_{b=1}^{B_s}({p_2}^s{}_{2,2} - \overline{{p_2}^s{}_{2,2}})_b * \left(\tilde{p}_2{}^s - \overline{\tilde{p}_2{}^s}\right)_b)}{\sqrt{\sum_{b=1}^{B_s}{p_2}^s{}_{2,2} - {p_2}^s{}_{2,2})_b{}^2} * \sqrt{\sum_{b=1}^{B_s}\left(\tilde{p}_2{}^s - \overline{\tilde{p}_2{}^s}\right)_b{}^2}}$$

The results are given on page 17, figure 3.

1.3.3.2. The Moran's I, our third and last attempt $[I]^s$

Moran's correlation coefficient, I, is an extension of Pearson Product-moment correlation coefficient, ρ, to a univariate series. Indeed, ρ measures whether or not, on average, reporting rates in the central pentads and reporting rates in the eight surrounding ones, respectively $({p^s}_{2,2})_{b\in\llbracket 1,576\rrbracket}$ and $(\tilde{p}^s)_{b\in\llbracket 1,576\rrbracket}$, are associated. Whereas, I, will consider only one variable, ${p^s}_{(i,j)\in\llbracket 1,676\rrbracket^2}$, reporting rates in pentads.

Note that with ρ, $({p^s}_{2,2})_{b=k}$ & $({p^s}_{2,2})_{b=k'}$ are not associated since the pairs $(\,({p^s}_{2,2})_{b=k}\,,(\tilde{p}^s)_{b=k})_{k\in\llbracket 1,676\rrbracket}\,)$ are assumed to be independent of each other. (ibid. $(\tilde{p}^s)_{b=k}$ & $(\tilde{p}^s)_{b=k'}$)

In the study of spatial patterns and processes, we may logically expect that close observations are more likely to be similar than those far apart. Therefore, we associate a weight, $\delta_{i,j}$, to each pair $(({p^s}_{i,j})_{\substack{i=k\\j=l}}\,,({p^s}_{i,j})_{\substack{i=k'\\j=l'}})_{(k,k',l,l')\in\llbracket 1,676\rrbracket^4})$, in order to quantify this.

The simplest is, for these weights to take values 1 for close neighbors, and 0 otherwise. We also set :

- $\delta_{(i,j),(i'j')} = 0$ if $i = i'$ & $j = j'$
- $\delta_{(i,j),(i'j')} = \delta_{(i'j'),(i,j)}$

These weights are sometimes referred to as a neighboring function.

Moran's I is defined as follows:

$$I = \frac{N}{\sum_{i,j}^{\llbracket 1,26\rrbracket^2}\sum_{i'j'}^{\llbracket 1,26\rrbracket^2}\delta_{(i,j),(i'j')}} * \frac{\sum_{i,j}^{\llbracket 1,26\rrbracket^2}\sum_{i'j'}^{\llbracket 1,26\rrbracket^2}\delta_{(i,j),(i'j')}\left({p^s}_{i,j}-\overline{p^s}\right)\left({p^s}_{i'j'}-\overline{p^s}\right)}{\sum_{i,j}^{\llbracket 1,26\rrbracket^2}\left({p^s}_{i,j}-\overline{p^s}\right)^2}$$

Where $N = 676$, the number of spatial units indexed by i and j; $({p^s}_{(i,j)\in\llbracket 1,676\rrbracket^2})$ is the variable of interest; $\overline{p^s}$ is its mean; $\delta_{(i,j),(i'j')}$ is an element of a matrix of spatial weights.

Note that I takes on the classic form of any autocorrelation coefficient: the numerator term in each is a measure of covariance among the $\{{p^s}_{i,j}\}$ and the denominator term is a measure of variance.

The moments of I may be evaluated under H_0 either by assuming:

- Normality, assumption N
- Randomization, assumption R

In this study, Assumption R is chosen. Whatever the underlying distribution of the population(s), we consider the observed value of I relative to the set of all possible values which I could take on if the $\{{p^s}_{i,j}\}$ were repeatedly randomly permuted. There are $N!$ such values.

The Moran's I gives an information about the stability of reporting rate $\{{p^s}_{i,j}\}$, throughout the area, and about the spatial correlation. Three cases can be distinguished:

- Constant reporting rate
- Heterogeneous reporting rate with no spatial correlation

- Heterogeneous reporting rate with spatial correlation

Negative (positive) values indicate negative (positive) spatial autocorrelation. Values range from -1 (indicating perfect dispersion) to $+1$ (perfect correlation). Under the null Hypothesis of no autocorrelation, the expected value of I is not equal to zero but given by $I_0 = \frac{-1}{N-1}$.

The results are given on page 18, figure 4.

The 4°G area shows a distinct transition in habitat at the 26° South. Indeed, the Southern African geography and its relevance to birds show different altitudes, topography features and vegetation on either side (Harrison, J.A., 1997, Vol.1, Ixxi-xcvi).

Above the 26° South, is the grassland biome. The dominant vegetation comprises grasses, with geophytes and herbs also well represented (Low & Rebelo 1996). The altitude is lower, between 600 & 1200.

Beneath the 26°, is the woodland (savanna) biome. It is defined here as having a grassy understorey and a distinct woody upperstorey of trees and tall shrubs (Rutherford & Westfall 1986). The altitude is higher, between 1200 & 1800 (Harrison, J.A., 1997, Vol.1, Ixvi).

Therefore, we can expect to have two different patterns of bird distributions in the North and in the South. The Pearson product-moment correlation coefficient method and the Moran's I method have been used to diagnose different patterns in bird distributions in the North and the South of the 4°G area. We also realize a non parametric test of Wilcoxon to see if there is a significant change overall.

The results are respectively given on page 20, 21 and 22, figures 6, 7 and 8. You can also see Appendix 5, on page 36.

1.3.4. Ornithologists' expertise

We provided several ornithologists with the list of the 98 studied bird species and asked them to class them into five groups: strict generalist species, partial generalist species, species that are found in half of the habitats, partial specialist species and strict specialist species. Then, we looked at how well, their division into groups coincided with the statistics.

2. Results and critiques of the methods

All the calculations were done with the statistical software R: R 2.13.1

The results for the different previously quoted indexes, will be plotted twice against two variables:

- The total reporting rate: it is always related to a species (s) and a specific area (A). It is the sum of all the records of the species in question in A, over the sum of all the cards, in A. Here again, data in blocks that have no records for the species in question, haven't been taken into account in the calculation, to avoid large areas where the species is absent (c.f. 1.3.3.2. The Moran's I, our third and last attempt $[I]$, p12). Let $(B_s)_A$ be the set of block in the area A, in which the species s occur.
  The formula of the total reporting rate in (A), is as follows:

$$p^s{}_{tot_A} = \frac{\sum_{b \in B_s} x^s{}_b}{\sum_{b \in (B_s)_A} n^s{}_b}$$

- The number of occupied pentads: $Op^s{}_A$. In Gauteng, it varies from 338 to 676. (We only look at the species occurring in at least 50% of the pentads)

As far as the migrant species are concerned, we first thought of selecting only the data that have been collected while the species are present, to not underestimate their reporting rates. However, the calculation that we have done, does not enable us to work on a dataset for which we do not have information for every pentad. The reality is that some pentads have not been atlased at all while the migrant species were present, which gives no chance for the species to be recorded.

The two scatters will give similar results given that the correlation between the two is quite strong, especially regarding species with a large range distribution. We can expect those species to be habitat non specialist and therefore, the species of interest in this study. (c.f. appendix 3, page 35). Nevertheless, some species have high total reporting rate with a low total number of occupied pentads and vice versa. In this case, the two scatters can give different information. Examples will be given further.

2.1. Comparison of the three different methods

In this paragraph, we will plot the new index $(\theta, \rho \; or \; I)$ against the total reporting rate only.

2.1.1. Distance between reporting rates

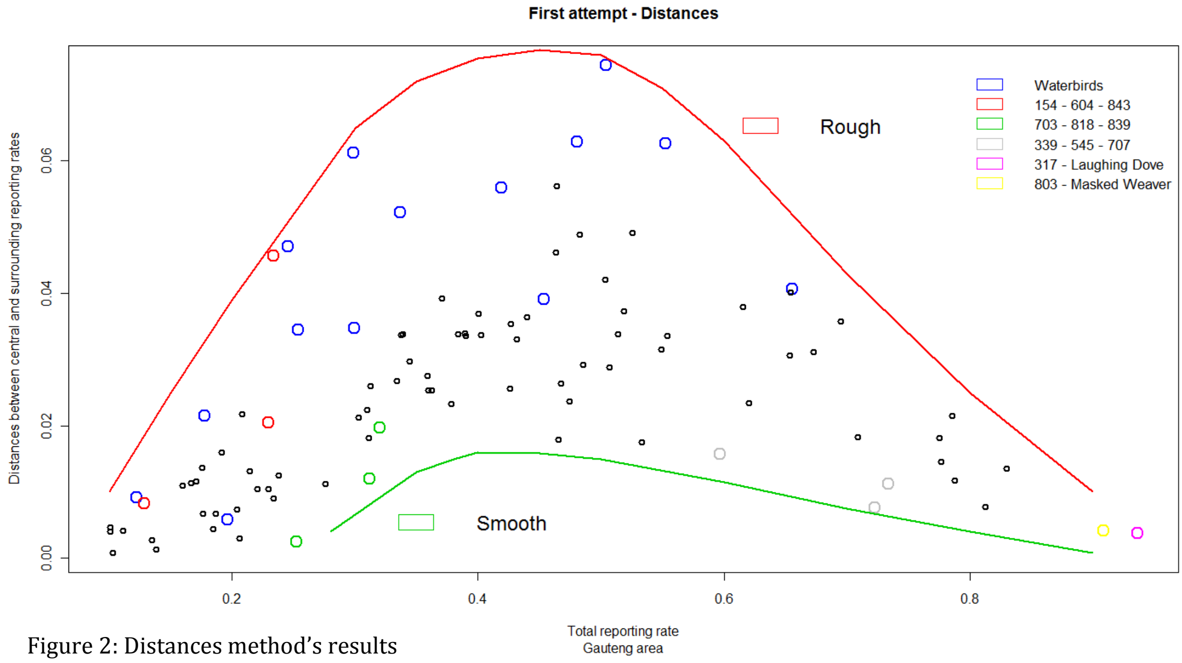


Figure 2: Distances method's results
The rest of the legend is given in appendix 4, on page 34

Note that the Distances method is a measure of roughness. The larger thêta is, the rougher the distribution is.

The points' pattern in Figure 1, let us think that the species with the lowest and highest total reporting rates have the smoothest distributions, as they have the lowest theta values. Because we couldn't explain it with a biological process, we assume that this trend is unlikely and that we probably missed something.

Our first observation was that waterbirds, known for being habitat specialist species and inclined to have rough distributions, are all concentrated on the upper part of the triangle, designing a curve. We then assumed that the closer to the red curve the point is, the rougher the distribution is. On the contrary, the closer to the green curve the point is, the smoother the distribution is. This hypothesis was confirmed after going through different colored maps of different species.

In conclusion, this method doesn't enable us to compare bird distributions according to the theta values only, we also need to take into account the total reporting rates. Indeed, only species with similar total reporting rate are comparable; and as we will see further, it is not advisable to compare reporting rates from a species to another, unless they have very similar abundance and conspicuousness.

