Statistical analysis

*Ingo Fetzer,^{1,2} Luigi Piemontese,¹
Juan C. Rocha^{1,4} and Berta Martín-López³*

¹STOCKHOLM RESILIENCE CENTRE, STOCKHOLM UNIVERSITY, STOCKHOLM, SWEDEN

²BOLIN CENTRE FOR CLIMATE RESEARCH, STOCKHOLM, SWEDEN

³LEUPHANA UNIVERSITY, LÜNEBURG, GERMANY

⁴BEIJER INSTITUTE OF ECOLOGICAL ECONOMICS, THE ROYAL SWEDISH ACADEMY OF SCIENCES, STOCKHOLM, SWEDEN

Key methods discussed in this chapter

Descriptive statistics, group comparison, regression models (linear, generalised linear), multi-variate analysis (including clustering, non-metric multidimensional scaling (n-MDS), principal component analysis (PCA), redundancy analysis (RDA), canonical correspondence analysis (CCA), factor analysis (FA) and multiple correspondence analysis), time series analysis.

Connections to other chapters

Statistical methods are used to analyse data from a wide variety of sources, such as ecological field data (Chapter 6) for understanding drivers and dynamics, data from participatory data collection (Chapter 8) or interview data from surveys (Chapter 7). Statistical methods can give an overview of spatial datasets (Chapter 24) but also help to discover patterns in time series results of agent-based modelling (Chapter 28). Statistical approaches are, among others, also used to estimate how well dynamic systems models (Chapter 26), expert models (Chapter 16) and ecosystem service models (Chapter 31) represent empirical data. Finally, statistics provide methods for historical assessments (Chapter 25) and quantitative pattern recognition (Chapter 17).

Introduction

Statistical methods are mathematical tools that can help to aggregate, present and explore complex datasets from various sources. These include field data, which are often used in social-ecological systems (SES) research. These methods can be very useful for understanding interactions, dependencies and relationships between social, environmental and ecological variables. Most statistical approaches and tests allow researchers to identify and determine the main significant controlling (environmental or social) parameters (Dalgaard 2008; Crawley 2015).

Although the use of statistical methods dates back to at least the 5th century BCE, the origin of statistics had little in common with what it is used for today, as the original use of statistics was limited to governance, mainly by providing demographic data. It was only during

SUMMARY TABLE: STATISTICAL ANALYSIS	
DISCIPLINARY BACKGROUND	KNOWLEDGE TYPE
<p><i>The methods in this chapter are derived from or have most commonly been used in:</i></p> <p>Ecology, Social Science, Political Science, Economics, Demography, Psychology, Earth Sciences</p>	<p><i>The methods in this chapter are primarily used to generate the following types of knowledge:</i></p> <ul style="list-style-type: none"> • Descriptive • Exploratory • Explanatory
RESEARCH APPROACH	PURPOSE OF METHOD
<p><i>The methods in this chapter originate from or most commonly adopt the following research approaches:</i></p> <ul style="list-style-type: none"> • Analytical/objective 	<p><i>The most common purposes of using the methods in this chapter are:</i></p> <ul style="list-style-type: none"> • System understanding
TEMPORAL DIMENSION	SYSTEMIC FEATURES AND PROCESSES
<p><i>The methods in this chapter are most commonly applied to the following temporal dimensions:</i></p> <ul style="list-style-type: none"> • Present (typically within the last 5–10 years) • Recent past (post-1700s) • Pre-industrial revolution (pre-1700s) • Future 	<p><i>While most methods can do many things, the methods in this chapter are particularly good (i.e. go-to methods) for addressing the following:</i></p> <ul style="list-style-type: none"> • SES components and linkages • Diversity • Social-ecological dependence and impact • Regime shifts • Exploring uncertainty
SPATIAL DIMENSION	
<p><i>The methods in this chapter are primarily either or both:</i></p> <ul style="list-style-type: none"> • Non-spatial • Explicitly spatial <p><i>The methods in this chapter are most commonly applied at the following spatial scales:</i></p> <ul style="list-style-type: none"> • Local • Regional (provincial/state to continental) • Global • Multiple places/sites around the world 	

the 18th century that the approach was extended to various scientific fields, when the newly developed probability theory (or ‘law of great numbers’) was integrated and recognised as a new branch of mathematics. Jakob Bernoulli, Abraham de Moivre and others introduced the idea of representing complete certainty by the number one, and probability (as a measurement for uncertainty) as a number between zero and one (Varberg 1963; Fienberg 1992).