2.1.2. Pearson product-moment correlation coefficient

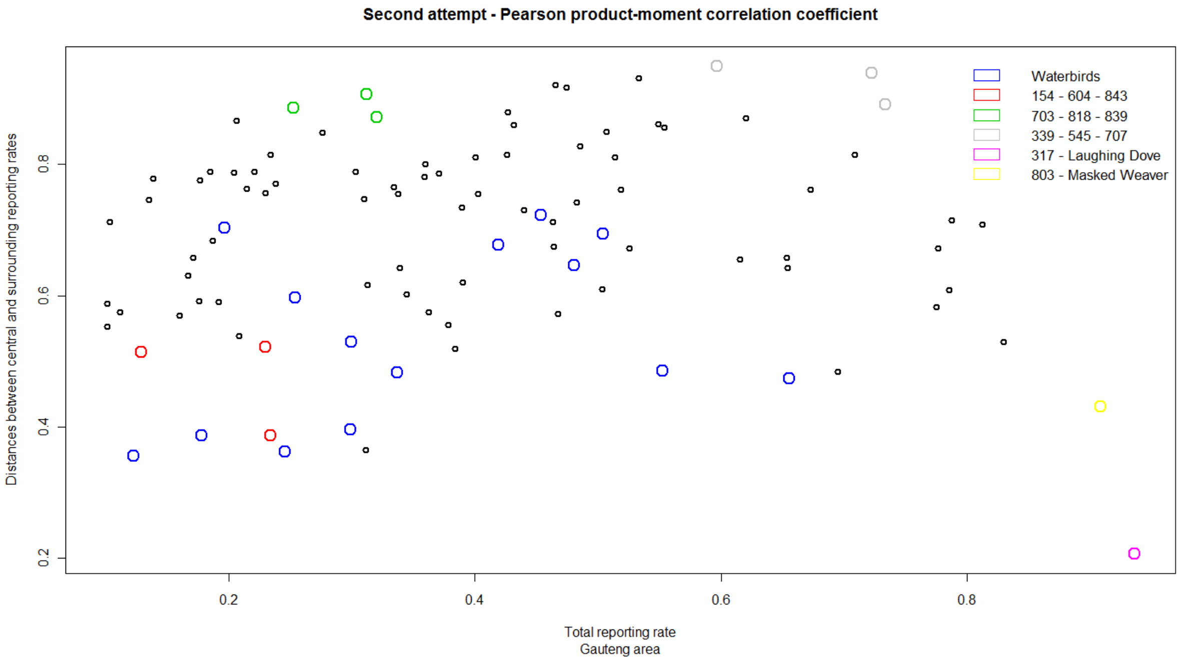


Figure 3: Pearson product-moment coefficient method's results
The rest of the legend is given in appendix 4, on page 34

Note that the correlation coefficients method is, contrarily to the previous one, a measure of smoothness. The closer to "1" the correlation coefficient is, the smoother the distribution is.

The first thing to say is that Figure 3 strengthens the conclusions given about the method of distances. The species with the highest coefficients and therefore the species with the smoothest distributions, are the species that we find at the bottom of the triangle in Figure 2, close to the green curve.

We also find the waterbirds mostly at the bottom in Figure 2.

2.1.3. Moran's I

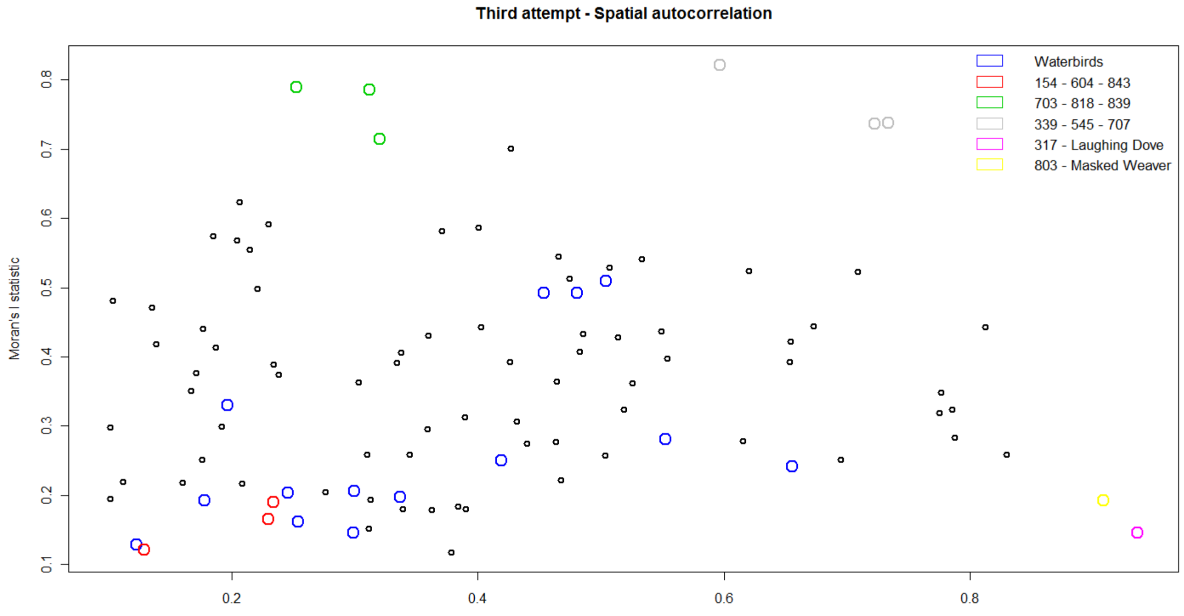


Figure 4: Moran's I method's results
The rest of the legend is given in appendix 4, on page 34

The first thing to notice is that the results given by the Moran's I method are very similar to the ones given by the Pearson product-moment correlation coefficient method. The Moran's I method considered all of the 8 pentads surrounding the central pentad to be single units giving single information, whereas in the Pearson method, we considered the 8 surrounding pentads to be a unique area. Considering this, the Moran's I method seems to be more accurate. However it doesn't take into account the number of repetitions done in each pentad, which is the case with the Pearson product-moment correlation coefficient method.

2.2. Results for the Pearson product-moment correlation coefficient in Gauteng, its Northern Part and its Southern Part.

In this paragraph, we will look at different scatters and at how the variables evolve in three different areas : Gauteng – the northern part (above the 26° South) – the southern part (underneath the 26°South). We will pay particularly attention to the species that have a high total reporting rate and a high total number of occupied pentads (in the defined areas). The following results are for the Pearson product-moment correlation coefficient only. The Moran's I results are very similar. The relative position of the species in the graphs are the same, only the values differ slightly making the points more scattered.

2.2.1. Results in Gauteng

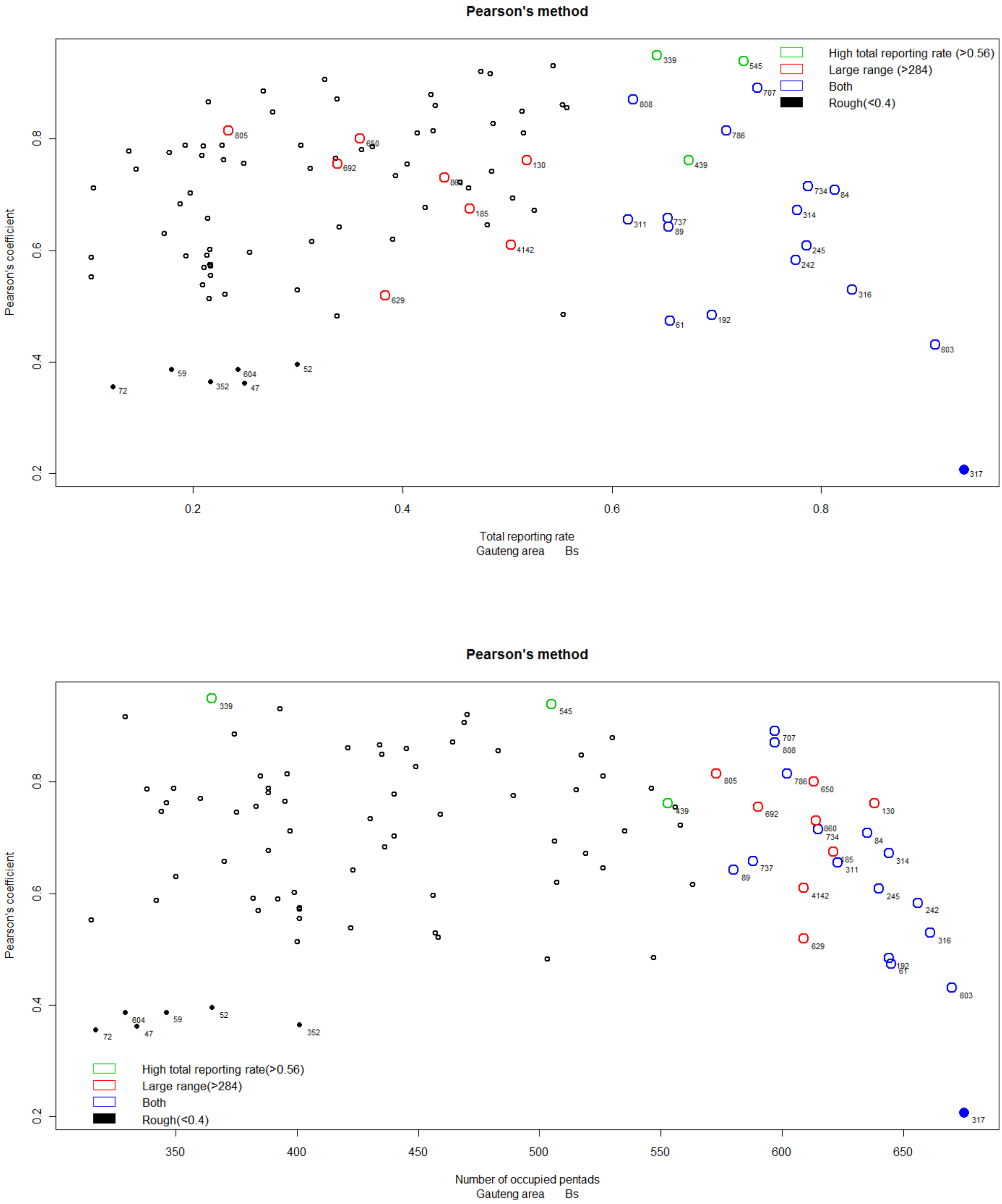


Figure 5: Pearson product-moment coefficient method's results in Gauteng
Top: Pearson's coefficient against the total reporting rate
Bottom: Pearson's coefficient against the number of occupied pentads

In the whole Gauteng, ρ^svaries from 0.2065 (317 – Laughing dove – *Streptopelia senegalensis*) to 0.9497 (339 – Grey Go-away-bird – *Corythaixoides concolor*).

Species with a relatively high abundance but a low total reporting rate (805 – 650 – 692) can be either :

- Species covering a large range with a low density
- Species covering a large range with a high density but that are inconspicuous and therefore not recorded.

The Red Billed Quelea (805 – *Quelea quelea*) is conspicuous and probably not really abundant.

Species with a high total reporting rate but a small number of occupied pentads in the whole Gauteng (339 - Grey Go-away-bird – *Corythaixoides concolor*) are probably abundant species in a specific habitat only.

Species with a high total reporting rate and a large range (317 - 61 - 192… in blue on the maps) are common conspicuous species. The analysis is therefore reliable. Among those, one, the Laughing dove (317 – Streptopelia senegalensis) has a rough distribution: $\rho^{317} = 0.2065$.

2.2.2. Results in the Northern part of Gauteng

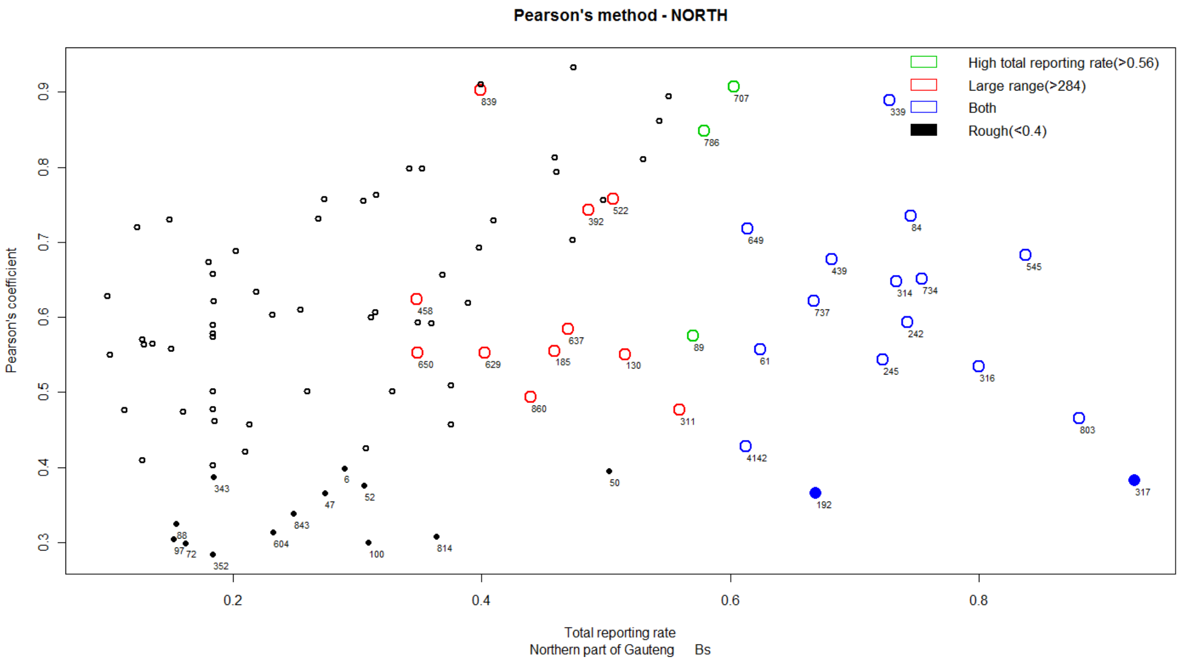


Figure 6: Pearson product-moment coefficient method's results in the Northern part of Gauteng
Pearson's coefficient against the total reporting rate

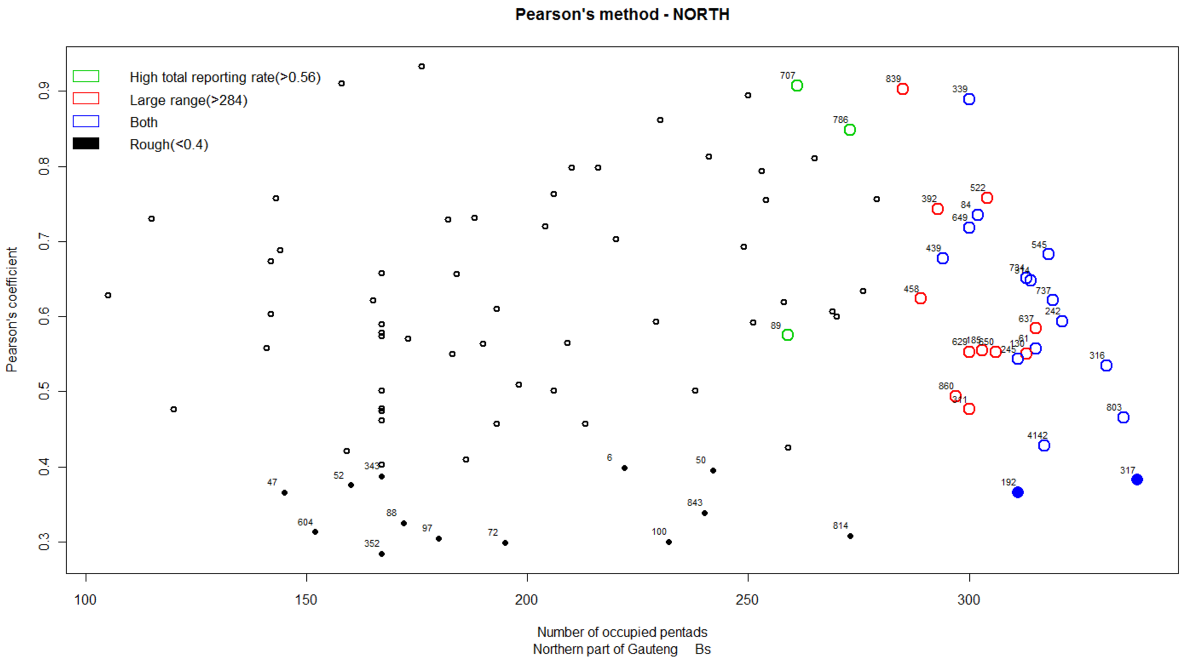


Figure 7: Pearson product-moment coefficient method's results in the Northern part of Gauteng Pearson's coefficient against the number of occupied pentads

In the whole Northern part of Gauteng, ρ^svaries from 0.2843 (352 – Cuckoo Diderick – *Chrysococcys caprius*) to 0.9330 (581 - Cape Robin Chat – *Cossypha caffra*).

The Grey Go-away-bird (339) is now a species covering a large range with a high total reporting rate. It is a species occurring chiefly in the North in savanna habitat.

Two species have a large range with a high total reporting rate and a rough distribution:

- Laughing Dove (317 – *Streptopelia senegalensis*)
- Helmeted Guineafowl (192 – *Numida meleagris*)

2.2.3. Results in the Southern part of Gauteng

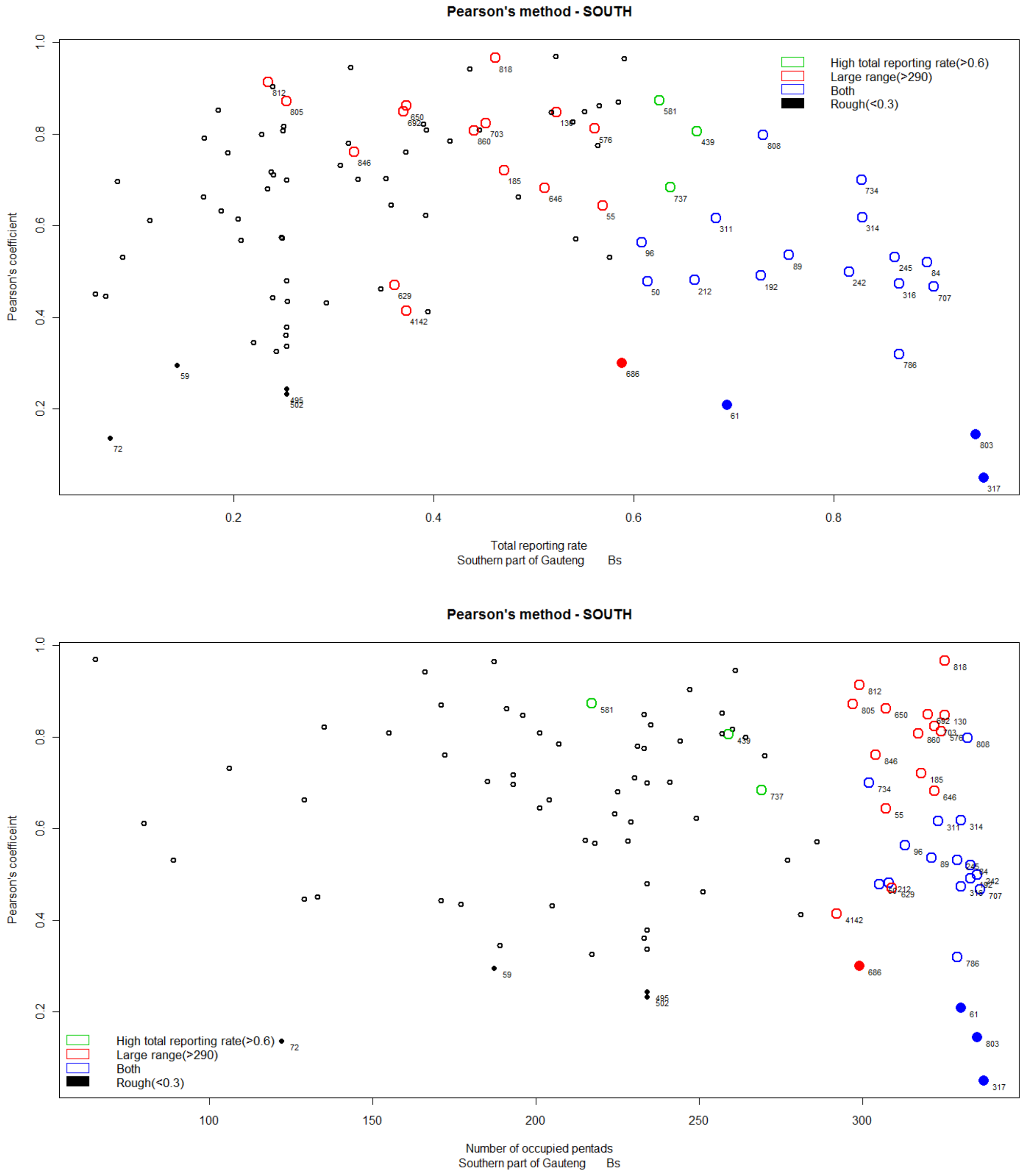


Figure 8: Pearson product-moment coefficient method's results in the Southern part of Gauteng
Top: Pearson's coefficient against the total reporting rate
Bottom: Pearson's coefficient against the number of occupied pentads

In the whole Southern part of Gauteng, ρ^s varies from 0.0497 (317 – Laughing dove – *Streptopelia senegalensis*) to 0.9698 (339 – Grey Go-away-bird – *Corythaixoides concolor*).

The Cape Robin Chat (581) is a species covering a medium range with a high total reporting rate. Given that this species did not appear in the previous scatters (Gauteng and North) we can conclude that it is a species that has a higher total reporting rate in the South but that still covers a few pentads.

Three species have a large range with a high total reporting rate and a rough distribution:

- Laughing Dove (317 – *Streptopelia senegalensis*)
- Southern Masked Weaver (803 – *Ploceus velatus*)
- Cattle Egret (61 – *Bubulcus ibis*)

The Cape Wagtail (686 – *Motacilla capensis*), covers a large range with a relatively low total reporting rate.