Modern statistical methods consist of a vast toolbox of different tests and analytical procedures. The choice of the most appropriate method requires: (a) a good understanding of the type of data (categorical or continuous data), (b) knowledge of the distribution pattern and shape of the data (e.g. normally distributed or skewed) in order to apply the appropriate statistical tests, and (c) an understanding of statistical methods to correctly interpret the outcomes. This is especially crucial because most statistical methods, as already mentioned above, do not provide an explicit outcome of ‘yes’ or ‘no’, but rather give the probability of an event happening or the likelihood of an item belonging to a certain group. So, researchers have to be clear about what is meant by assigning statements such as ‘unlikely’ or ‘likely’ to the outcome of an analysis, expressed by a ‘probability value’ (*p*-value) (Hon 2010; Crawley 2015).

Statistical tools generally test a hypothesis, usually derived from a specific scientific question, and investigate the likelihood for the hypothesis to be either accepted or refuted. Statisticians have agreed that an event is unlikely if it occurs less than 5% of the time ($p < 0.05$). Sometimes, for stricter tests (e.g. in medical studies), thresholds of 1% ($p < 0.01$) or lower are used. Thus, a probability value of less than 5% suggests that this event is unlikely to occur by chance. However, there is still a small probability that two unrelated events could take place simply by chance (leading to a ‘false positive’ or type I error), or that a relationship between two related events is not detected by the test (leading to a ‘false negative’ or type II error). Thus, poor insight into basic statistical principles combined with a lack of knowledge about data quality, biased data or too few data may lead to wrong conclusions (Dalgaard 2008; Hon 2010; Crawley 2015).

Generally, statistical methods do not directly test the likelihood that the question being asked is true, but rather the likelihood of a ‘null hypothesis’ being true. A null hypothesis is a general statement or default position that there is ‘no relationship’ between two measured phenomena, or ‘no association’ among groups. Thus, dynamics between them are uncoupled and random. The null hypothesis often starts in the opposite position of the question being asked! In other words, the null hypothesis is the logical opposite of the scientific question that one is interested in. The logic is that if it turns out that the null hypothesis is very unlikely to be true (e.g. with a probability of $p < 0.05$ (Dalgaard 2008; Crawley 2015)), then its logical opposite (the alternative hypothesis reflecting the research question) must be valid. The real art is to pose a question that can lead to a research hypothesis of which its logical opponent, the null hypothesis, can be formulated. Finally, the null hypothesis has to be tested with the appropriate test (Quinn and Keough 2002; Hon 2010).

Example

Research question: Is there a difference between measurements of Group A and B?

Research hypothesis: Measurements of Group A and B differ

Null hypothesis: Measurements of Group A and B do not differ

Only if a null hypothesis needs to be refuted by a statistical test then the research hypothesis, as the only logical alternative, is accepted to be true.

Social-ecological systems research uses many popular statistical methods to analyse quantitative and qualitative data, including statistical visualisation, analyses and modelling.

- **Statistical visualisation:** Visualisation of statistical results, mainly through plots and diagrams, is used to aggregate data and understand patterns (e.g. social-ecological interactions) and trends (forecasting and development), and to help build conceptual models (e.g. causal loop diagrams).
- **Statistical analyses:** Statistical analyses include traditional aggregation of datasets (e.g. by estimating mean, variance, standard error) that help to summarise data, understand patterns, estimate dynamics, dependencies and correlations between or within elements of data, and calculate differences between groups. Different statistical techniques (e.g. confidence intervals, errors) are often used to review the quality and certainty of data collections and inquiries, and to understand relationships between different components of complex SES and their contexts (e.g. estimating covariance, correlation strengths). These analyses can be used for data-based scientific and political decision support, for instance by creating statistical models or calculating metrics as means and modes from data points by comparing deviation from similar studies done elsewhere or earlier (e.g. the number of disease outbreaks in a community before and after the introduction of a waste-water treatment plant).
- **Statistical modelling:** Statistical modelling includes modelling and forecasting future trends or filling in gaps for missing data. Statistical models are often used to improve strategic and political planning, e.g. fitting empirical data to a function to allow one to make general statements, even for situations where no data exist. Clustering multifactorial data into an easy-to-read visual plot to understand which factors are relevant is another common use of statistical model outputs.