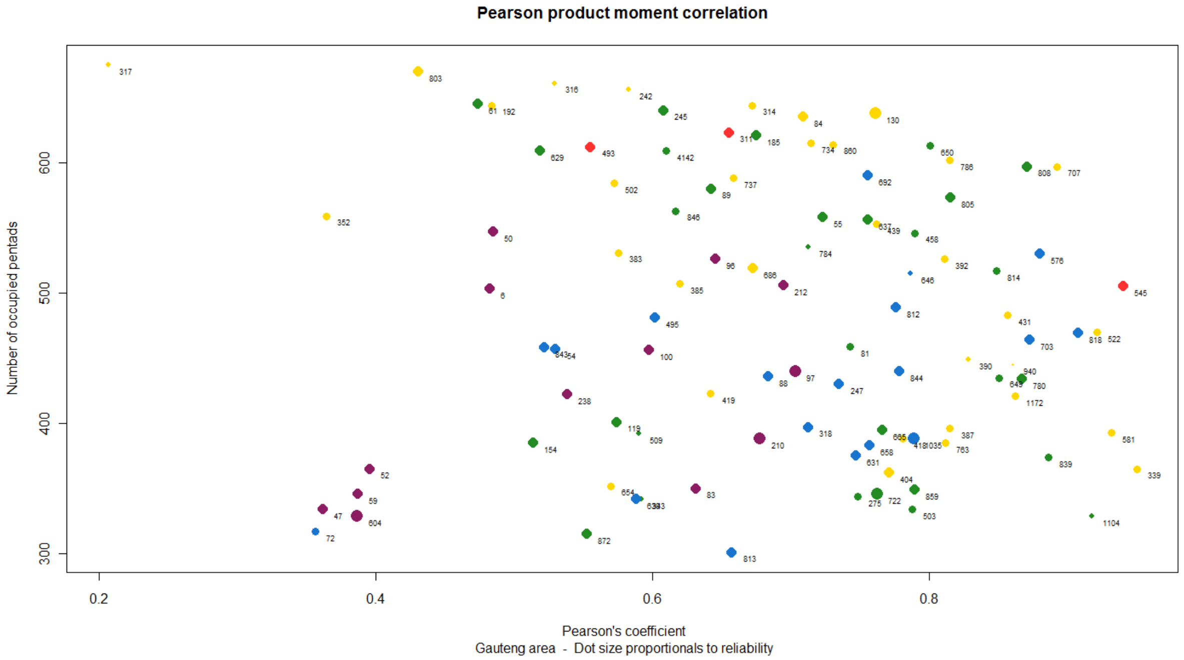
2.2.4. North vs South

Differentiating distribution in the North and the South was relevant that is why we also plot $\rho^{s}{}_{South}$ against $\rho^{s}{}_{North}$. The scatter is given in appendix 5, on page 36.

2.3. Ornithologists' survey

Four ornithologists answered to our request. They are all living in Gauteng and their answers rely on their experience in this area principally.

Some species have been easy to classify whereas, others have been put in very different groups. To take this into account, the scatter shows points with different sizes. The closer the answers were, the bigger the point is.



Within the same category, the results can be quite different. Although we can differentiate different zones, the overlap is significant.

These data can help create a typology for each of the four defined groups. (C.f. Appendix 7, page 36).

3. Discussion

This study relies on a database which tells us if a species is “present” or “absent” in a specific area. As we have already mentioned earlier, it is actually more informative than a simple binary variable because inventories have been repeated several times in the same area throughout the four years that the project has been running. Reporting rates are quantitative information which is not proportional to bird density but which provide an index that fluctuates with changes in density.

As every experimental studies, there is a gap between the experimental design and the actual experimentation that leads to uncertainties.

Many factors influence reporting rate, only one of which is relative abundance. Therefore its use as an index of relative abundance is subject to distortion.

The first hypothesis is to say that there is no annual effect and that no major changes occurred within those four years.

For this study, the gap is mostly related to the observation process which is obviously impossible to standardize perfectly. In consequence, the variability between “true” and estimated reporting rates is hardly assessable and constitutes the main weakness of this study. The trend is obviously an underestimation of the “true” reporting rate. Indeed, if a species is not reported, it could be truly absent or simply missed. Which means that the results depend on the detection probability, a probability depending itself on numerous factors more or less measurable.

Below, you will find the enumeration of the factors in question with a few examples :

If the bird is absent, then the detection probability is 0. If the bird is present, the probability to record it varies :

- ^{d} **Observer’s effects:**
  Affect $P_{id}{}^{s}{}_{i,j}$, the probability to identify the present species “s” in the pentad$_{(i,j)}$.
- Species which observers find easy to identify are recorded more frequently than those which are more challenging to identify, and the ability to identify can vary seasonally with plumage and behavior. (This is closely related to the species factor, discussed further).

- The fact that this ability to identify affects reporting rate means that the level of observer skill and experience will also affect this statistic. SABAP2 is a very ambitious project which aims to make as many inventories as possible of an area covering three countries. This kind of project cannot run without the involvement of citizen scientists who volunteer to collect the data all over the country. 4 years after its starts, 921 atlasers made more than 55252 inventories. Another objective of the SABAP2 project aims to educate about wildlife and especially about birds, which means that the more people get involved, the more effect the project has.

In consequence, atlasers are predominantly volunteers who decided to get involved in the Southern African Bird Atlas Project as their own initiative. They do not require any qualification to be able to get involved in it. In fact, anyone who is interested can register with the ADU.

Although atlasers are mostly passionate and skilled birders who are able to identify Southern African birds, they obviously all have different abilities. For example, not all of them are able to identify a species according to its calls. Variability can be expected, especially between novices and professional fieldworkers. If they have any doubt, they won't record the bird species and its reporting rate will be underestimated. In conclusion, standardizing the observer's skills is hardly feasible (as for the observer's effort that we will discuss further).

- The longer the time period spent compiling a checklist, the greater the likelihood of rare and secretive species being recorded. SABAP2 imposed a minimum of two hours atlasing in order to standardize as much as possible atlasers' effort. Nevertheless, some pentads are covered with numerous different habitats that increase the minimum effort time required to identify most of the present species. Even if atlasers are encouraged to cover all the different habitats, this minimum effort time is not always adapted to the type of pentad.

- Note another observer's effect which is their preconceptions. Some interesting results given by SABAP1 show that atlasers are influenced by their preconceptions regarding the kind of species that it is likely to find in the inventoried area. This effect prompted incoherent results in some of the bird distributions. (European Swift (Apus apus) and Black swift (*Apus barbatus*) distribution in SABAP1)

- In a specific pentad, if only one observer has been atlasing, the observer' effects might be even stronger. The mistakes that can be related to his weaknesses will be repeated in time instead of being compensate by another atlaser's skills.

- **Species:**

  Affect $P_{id}{}^{s}{}_{i,j}$, *the probability to identify, and* $P_{e}{}^{s}{}_{i,j}$, *the probability to encounter, the species "s" in the pentad*$_{(i,j).}$

- Given the same abundance, a relatively conspicuous species is recorded more frequently than an inconspicuous and secretive species. Generally speaking, reporting rates should only be used within species; comparisons based on reporting rate may be made between areas and seasons for one species, but not between species. Sizes, colors and calls are very different from one species to another. This variability makes them more or less easy to identify.

- The way in which individuals are grouped has an effect on reporting rates. Many species tend to flock in the non-breeding season and to disperse as pairs in the breeding seasons, e.g. the Blue Crane. Consider a species for which even single individuals are conspicuous. If the species forms non-breeding flocks or breeding colonies, and the birds become clustered, the probability of the species being encountered is smaller while it is clustered, thus reducing reporting rates. On the other hand, if single birds are cryptic and clusters are more readily observed and identified, reporting rates will be greater during the period of clustering.

- Note that when a species is recorded, it is not necessary making use of the habitat. It could only be flying over. Species moving a lot, moving far and being easily identifiable on flight, could appear to have a larger range distribution than it is actually. However, experienced atlasers assert that the proportion of species in a card being recorded while they were flying over is very low. Therefore, we can assume that this effect is negligible.

- **Abundance:**

Affect $P_e{}^s{}_{i,j}$, the probability to encounter the species "s" in the pentad$_{(i,j)}$.

- An abundant species will have better chance to be seen than a rare one. The latter has great chances to be missed even if it is actually present in the pentad.

- [b] **Habitat:**

Affect $P_e{}^s{}_{i,j}$, the probability to encounter the species "s" in the pentad$_{(i,j)}$.

- According to the type of habitat, visibility changes and can make inventories more or less complicated. In thick habitats, some birds are hidden in bushes or tall grasses. Here, there is a strong interaction with the species factors. Indeed, some species are habitat specialists and will always occur in the same type of habitat.

- Seasonal changes in habitat structure, such as reduced foliage in woodland during winter, can affect the conspicuousness of birds, and hence reporting rate.

- Some pentads have good networks of roads allowing access to all parts ; others have few roads making access to some important habitats difficult. Mountain tops, isolated wetlands and forest patches are difficult to reach in many pentads.

  In consequence, the proportions of identified species after two hours of atlasing varies from one habitat to another.

- SABAP2 aims to involve as many people as possible and, in order to not discourage potential volunteers, the SABAP2 team only encourages atlasers to explore all the different habitats occurring in the pentad, and does not make it obligatory. Therefore, the atlasers' effort cannot be standardized as much as it should be. The same kind of comments can be made about the "favorable conditions" that we talk about in the experimental design.

- [a] **Annual effect:**

Affect $P_e{}^s{}_{i,j}$, the probability to encounter the species "s" in the pentad$_{(i,j)}$.

- Saying that there is absolutely no variation from a year to another is obviously wrong, there always is. However, within a period of 4 years, we can expect not to see any significant change in the bird distributions

- ᶜ **Seasonality:**

Affect $P_{id}{}^{s}{}_{i,j}$, the probability to identify, and $P_{e}{}^{s}{}_{i,j}$, the probability to encounter, the specie "s" in the pentad$_{(i,j).}$

The effect can be considered at two different levels:

- The observation process: Throughout the year, a species can display a new behavior and look physically totally different. We often observe an increase in the activity and important changes of the plumage colors and shape that makes a species more or less conspicuous throughout the year. (Bishop's distribution in SABAP1). On the contrary, some species are really quiet during incubation, trying not to attract the attention of predators. During the non breeding season, some species look really similar and cannot be differentiated without taking any measurements. Many passerines undergo moult soon after breeding and then behave more quietly and secretively. In those situations, the species is not recorded and its reporting rate is underestimated.

- The biological process: bird species show significant movements to different extents, which can locally affect the species occupancy.

  ❖ Non migrant birds fly from places to others throughout the year (Example: dispersal after breeding). Sometimes, departures are compensated by arrivals but not always. Some are nomadic species and their movement patterns are closely related to the weather. Therefore, they are seasonal effects on bird distributions, even for non migrant species, which are more or less significant according to the species.

  ❖ Migrant species, show considerable seasonal effects on occupancy. Strict migrants are absent during several months and, with this protocol, their presence will be underestimated. This is something that has been taken into account in the calculation of the total reporting rate but not in the measure of the index for the reasons that have been previously quoted.

Seasonal variations have to be taken carefully. They can reflect the truth (Bird dispersing after breeding for example), and sometimes reflect a variation in conspicuousness (See the Red Bishop description in the species manual).

- Seasonality probably interacts with other factors such as habitat, but in far less proportions compared to its interaction with the species factor.

- The protocol does not give any directions about when during the year the inventories should be done. In consequences, we can locally miss information.

However, over the whole area, we have a fairly good idea of how reporting rates vary throughout the year.

- **e Arithmetic effect:**
- This relates to the number of checklists available for a given pentad, in other words, the denominator in the calculation of the reporting rate statistic. The larger it is, the more accurate the reporting rate is. If there is one checklist, the only possible values for the reporting rate are 0% and 100%. If there are two checklists, values of 0%, 50% and 100% are possible. If n is the number of checklists available, then, there are $n + 1$ possible values for the reporting rate. The implication is that if a relatively rare species is recorded in a pentad with few checklists, it will have a high reporting rate in the pentad, misleading impression of relatively high abundance in that area.
- A large number of different observers and of available checklists blend the observer's effects and increase the probability to encounter the species. However, the protocol does not impose a certain number of different atlasers in the same pentad. Actually, they are even sometimes influenced to regularly monitor the same site which is easy for them to access. This would give valuable data to better understand how bird distributions change from year to year. The SABAP2 protocol was not only done to answer one question but several and what could be a strength for a study, could be a weakness for another. Without taking the first row into account, 12 pentads have been inventoried by one atlaser only, so far.
- We mentioned that some pentads have not been inventoried when the migrant species were present (more or less from October to April). This probably affects the index significantly and might not allow us to conclude anything regarding the migrant species. Twelve strict migrant species have been studied and the maps showing the "holes" of information are given in the species manual.

- **f Data submission:**
- Among the thousands of submitted datasets, some will get mistakes (while copying from the datasheet for example).