SES problems and questions

A number of common SES features and questions can be explored using statistical analyses of empirical SES data, such as ecological field data (Chapter 6), interview and survey data (Chapter 7), or data obtained from participatory data collection (Chapter 8) or via data-mining techniques (Chapter 17). Statistical methods allow an understanding of a huge range of phenomena, such as social behaviour, ecological or environmental changes, and social-ecological interactions. Typical questions include:

- What are the typical attributes of key SES features in a particular region? (e.g. what is the average amount of fuelwood harvested or used by communities in a particular region, and how much does it vary between households? (Paumgarten and Shackleton 2011; Keane et al. 2019)). Statistical analysis can be used to determine the mean or central value of a dataset containing multiple data points, as well as the range and variability in the dataset, and allows comparing the values.
- How do SES features in a particular area compare with those of different areas or with other studies from literature? (e.g. how does the average income of rural and urban communities differ (Keane et al. 2019; Robinson, Zheng, and Peng 2019); how does mean summer precipitation values differ between regions?). Statistical methods can be used to determine whether there are significant differences between different regions or groups.
- How are different SES features related? (e.g. regional farm assessments investigate how environmental factors such as irrigation water availability, soil type, mean summer temperatures and growing periods are related to aspects such as farm size, income, distance to markets and the number of family members (Keane et al. 2019; Robinson, Zheng, and Peng 2019)). Regression analyses can be applied to determine relationships, and the

strength of relationships, between system components (e.g. how does the application of a pesticide relate to its effect on crops while increasing income but also affecting human health or an ecosystem (Carvalho 2017), or how do different levels of land cover change the impact on biodiversity, ecosystem services provision and human well-being? (Iniesta-Arandia et al. 2014; Meacham et al. 2016; Martín-López et al. 2017)). Once these relationships are established, regression models can be used to forecast effects on the impact of future changes, or to help fill gaps in empirical data (Pollnac et al. 2010; Pinsky and Fogarty 2012).

- Which places have similar SES characteristics? (e.g. which farms are more similar in their socio-economic and environmental characteristics? (Martín-López et al. 2012; Meacham et al. 2016; see also Case study 18.1)). Multivariate analysis can be used to bring all the SES characteristics (with all their different units) together, allowing researchers to identify clusters of similar SES sites. Similar sites often either share similar problems (Meacham et al. 2016) or can be used as archetypes to identify other places sharing similar SES conditions where a particular intervention could also be successfully applied (Piemontese et al. 2020).
- How do key SES features change across years? (e.g. time series analyses allow researchers to estimate how early identification of changing rainfall conditions can be a critical indicator to ensure national food security; these analyses can also allow conducting risk assessments to determine the likelihood and potential impacts of droughts and flooding for small-scale farmers (Shongwe et al. 2011; Husak et al. 2013)).

Brief description of key methods

Prior to any analysis, a researcher must be aware of the characteristics of the investigated data (Hon 2010), as the selection of appropriate statistical methods depends on the data types. This includes both the type and the distribution of data. A first step in any analysis is usually to summarise and visualise the data to gain an understanding of the dataset and detect obvious data entry errors. Following this, the data may be analysed using a variety of statistical methods.

Data types

Two main characteristic types of data exist: discrete and continuous. It is essential to make this distinction because each data type is analysed using different statistical methods and tests (so-called ‘parametric’ and ‘non-parametric’ statistical approaches – see details below).

Discrete (categorical) data are descriptive data that can be counted according to group attributes (or their ‘quality’). Group attributes can be categorical factors which are either non-rankable (e.g. sex, colour) or rankable (e.g. age classes, size groupings). The number of male and female participants in surveys can be counted, for example. The total number of male and female participants cannot be ranked, whereas the numbers of females and males between several survey batches can be ranked. Discrete data that cannot be ranked are referred to as ‘nominal data’, whereas data that can be ranked or ordered are referred to as ‘ordinal data’ (see ‘Statistical notation’ in Table 18.1).

Continuous data are data that cannot be counted, e.g. temperature can be measured using a thermometer but cannot be counted. Unlike ordinal data, continuous data include the size or a magnitude (the ‘quantity’ of a parameter that can be put on a standardised equidistant (= interval)

Table 18.1 Overview of data types, characteristics and terms used for the statistical categorisation of data

<i>Data type</i>	<i>Characteristics (and scale)</i>	<i>Statistical notation</i>	<i>Category</i>	<i>Distribution</i>	<i>Example</i>
Count data (<i>n</i>)	Discrete (= countable)	Nominal data	Qualitative	Non-parametric	15 students
Count with 2 groups (dichotomous data)	Discrete categorical Non-rankable	Nominal data	Qualitative	Non-parametric	yes–no, male–female
Count data with > 2 groups	Discrete categorical Non-rankable	Nominal data	Qualitative	Non-parametric	7 black, 4 white, 6 red
Count data with rankable groups	Discrete categorical Rankable	Ordinal data	Qualitative	Non-parametric	2 small, 4 medium, 5 large
Measurements (but no true zero)	Continuous	Interval data	Quantitative	Parametric	2.0 °C, 2.1 °C, 3.2 °C, etc.
Relative measurements (with true zero)	Relative continuous	Interval ratio data	Quantitative	Parametric	23 kg 34 kg 45 kg, etc.
Time series	Continuous data linked by a time vector	Interval data along a time axis	Quantitative	Parametric	Day 1: 2.3; Day 2: 3.1; Day 3: 4.3; etc.

scale). Continuous data therefore always take numerical values: the distance between two points, when put on a scale, is placed at an equal representative distance from each other. Continuous data are therefore statistically categorised confusingly as ‘interval data’. With no zero reference point, values below zero are possible (e.g. for temperature, altitude).

For continuous data with a zero reference point present, all data are placed relative (or as a ‘ratio’) to this zero reference point. Values below zero are thus not possible (e.g. percentage, height or weight). These data are termed ‘ratio data’ (see ‘Statistical notation’ in Table 18.1). When working with ratio data, but not interval data, the ratio of two measurements has fundamentally different consequences: for the ratio variable ‘weight’, a weight of 8 grams is always exactly twice as heavy as a weight of 4 grams. However, for the interval variable ‘temperature’, a temperature of 10 °C need not necessarily be considered to be twice as hot as 5 °C.

Continuous data always contain a higher ‘information quantity’ than rankable, discrete (ordinal) data, which in turn contain a higher information quantity than discrete (nominal) data (Table 18.1). Data with a higher information quantity can be easily converted into data with a lower information quantity (e.g. continuous temperature measurements can be placed

into groups of ‘cold’ and ‘warm’ and events can be counted), but the reverse is rarely possible. However, data of high information quantity usually require more effort and resources to gather and, depending on the system under investigation, may often be even impossible to attain. Moreover, comparison of data is often only possible between datasets of similar types. In their research planning, researchers have to balance the data with high ‘information quantity’ with the amount of ‘data quantity’ actually needed to answer the research question (Dalgaard 2008; Hon 2010; Crawley 2015).

Data distribution

In addition to knowing the type of data, it is critical to understand the distribution of the data in order to select an appropriate statistical method. The data distribution can be described by a continuous distribution function, which is called the ‘probability density function’. Continuous data in statistical data analysis are often assumed to be symmetrically or ‘normally’ distributed around a central value (the ‘mean’) with some variability. However, the assumption that data are normally distributed needs to be tested (e.g. by running a ‘normality’ test) before any further analysis, as the results may be invalid if this assumption does not hold. As discrete data are usually not normally distributed, most standard statistical tests cannot be applied. ‘Parametric’ data refer to data that follow a normal or other specified distribution, allowing the more statistically reliable parametric tests to be applied, while data that do not follow a specified distribution are termed ‘non-parametric’ (or distribution free). Here only the less robust non-parametric class of tests is applicable.

Table 18.1 provides an overview of data types and technical statistical terms used to categorise data.