The sources of mistakes are numerous. However, the accuracy increase substantially with the number of repetitions and observers. Excluding the row framing the studied area, all the pentads have at least four cards and very few have one observer only. (C.f. Species manual).

In this study, we pay particularly attention to the habitat-non-specialist species that are prone to have a large distribution range.

After analyzing the scatters of the Pearson product-moment correlation coefficient method, we noted that among those species with a large distribution range, some had a high total reporting rate which make them abundant conspicuous species. The results regarding those species are therefore fairly significant.

We noted down five of them that have rough distribution:

- **Laughing Dove (317 – *Streptopelia senegalensis*)**; which is rough in the three studied areas: Gauteng – Northern part and Southern part of it.
  Among the 676 pentads, only one has no record for this species (C.f. Manual). It is the pentad 2700W – 2815S. This pentad is in the row at the periphery of the studied area. It has one card. The observer that inventoried this pentad has also inventoried nine others where he has always recorded the Laughing Dove. In conclusion, the absence of record is not due to the observer's skill. Either the species is truly absent or simply missed. More cards need to be done to have a fairly good estimate of the reporting rate in this pentad.

- **Helmeted Guineafowl (192 – *Numida meleagris*)**; which is particularly rough in the Northern part of Gauteng.
  There are 32 pentads with no records for this species. 53% of them are in the peripheral row. Twenty four atlasers inventoried those pentads and appendix 7, on page 39 shows that they are all able to identify the species, except maybe the observer 13090. However, given that he only has atlased once, it could have simply not encountered the bird species.

- **Southern Masked Weaver (803 – *Ploceus velatus*)**; which is particularly rough in the Southern part of Gauteng.
  There are 6 pentads with no records for this species. 100% of them are in the peripheral row. Two atlasers inventoried those pentads and appendix 8, on page 40 shows that they are all able to identify the species.

- **Cape Wagtail (686 – *Motacilla capensis*)**; ibid. Southern Masked Weaver
  There are 157 pentads with no records for this species. 32% of them are in the peripheral row. Seventeen atlasers inventoried those pentads and appendix 9, on page 41 shows that they are all able to identify the species, except maybe the observer 11637. However, given that he only has atlased once, it could have simply not encountered the bird species.

- **Cattle Egret (61 – *Bubulcus ibis*)**; ibid. Southern Masked Weaver
  There are 31 pentads with no records for this species. 67% of them are in the peripheral row. Twenty two atlasers inventoried those pentads and appendix 10, on page 42 shows that they are all able to identify the species, except maybe the observer 11637 again. However, given that he only has atlased once, it couls have simply not encountered the bird species.

Ideally, those species should be replaced in their group (Generalist – Partial generalist – Half habitat – Partial specialist – Specialist) and see if their calculate index differ significantly from the ones given by the density function. (C.f. Appendix 6).

Only four ornithologist answered our request. Some species were classified in the Generalist group by one ornithologist and in the Partial specialist group by another one. Although it is a minority, it prompts mistakes that we cannot neglect.

As we have said earlier, the studied area can be relevant for a species and not for others, that are absent for example. This is also a variable that can have a significant effect. (C.f. Appendix 11, results are given for the Northern part of Gauteng only).

Conclusion

The Second Southern African Bird Atlas Project is an ambitious project that aims to answer different research questions and to raise awareness regarding avifauna. It relies on the work of volunteers, therefore the protocol cannot be perfectly standardized. The sources of mistakes are several, however, the accuracy of the data and the relevance of the study increase rapidly with the number of repetitions and observers.

Two methods among the three that have been suggested, appear to be meaningful in order to measure the smoothness of bird distribution and assess whether or not a species gets into difficulty: the Pearson's product moment correlation coefficient and the Moran's I. Each of them, has its own advantages and shortcomings. There are leads to set up a new one that will take into account more factors and therefore will be more complete. One possibility would be to modify the matrix of weights of the Moran's I method. Every link between two pentads would be weighted by another coefficient as follows: $\sqrt{\frac{1}{n_{i,j}}+\frac{1}{n_{i',j'}}}$ with $n_{i,j}$ & $n_{i',j'}$, respectively, the number of cards in the pentad$_{\text{i,j}}$ and in the pentad$_{\text{i'j'}}$.

The studied area has to be defined carefully according to the species we want to look at. If it contains different types of habitat, some species might occur in some of them only and it might not be relevant to consider the whole area for these species.

The index of smoothness gives an idea of relative fragmentation but does not tell us how serious the fragmentation is. As we have already mentioned earlier, we expect habitat specialist species to get lower values than generalist species and this does not mean that habitat specialist species are more in difficulty than the generalists'. Nevertheless, within the same category, it already gives a list of species that should be looked at more carefully: Laughing dove …… A species typology made according to this index could help orientate conservation plans.

The index will have a very important role in the future when more data will be collected and comparisons throughout time will be feasible. A this stage, we will look at the dynamic of the species population. New opportunities in order to say whether or not the distribution is prone to fragment will be available. This, strengthens the fact that SABAP2 should become an ongoing project.

The datasets do not enable us to suggest proper explanations regarding a possible fragmentation of some species. Although many factors could affect bird distributions, habitat plays a key role and more information about it would be extremely valuable. Enticing volunteers to collect information about habitats in ever pentad would help a lop understand the reasons behind the tendency.

Although SABAP2 had reduced the area of each experimental unit to 5'*5' in order to increase the accuracy of the atlas, some places have still never been explored because they are not reachable. Some of those places are gardens where a lot of birds occur. A new initiative has just been launched, "My bird patch". The objective is to motivate birders to regularly inventory their gardens and therefore extend the percentage of the atlased area.

Acknowledgement

I would like to thank the ADU for giving me the opportunity to be involved in wildlife research studies that are conservation based. I thank the whole team for its help and the wonderful working atmosphere. I particularly thank Leslie Gordon Underhill for trusting me and giving me responsibilities; but also for giving me the opportunities to get more experience in data analysis and fieldwork.

References

Cliff, A.D. & Ord, J.K., 1981, Spatial processes – models & applications, p13,14,42

Cox DR, Snell EJ 1989. Analysis of Binary Data, 2nd edn. Chapman & Hall, London

Harrison, J.A., Allan, D.G., Underhill, L.G., Herremans, M., Tree, A.J., Parker, V. & Brown, C.J.(eds). 1997. The atlas of Southern African Birds. Vol. 1: Non-passerines. BirdLife South Africa, Johannesburg.

Harrison, J.A., Allan, D.G., Underhill, L.G., Herremans, M., Tree, A.J., Parker, V. & Brown, C.J.(eds). 1997. The atlas of Southern African Birds. Vol. 2: Passerines. BirdLife South Africa, Johannesburg.

Low, A.B. & Rebelo, A.G. (eds) 1996. Vegetation of South Africa, Lesotho and Swaziland. Department of Environmental Affairs & Tourism, Pretoria.

Paradis, E., June 16, 2011, Moran's Autocorrelation Coefficient in Comparative Methods http://cran.r-project.org/web/packages/ape/vignettes/MoranI.pdf

Proschan, M.A., February, 2008, The American Statistician, Vol. 62, No.1

Ord, J.K., 1981, "Tests of significance using non-normal data" Geographical Analysis

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Rutherford, M.C. & Westfall, R.H., 1986. Biomes of Southern Africa – an objective categorization. Memoirs of the Botanical Survey of South Africa 54: 1-98

Harebottle, D.M., Smith, N., Underhill, L.G., Brooks, M., 2007, Southern African Bird Atlas Project 2, Instruction manual.http://sabap2.adu.org.za/docs/sabap2_instructions_v6.pdf

http://adu.org.za/pdf/Little_F_2003_PhD_thesis.pdf, page 12

Summary

Birds are precious bio-indicator of ecosystems health. Studying their distribution throughout time would help anticipate biodiversity erosion. This study relies on the data collected in order to create the Second Southern Bird Atlas Project (SABAP2). The latter is an initiative that aims to learn about bird distributions and also to raise awareness about biodiversity and particularly about avifauna.

The dataset is substantial and allows statistical analysis. Three different methods are developed to measure the fragmentation of bird distribution. Two of them appear to give similar results: The Pearson's product-moment correlation coefficient and the Moran's I. The data collection has been done by volunteers, therefore the protocol cannot be entirely standardized. It follows that shortcomings exist and that they must be taken into account. This is only partially done in the two selected methods which gives them different advantages and limits. The Moran's I method can probably be improved to get a more comprehensive approach.

The newly created index gives an information regarding the level of fragmentation, it does not directly say whether or not the species is in difficulty. Indeed, we expect different levels of fragmentation between specialist and generalist species. The results regarding Southern African species surprisingly shows that habitat non specialist species have some of the roughest distribution. Moreover, being most of the time abundant and common species, the data collection is fairly correct.

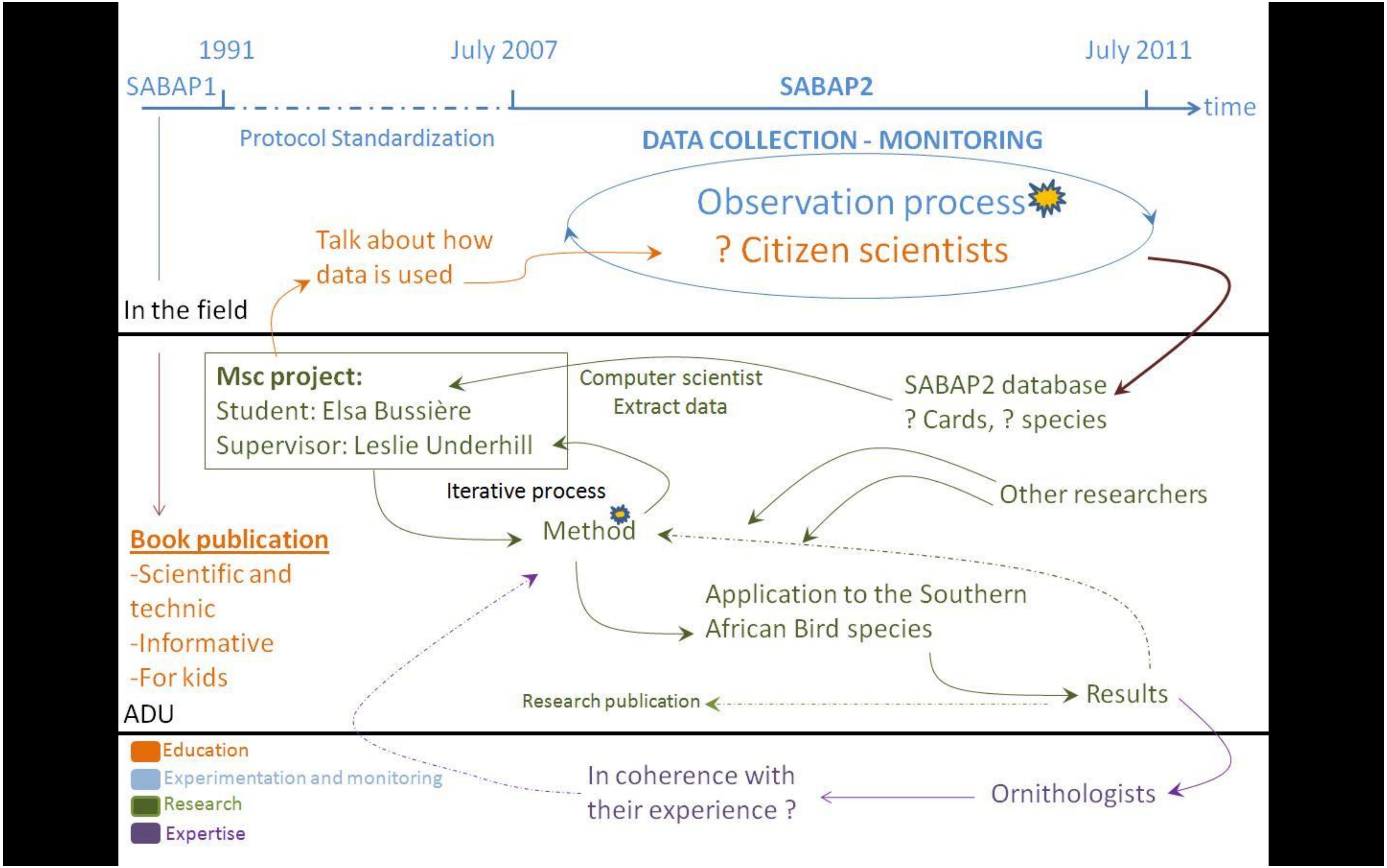
Appendixes

Appendix 1: Simplified calendar

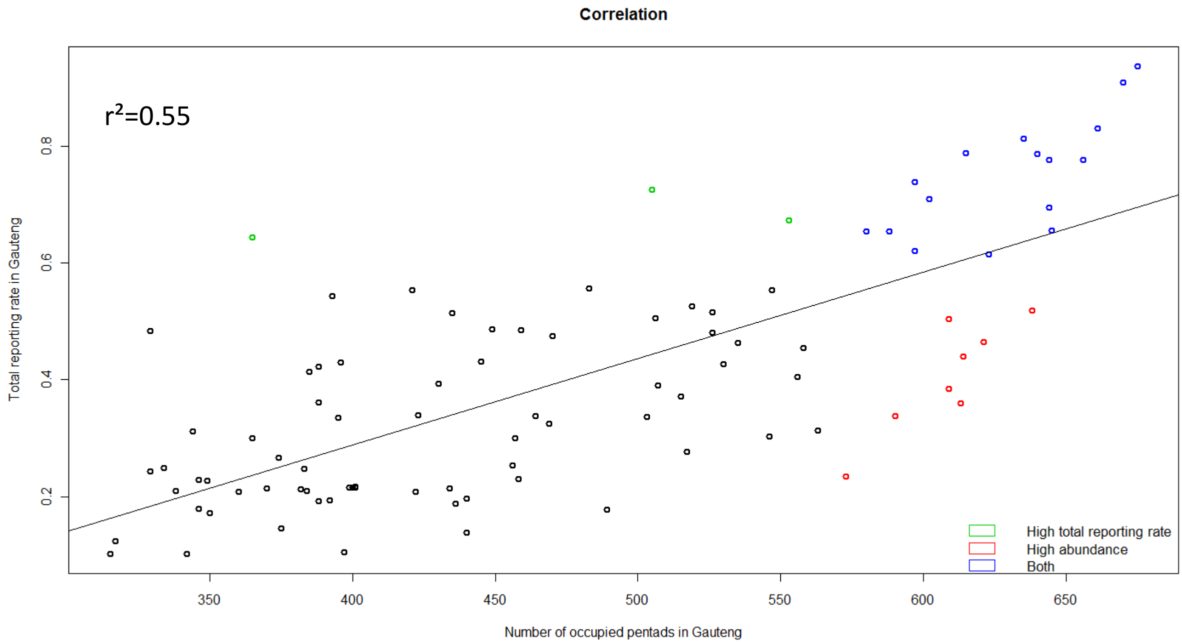
0 – 2nd	✣Fieldwork (atlasing) – Bibliographical work (SABAP1 & 2) – Statistics revisions
3rd – 6th	✣First Approach: Method of Distances
	➢Formalisation (in collaboration with Bergit Erni and Leslie Underhill)
	➢Writing a R-code + test on a fictitious dataset
	➢Downlaod real data via a computational tool (computer scientist)
	➢Data manipulation on Excel (Error)
	➢Adaptation of the R-code to the real dataset
	↳ showed method's first weaknesses (other researchers' help required)
	➢First results: showed more shortcomings
8th	✣R-code to display bird distributions with 3D-graphs
8th	✣Start writing the potential research article
9th	✣Start second project on potential climate change effect on bird migration
9th	✣Second Approach: Correlation coefficients
	➢New R-code
	➢Adaptation to exclude large areas where the species doesn't occur
	➢Creation of a new tool to separately downlaod the data for the migrants.
	➢Close look at the results (species after species)
10th	✣Rewrite all the R-code to avoid data manipulation in Excel (repeatability)
12th	✣Start thinking of how taking into account the observation process
12th	✣Give a talk à Kirstenbosch
13th	✣Discussion with a Post Doc about occupancy models to include observation process
13th	✣Start writing report
13th-14th	✣R-code to find a better way to display bird distributions (3D graphs → colored maps)
14th	✣Bibliographical work to list all the weaknesses
14th	✣Third Approach: Moran's I
	➢New R-code
	➢Results for Moran's I and correlation coefficients
15th	✣Find a clever way to illustrate different kind of effects
15th	✣ Write new R-code to get information about the observation process (N°obs/pentad)
16th	✣ Ask highly skilled ornithologists for information to confirm or invalidate the results
16th	✣ Ask the computer scientist for map distributions in South Africa
16th	✣Rewrite the two R-codes to take into account (deletion of blocks with no record, new data for migrants, the analysis in the South and the North)
17th	✣Analyse the questionnaires given to the ornithologists and try to find a way to create a typology
18th	✣ Start working on an improvement of the Moran's I method to take into account the arithmetic effect
19th	✣ write the reports, research paper
>20th	

weeks

Appendix 2: Interactions between actors



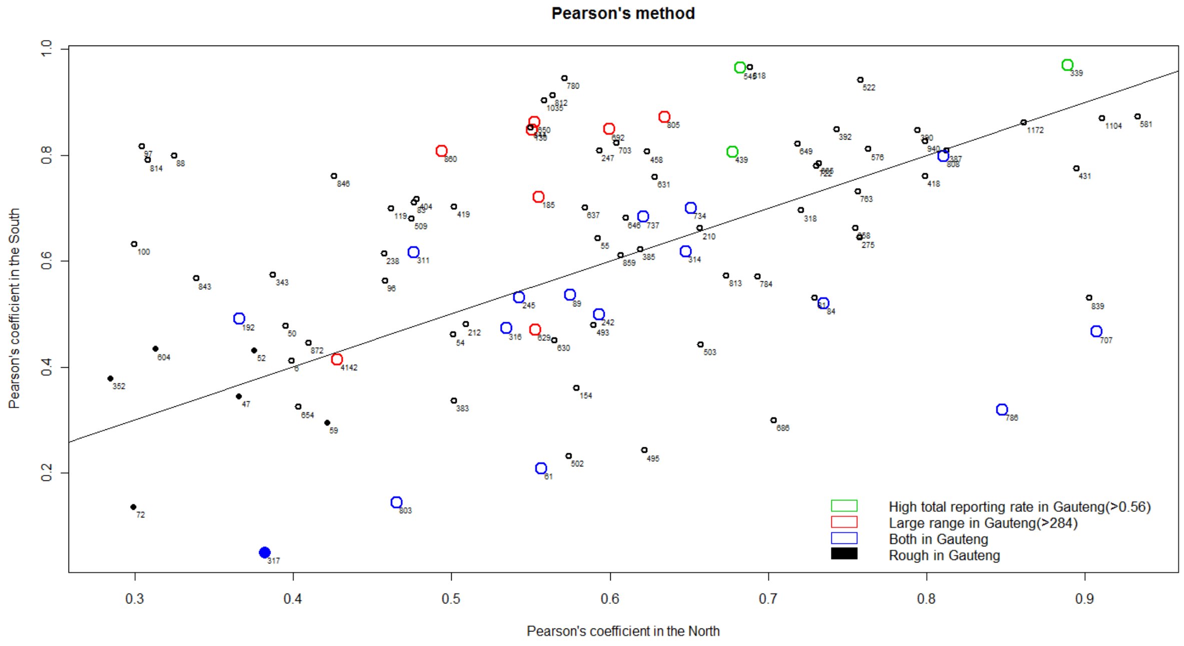
Appendix 3: Correlation between total reporting rate and total number of occupied pentads in Gauteng.



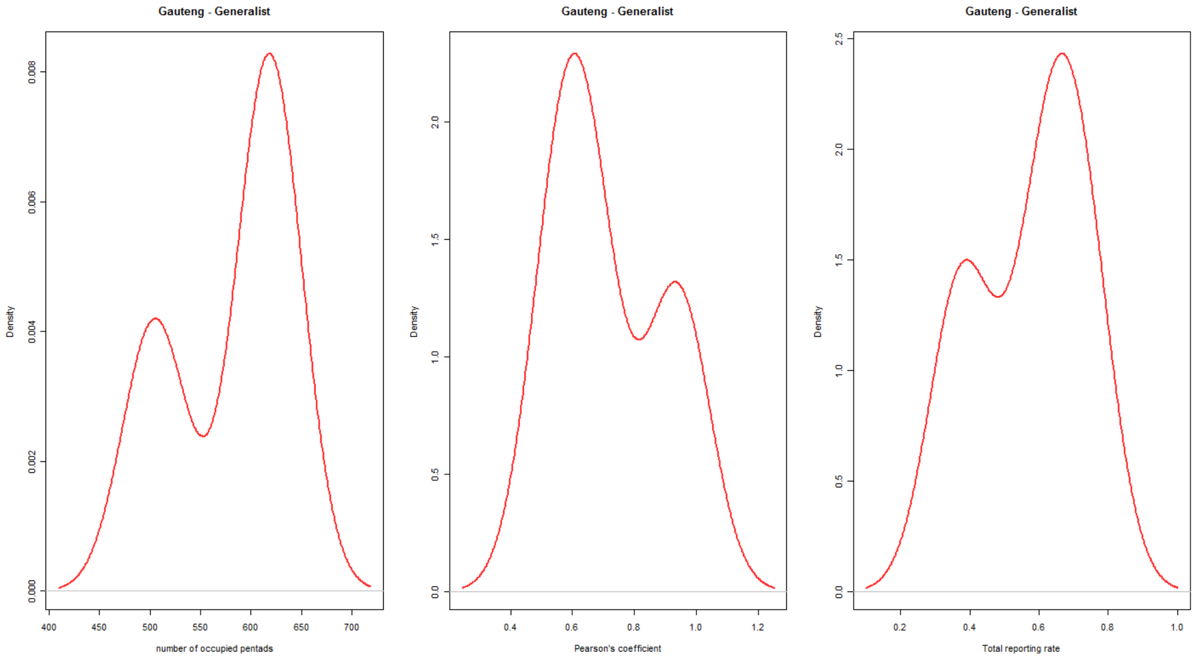
Appendix 4: Legend for the three scatters p16, 17 and 18

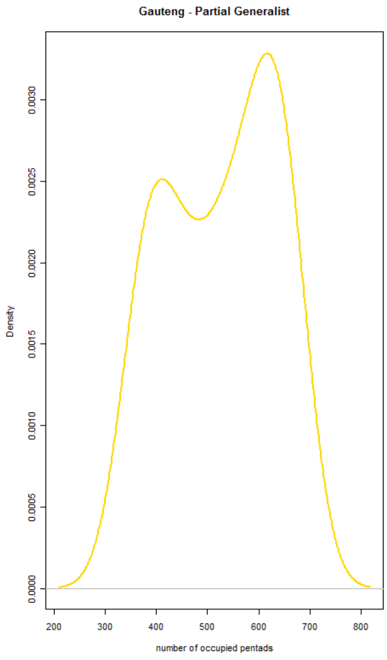
154: Steppe Buzzard, *Buteo vulpinus*
339: Grey Go-away-bird, *Corythaixoides concolor*
545: Dark-capped Bulbul, *Pycnonotus tricolor*
604: Lesser Swamp-Warbler, *Acroephalus gracilirostris*
703: Cape Longclaw, *Macronyx capensis*
707: Common Fiscal, *Lanius collaris*
818: Long-tailes Widowbird, *Euplectes progne*
839: Blue Waxbill, *Uraeginthus angolensis*