Summarising and visualising data

The first step in statistical analysis is typically to summarise and visualise the data. Parameters such as mean, variance and standard deviation are typical overall descriptors for quantitative, normally distributed parametric data. One may, for example, summarise mean (average) crop production across different study sites or conditions, and the variability in crop production using other characteristic metrics such a range, variance or standard deviation. For non-parametric data and discrete data, descriptive parameters such as median, quartiles or mode (relative maximum) are typical descriptive indices (Crewson 2006; Brown and Saunders 2007).

It is also useful to explore the data by visualising them in order to get a ‘feel’ for the data and spot any obvious data-entry errors. Discrete data are usually presented as numbers in a contingency table, or graphically as bar or pie plots. If the discrete (countable) data are several repeated measurements (e.g. number of participants in several survey batches), then overall descriptions around a median or mode can be done. Variability of this type of data is described using quantiles or percentiles and is visualised in ‘box-and-whisker’ plots (Dalgaard 2008; Hon 2010; Crawley 2015).

Continuous ungrouped data are typically visually represented as ranked or non-ranked scatter plots, single-line plots or simple bar plots. With grouped continuous data, stacked bar plots, sequential means with error bars or overlapping line plots are often applied. For the representation of size distributions, bar graphs with binned data or frequency histograms are useful. These are often also used for visual detection of the data’s distribution for normality, skewness and biases (often with a standard normal plot overlay). Alternatively, normal

probability plots can be applied directly to visualise how far all or single extremes of the measurement values deviate from a normal distribution (Dalggaard 2008; Hon 2010; Crawley 2015).

Analysing data

Many SES datasets consist of multiple variables. Each of these variables can be analysed separately using descriptive statistics to understand their central tendency and variability (Table 18.2). Where the same variables (e.g. height, species abundance) have been measured across different sites or groups, they can be compared using various methods for group comparison techniques (e.g. to test whether bird abundance differs between different sites). If the prerequisites for parametric data are not fulfilled, then less sensitive non-parametric alternatives for the parametric test have to be used instead.

Alternatively, relationships between different variables can be explored. The simplest relationship is an explanatory–response relationship, where a change in an explanatory variable ‘determines’ a change in a response variable. Regression models can test the relationship between a response and single or multiple explanatory variables. Most regression models assume that the response and explanatory variables are normally distributed, but there are also options for variables with other distributions. In SES research, regression models are commonly used to: (a) determine the strength of predictors, (b) forecast an effect, or (c) find trends. Linear regression models are also used to fill in missing values in time series, and other types of datasets estimate gaps in measurements using values anticipated by the model. With a well-fitted model, back- and forecasting estimates, across the minimum and maximum limits where own measurements have been made, are possible.

In contrast to ‘univariate’ analyses that focus on predicting a single outcome, ‘multivariate’ analyses investigate how multiple outcome variables co-occur. The aim of these approaches usually centres on comparing how similar two or more objects (e.g. sites, households, ecological state) are, allowing the combination of relevant characteristics with different units and data classes. Multivariate techniques include, among others, hierarchical clustering, principal component analysis (PCA), factor analysis (FA), multiple correspondence analysis (MCA) or non-metric multidimensional scaling (n-MDS) (Hon 2010).

In an SES context, variables in multivariate analyses can contain combinations of social and/or environmental aspects simultaneously, but they should all be on quasi-identical scale; otherwise, the variable with the largest range will dominate the outcomes. To avoid this, normalisation of the data (by converting them into relative values of a range from 0 to 1) is usually necessary. Multivariate analyses typically entail a mathematical procedure of ‘dimensionality reduction’. This allows the dimensionality of the data to be reduced to a two-dimensional table matrix or graph that can be plotted and is easier to understand. Thus, multivariate analysis allows interpretation of complex interactions of multiple variables where no direct explanatory–response relationship is possible.

Many events in SES research are ordered in time and can be analysed using time series analysis (TSA) models. These models can have different purposes. Some aim to interpret plausible descriptions of sampled data; others to test hypotheses about the drivers behind the observed variability or to quantify response similarity (e.g. synchrony of two or more events). Time series analysis models are sometimes also intended to forecast and simulate scenarios, e.g. conducting risk assessment or projecting trajectories into the future for different management policies. Table 18.2 provides a summary of the key methods used in statistical analysis.

Table 18.2 Summary of key methods used in statistical analysis

<i>Method</i>	<i>Description</i>	<i>References</i>
Descriptive statistics	<p>Descriptive statistics are indices that summarise a given dataset so that it can be easily understood. The coefficients can be divided into measures of:</p> <ul style="list-style-type: none"> • Central tendency, including mean and median mode • Variability (spread), including range, quartiles, variance and standard deviation • Correlation 	<p><i>Key introductory texts</i> Dalgaard 2008; Hon 2010</p> <p><i>Applications to SES</i> Scheffer et al. 2012; Dearing et al. 2014; Jaramillo and Destouni 2014; Oteros-Rozas et al. 2014</p>
Group comparison	<p>Several statistical tests are available to compare whether the means of two or more groups differ. To compare two groups, the most common tests are the T-test (normal data), the Mann–Whitney <i>U</i> test (non-parametric ordinal data) and Chi-Square (non-parametric nominal data). To compare more than two groups, the ANOVA (normal data), Kruskal–Wallis (non-parametric ordinal) or Chi-Square (non-parametric nominal) test can be used. To determine which of more than three groups differ, a ‘post hoc’ test, e.g. a Tukey or Scheffé test, needs to be applied.</p>	<p><i>Key introductory texts</i> Dalgaard 2008; Crawley 2015</p> <p><i>Applications to SES</i> Cummings et al. 2010; Inieta-Arandia et al. 2014; Martín-López et al. 2017</p>
Regression models	<p>Regression models (simple linear, multiple linear, non-linear, generalised linear) are statistical methods allowing one to summarise and study the relationships between two or more variables. Simple regression models are used to study explanatory–response relationships between two continuous variables. Linear models are the most common regression models, used for normally distributed continuous data. Multiple regression models estimate the response of a single variable from a combination of several predictor variables.</p> <p>Using least-square fitting to non-linear distribution models allows one to understand the characteristics of non-linear data. Data transformation (e.g. log, square root, double square root) of non-linear data may make it possible to linearise data and allow the application of linear models.</p> <p>Generalised linear models can be used for non-normally distributed, binary (logistic regression) and count (Poisson regression) data. Model optimisation (e.g. finding the correct set of variables that best explains the behaviour of the variable of interest) is often supported by using an Aikaike (AIC) or a Bayesian (BIC) information criterion and avoids critical over-fitting.</p>	<p><i>Key introductory texts</i> Dalgaard 2008; Hon 2010; Crawley 2015</p> <p><i>Applications to SES</i> Pollnac et al. 2010; Martín-López et al. 2012; Pinsky and Fogarty 2012; Meacham et al. 2016; Kim and Kim 2018</p>

<i>Method</i>	<i>Description</i>	<i>References</i>
Multivariate analysis	Multivariate clustering methods originate from ecology. These techniques include clustering, non-metric multidimensional scaling (n-MDS), principal component analysis (PCA), redundancy analysis (RDA), canonical correspondence analysis (CCA), factor analysis (FA) and multiple correspondence analysis (MCA). They are used for the classification (grouping) of multivariate data objects (i.e. datasets with multiple response or outcome variables) into sets with similar characteristics ('clusters') according to the similarity of their overall variable values. These techniques are often used with large datasets with many variables and help to transform high-dimensional data into more easily understandable two-dimensional graphs or tables. Similarity of objects is determined based on how similar objects are with regard to their variable composition and values. While clustering and n-MDS are based on a similarity index (typically Euclidean distance), PCA, RDA, CCA, FA and MCA define the dimension reduction based on the correlation or covariance between objects.	<i>Key introductory texts</i> Wilmink and Uytterschaut 1984; Quinn and Keough 2002; Oksanen 2015 <i>Applications to SES</i> Cinner et al. 2013; González-Orozco et al. 2014; Iniesta-Arandia et al. 2014; Kotzee and Reyers 2016; Maione, Nelson, and Barbosa 2018
Time series analysis	Time series analysis (TSA) is applied to time series data. The most common statistical models of time series are multivariate autoregressive state-space models (MARSS), also known as vector autoregression in econometrics. These models are typically composed by a drift term that represents a deterministic dynamic (e.g. a growth rate) and a process error term that is stochastic (e.g. some random variable that represents environmental variation). The latter is important because even in a model with a positive growth rate, a population can collapse due to the environmental noise. TSA is often a necessary condition to identify causality, although not all time series data will allow for the detection of causal effects.	<i>Key introductory texts</i> Angrist and Pischke 2009; Holmes, Ward, and Wills 2012; Pearl and Mackenzie 2018 <i>Applications to SES</i> Sabo et al. 2017