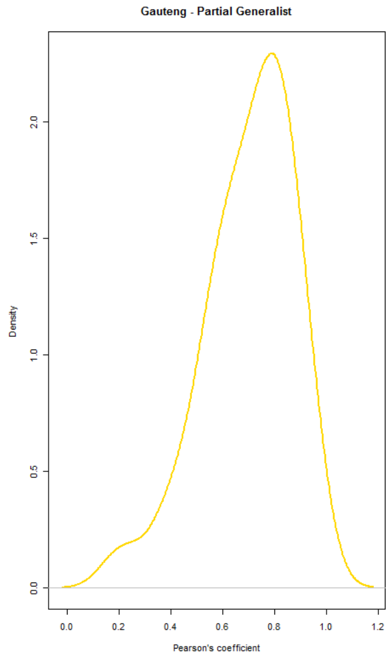
Appendix 5: Scatter North vs South

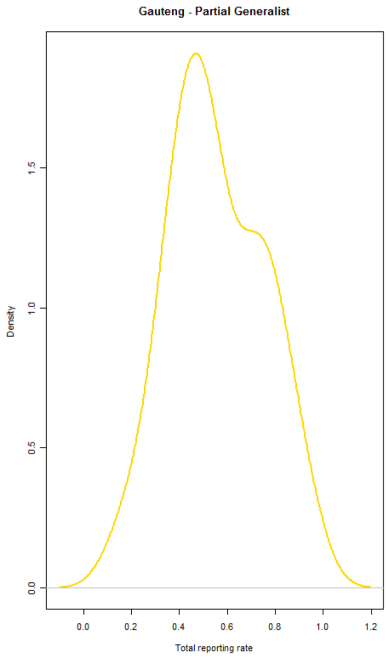


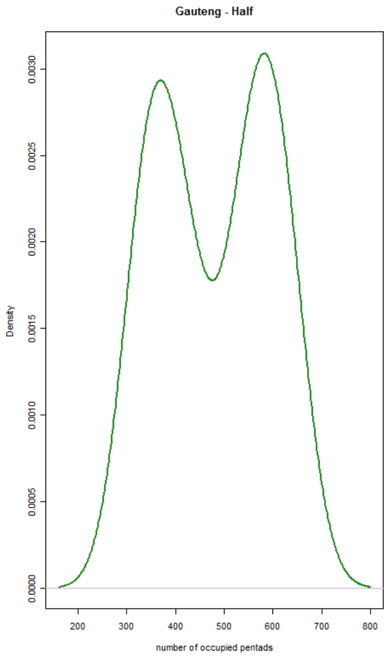
Appendix 6: Typology

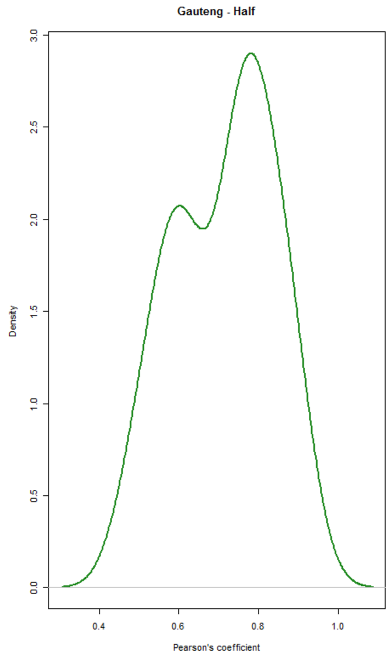


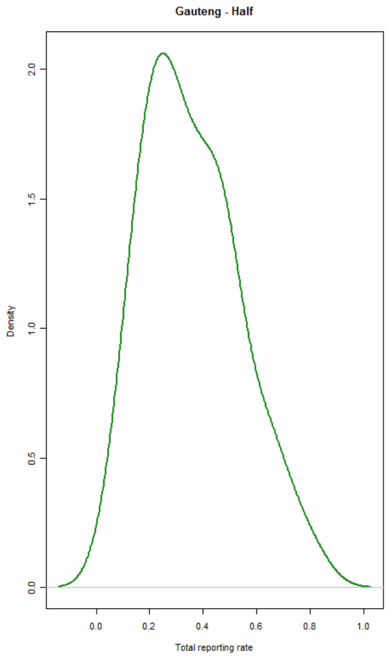

Gauteng - Partial Generalist
Density
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
200
300
400
500
600
700
800
number of occupied pentads

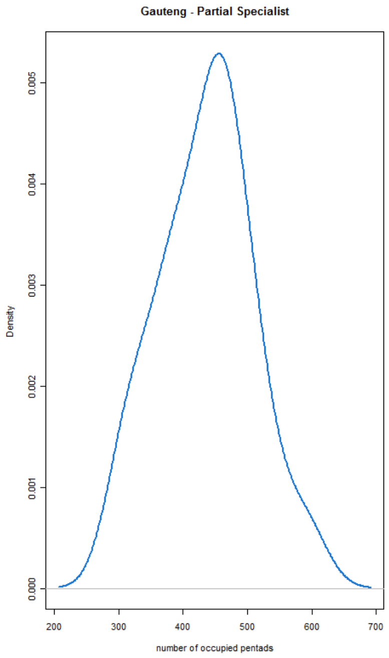

Gauteng - Partial Generalist
Density
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Pearson's coefficient

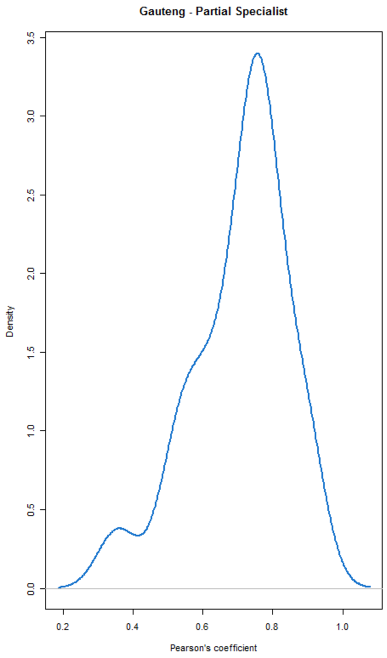

Gauteng - Partial Generalist
Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Total reporting rate

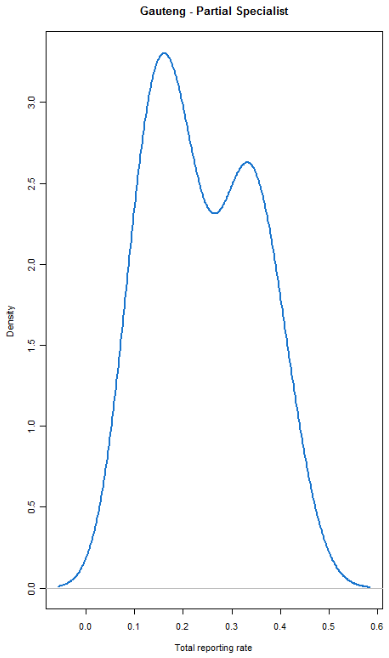

Gauteng - Half
Density
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
200
300
400
500
600
700
800
number of occupied pentads

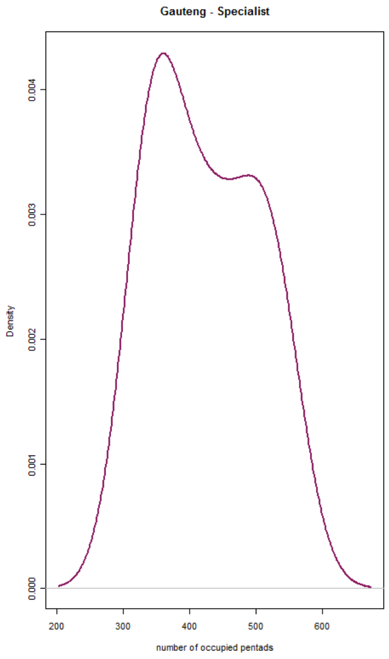

Gauteng - Half
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.4
0.6
0.8
1.0
Pearson's coefficient

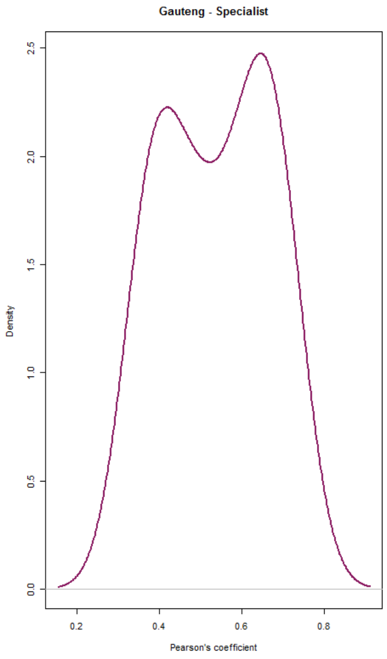

Gauteng - Half
Density
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
Total reporting rate

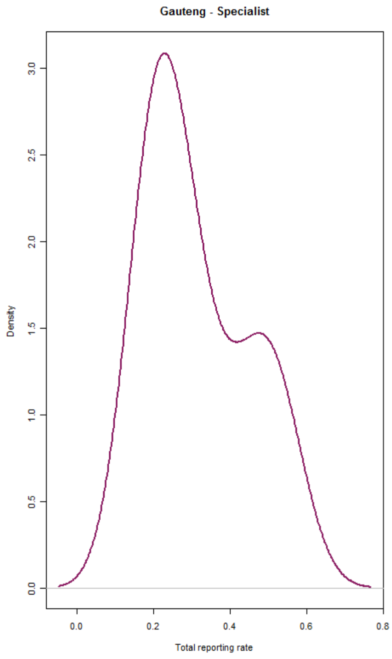

Gauteng - Partial Specialist
Density
0.000
0.001
0.002
0.003
0.004
0.005
200
300
400
500
600
700
number of occupied pentads


Gauteng - Partial Specialist
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.2
0.4
0.6
0.8
1.0
Pearson's coefficient


Gauteng - Partial Specialist
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Total reporting rate


Gauteng - Specialist
Density
0.000
0.001
0.002
0.003
0.004
200
300
400
500
600
number of occupied pentads


Gauteng - Specialist
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.2
0.4
0.6
0.8
Pearson's coefficient


Gauteng - Specialist
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
Total reporting rate

Appendix 7: Information about the species 192

observer	number of inventoried pentads	species	number of pentads with the species	[redacted]	Pent. Coord.	N. of cards	Pent. Coord.	N. of cards
10239	41	192	27	[redacted]	2455_2700	2	2530_2845	8
11827	67	192	45	[redacted]	2455_2745	1	2535_2750	7
11366	3	192	1	[redacted]	2455_2845	2	2540_2655	1
10706	18	192	13	[redacted]	2455_2850	1	2600_2655	1
10622	13	192	8	[redacted]	2455_2855	1	2605_2655	2
2220	7	192	6	[redacted]	2455_2900	1	2620_2705	4
10565	79	192	50	[redacted]	2500_2655	1	2700_2710	1
10718	45	192	30	[redacted]	2500_2745	4	2700_2845	1
51	125	192	86	[redacted]	2500_2810	4		
1692	106	192	65	[redacted]	2500_2900	2		
1691	107	192	84	[redacted]	2505_2655	2		
10101	59	192	46	[redacted]	2505_2900	2		
1706	42	192	26	[redacted]	2510_2655	5		
10824	40	192	25	[redacted]	2510_2900	1		
13090	1	192	0	[redacted]	2515_2810	6		
10005	104	192	77	[redacted]	2515_2855	4		
10085	122	192	83	[redacted]	2520_2845	4		
10281	79	192	52	[redacted]	2520_2850	5		
10210	87	192	59	[redacted]	2520_2855	4		
10377	30	192	26	[redacted]	2525_2710	4		
10848	136	192	113	[redacted]	2525_2750	4		
10821	52	192	42	[redacted]	2525_2850	5		
135	43	192	27	[redacted]	2530_2715	6		
10857	176	192	110	[redacted]	2530_2735	4		