Limitations

Social-ecological systems are intrinsically difficult to understand due to their complex relationships, interactions between internal and external parts of the system, and the effect of random fluctuations and chance events. A key objective in SES research is to understand how different elements of an SES are related and how relationships change over time. However, it is often difficult to distinguish between the main influencing factors in a system and so-called 'noise'. Noise can derive from either insignificant influencing elements or random processes (Crewson 2006; Dalgaard 2008). Random processes are patterns that are measured but lack

structure or predictability. Random noise can often be a large component in data, and it can be difficult to determine its magnitude. It decreases the quality of data by obscuring the main signal, i.e. the effect of the main influencing factors.

Data quality can often be already increased during the data-gathering process, e.g. through excluding avoidable random processes (e.g. taking unequal numbers of samples at different points during a field study) and biases (e.g. accidentally avoiding certain groups during surveys and favouring others) by a good consistent sampling protocol, and well-structured experimental design (Dalgaard 2008; Hon 2010).

Statistical methods strongly rely on the concept that the observed dataset that is sampled is a valid representative of a larger (unknown) ‘population’. A good sampling design (also called ‘casing’ in social sciences) helps to avoid known biases. Biased data are data with a disproportionate weight in favour of or against one thing, person or group, so that the data are no longer a perfect representative of the larger population about which conclusions are being drawn. Sampling biases often result from uneven or patchy sampling, or from an underlying preconceived idea of the structure of the unknown population that results in the wrong type of sampling strategy. In time series analysis, bias can also arise when the variance of errors (random noise) changes over time, or when the data distribution is not normal, potentially invalidating results when the wrong analysis method is applied. Reducing bias helps to enable better discovery of underlying patterns and detection of relationships in an SES. Systematic biases may be reduced through increased sample sizes, increased randomisation of sampling (locations) or better selection of the sampling strategy (e.g. avoidance of explicit inclusion or exclusion of certain groups or system constellations) that could bias the data (see Crewson 2006; Dalgaard 2008; Hon 2010; Crawley 2015).

Thus, the overall aim should be, generally, to reduce the so-called ‘sampling error’, which denotes the gap between the sampled data as a representative and the current real status of the system. Because statistical methods are very sensitive to biased, patchy and noisy data, the outcome of most statistical approaches depends mainly on the overall quality of the investigated data.

Data collection therefore aims to collect unbiased data with as little noise as possible (Brown and Saunders 2007). However, high-quality data are often difficult to achieve in appropriate (temporal and spatial) resolution with fully excluding potential (often unknown) biases. Empirical data obtained from fieldwork often barely cover the minimum required representative data needed, as data are often either difficult to obtain given personnel capacity and time constraints, are simply not available, or are available in limited quantities.

In some situations, accurately estimating the effect or strength of specific explanatory variables requires experimental comparison (in the form of ‘before–after’ situations or as intentional experimental manipulations). However, in many real-world situations, and at the scale that most SES operate, experimental exclusion of a given factor is realistically impossible (e.g. removing the effect of global warming in order to see its influence, or excluding all insects from a large region to estimate the contribution of pollinators).

Moreover, results of most statistical methods focus on average outcomes, patterns or relationships in a dataset and cannot be used to draw conclusions about individual characteristics or cases. Another problem with statistical analysis is the tendency to jump to unjustified conclusions about causal relationships. One can often find evidence that two variables are highly correlated, but that does not prove that one variable actually causes another. A strong statistical relationship in the past can also not be assumed to hold in the future if conditions change. Statistical analysis is not capable of proving a causal relationship between two variables but requires a good understanding of the system investigated by the investigator (Pearl 2009).

Lastly, even with perfect data, the incorrect application of a method may result in drawing incorrect conclusions, as every method can only be applied to certain data types and under certain assumptions. Thus, the analysing researcher must have knowledge of the data type and structure to ensure that the method or test is appropriate and valid.