Appendix 8: Information about species 803

Observer	Number of inventoried pentads	Species	Number of pentads with the species	[redacted]	Pentad coordinates	Number of cards
10210	87	803	75	[redacted]	2455_2655	1
1867	9	803	6	[redacted]	2455_2735	2
				[redacted]	2455_2745	1
				[redacted]	2640_2900	1
				[redacted]	2700_2800	1
				[redacted]	2700_2820	2

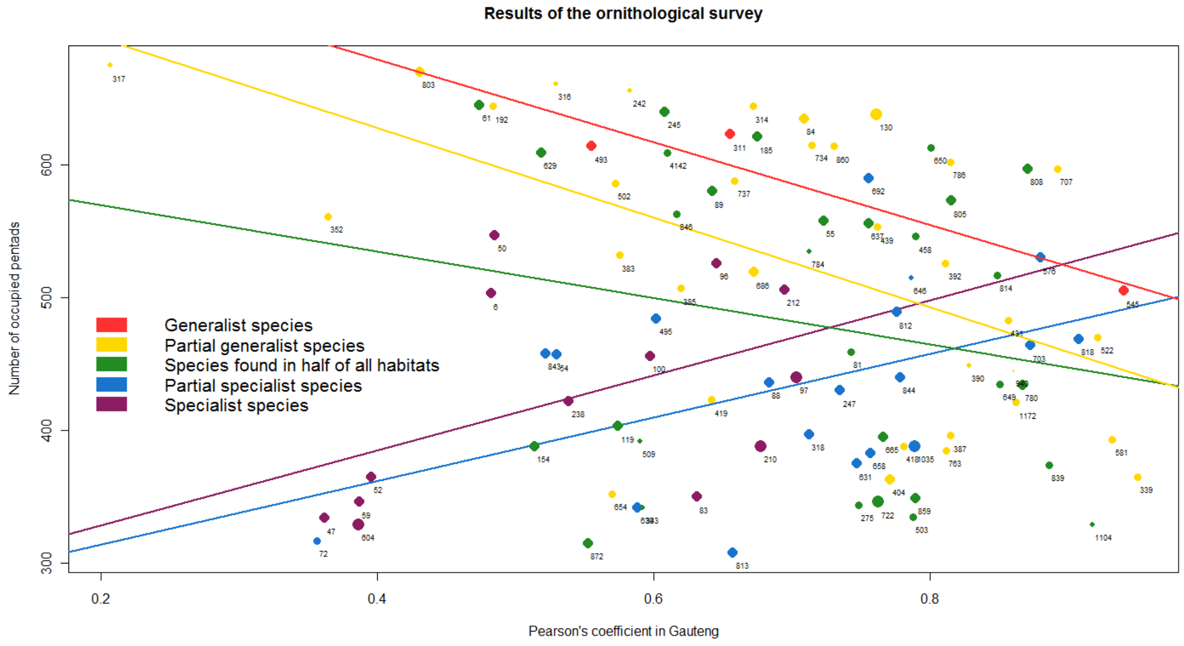
Appendix 9: Information about species 686

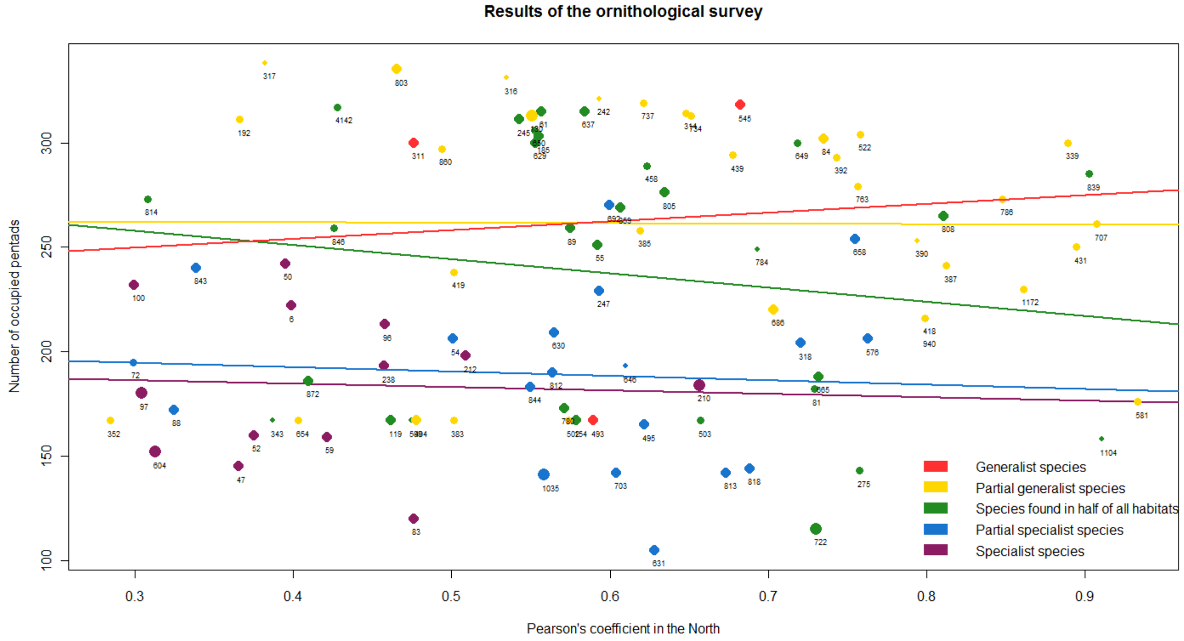
Observer	Number of inventoried pentads	Species	Number of pentads with the species	Pent. Coord.	N. of cards	Pent. Coord.	N. of cards	Pent. Coord.	N. of cards	Pent. Coord.	N. of cards	Pent. Coord.	N. of cards
10210	87	686	30	2455_2655	1	2500_2815	6	2510_2850	7	2530_2725	4	2620_2710	4
10239	41	686	19	2455_2700	2	2500_2825	8	2510_2855	4	2530_2900	3	2625_2655	1
11637	2	686	0	2455_2710	6	2500_2840	4	2515_2655	1	2535_2655	3	2625_2700	4
2220	7	686	1	2455_2720	3	2500_2845	4	2515_2715	4	2535_2700	4	2625_2715	4
51	125	686	46	2455_2725	3	2500_2850	5	2515_2720	14	2535_2725	7	2635_2655	4
11366	3	686	1	2455_2735	2	2500_2855	4	2515_2740	4	2535_2730	6	2635_2710	4
10768	70	686	35	2455_2740	3	2500_2900	2	2515_2750	4	2540_2655	1	2635_2735	5
10622	13	686	2	2455_2745	1	2505_2655	2	2515_2800	6	2540_2700	4	2640_2655	4
10706	18	686	3	2455_2800	1	2505_2700	4	2515_2805	7	2540_2705	4	2640_2730	5
10824	40	686	5	2455_2805	1	2505_2705	4	2515_2835	8	2540_2725	5	2640_2900	1
1868	40	686	15	2455_2810	6	2505_2710	4	2515_2840	4	2545_2655	2	2645_2655	4
11080	51	686	34	2455_2815	4	2505_2715	4	2515_2845	4	2545_2700	5	2650_2705	4
1692	106	686	46	2455_2830	6	2505_2725	4	2520_2655	4	2545_2705	4	2650_2900	3
10005	104	686	53	2455_2835	4	2505_2735	7	2520_2710	5	2545_2720	5	2655_2705	5
11827	67	686	12	2455_2840	2	2505_2810	71	2520_2720	4	2550_2705	6	2655_2720	5
1858	64	686	31	2455_2845	2	2505_2815	4	2520_2725	4	2555_2700	5	2655_2730	4
10857	176	686	79	2455_2850	1	2505_2820	5	2520_2745	7	2555_2705	6	2655_2830	4
				2455_2855	1	2505_2825	5	2520_2810	5	2555_2710	5	2655_2900	4
				2455_2900	1	2505_2830	6	2520_2825	7	2555_2900	2	2700_2710	1
				2500_2655	1	2505_2835	6	2520_2830	5	2600_2655	1	2700_2715	1
				2500_2700	4	2505_2840	5	2520_2835	10	2605_2655	2	2700_2725	1
				2500_2705	4	2505_2900	2	2525_2655	3	2605_2700	4	2700_2745	2
				2500_2710	4	2510_2655	5	2525_2700	35	2605_2900	2	2700_2750	1
				2500_2715	4	2510_2720	4	2525_2705	16	2610_2655	1	2700_2800	1
				2500_2720	5	2510_2725	4	2525_2725	5	2610_2700	4	2700_2850	1
				2500_2725	4	2510_2740	5	2525_2730	4	2610_2715	4		
				2500_2735	4	2510_2745	18	2525_2750	4	2610_2720	4		
				2500_2740	4	2510_2755	4	2525_2800	5	2610_2730	4		
				2500_2750	5	2510_2810	10	2525_2825	9	2610_2900	2		
				2500_2755	4	2510_2820	5	2525_2835	7	2615_2655	1		
				2500_2800	7	2510_2825	13	2525_2845	6	2615_2715	4		
				2500_2805	6	2510_2840	4	2530_2655	9	2615_2740	5		
				2500_2810	4	2510_2845	6	2530_2700	4	2620_2705	4		

Appendix 10: Information about species 61

Observer	Number of inventoried pentads	Species	Number of pentads with the species	[redacted]	Pent. Coord.	N. of cards	Pent. Coord.	N. of cards
10210	87	61	45	[redacted]	2455_2655	1	2545_2705	4
10239	41	61	24	[redacted]	2455_2700	2	2610_2655	1
11637	2	61	0	[redacted]	2455_2720	3	2610_2700	4
1692	106	61	73	[redacted]	2455_2725	3	2620_2655	1
10610	15	61	7	[redacted]	2455_2730	2	2640_2900	1
2220	7	61	3	[redacted]	2455_2735	2	2700_2655	5
51	125	61	96	[redacted]	2455_2740	3	2700_2700	1
11366	3	61	2	[redacted]	2455_2745	1	2700_2715	1
10824	40	61	27	[redacted]	2455_2750	2	2700_2720	1
11010	41	61	35	[redacted]	2455_2850	1		
10775	31	61	15	[redacted]	2500_2705	4		
19999	8	61	3	[redacted]	2500_2725	4		
10821	52	61	37	[redacted]	2500_2735	4		
1870	23	61	18	[redacted]	2500_2740	4		
10768	70	61	61	[redacted]	2505_2655	2		
10857	176	61	119	[redacted]	2510_2655	5		
10848	136	61	102	[redacted]	2515_2740	4		
10777	5	61	2	[redacted]	2535_2700	4		
11827	67	61	40	[redacted]	2540_2655	1		
10101	59	61	42	[redacted]	2540_2700	4		
272	21	61	16	[redacted]	2545_2655	2		
13090	1	61	0	[redacted]	2545_2700	5		

Appendix 11: Importance of the demarcation of the studied area


Results of the ornithological survey


Results of the ornithological survey