Resource implications

Complex statistical methods require computers and specialised software to conduct calculations or create graphical visualisations of the data and results. In the past, accessibility and computational power were the main constraints for rigorous and complex analyses of large datasets or the application of advanced statistical methods. However, given the technical capabilities of current computers, the application of statistical methods on even medium- to large-size datasets (up to several gigabytes) generally does not require specialised hardware or computational power any more. Also, given the multitude of free statistical software packages, specialised commercial software is rarely needed, as most free-of-charge statistical software packages now cover a wide range of statistical methods typically applied in SES research. Freely available open-source software includes, for example, R, Octave and Pandas, while many universities also provide access to commercially available products such as Statistica, SPSS, SAS, JMP and Strata. More specialised software packages such as JmulTi and Microfit (econometric analysis), OpenEpi (epidemiology), Simfit (simulations) or OpenNN, Torch or Weka (machine learning) are also freely available (see en.wikipedia.org/wiki/List_of_statistical_packages for a full list).

Many universities also offer statistical resource centres to assist students and researchers, given the specialised and highly technical aspects of applying many statistical methods. These include: (a) posing a correctly formulated research question (that can create a null hypothesis, that then can be falsified), (b) choosing the appropriate sampling design and type of data needed to answer the question, (c) having expertise in data acquisition, manipulation and interpretation, (d) selecting the appropriate statistical method for the type of data, and (e) being able to correctly interpret statistical results, which often requires significant expertise and understanding of statistical methods.

There are also a multitude of information courses (e.g. at DataCamp, Udemy and other online learning platforms) and tutorials available online (e.g. KhanAcademy.org, R-project.org).

New directions

Increasing computational power and accessibility of data from large online databases enables the analysis of ‘big data’ to test SES relationships and novel hypotheses (see also Chapter 17). ‘Big data’ are created by increasing numbers of electronic sensors, the automatic collection of data available through the Internet, and more long-term studies and national statistics data (e.g. the World Bank, the UN and other databases) being made available online (Bhadani and Jothimani 2016). The increasing number of people with mobile devices, active participation in social media and Internet use means data on human social interactions have become much more accessible. Analysis of social network data has recently radically transformed the way applied social science is done (Foster et al. 2016), allowing for better predictions of social dynamics, human decision-making and the potential consequences (Thai, Wu, and Xiong 2017).

A completely different field is the increasing accessibility of geophysical data by monitoring satellites and satellite-based sensor techniques (provided by e.g. NASA, ESA or platforms such

Case study 18.1: Drivers of ecosystem vulnerability in Andalusia, Spain

Different statistical methods can be used to unravel information about social-ecological dynamics and associations between social and ecological aspects. This is well illustrated by SES research conducted in the semi-arid watersheds of Adra and Nacimiento in Andalusia, Spain. The aim of this research was to find patterns of social-ecological interactions between stakeholders' perceptions of ecosystem services, stakeholder well-being and drivers of change (Iniesta-Arandia et al. 2014).

Altogether 381 face-to-face questionnaires were administered from May 2009 to February 2010. Based on these surveys, different groups of stakeholders who used and managed ecosystem services in each of the watersheds were identified by conducting hierarchical cluster analysis (HCA) and principal component analysis (PCA) (Figure 18.1A). Five stakeholder groups were identified: (a) local actors who depend on provisioning ecosystem services, (b) local actors not directly dependent on provisioning ecosystem services, (c) environmental and local development professionals, (d) rural tourists, and (e) nature tourists.

Next, different perceptions of vulnerable ecosystem services according to the five groups of stakeholders were explored by conducting the non-parametric Kruskal-Wallis test. This test showed that there were statistically significant differences in the perceptions about vulnerable ecosystem services among stakeholders. Local actors perceived a higher degree of vulnerability of provisioning ecosystem services, whereas environmental professionals perceived a higher degree of vulnerability of cultural ecosystem services than other stakeholders (Figure 18.1B).

Finally, to determine whether there were associations between vulnerable ecosystem services, human well-being and the effect of drivers of change according to the perceptions of different stakeholders, a principal component analysis (multivariate ordination analysis) was conducted. This statistical analysis showed that whereas local actors perceived the importance and vulnerability of provisioning ecosystem services for their well-being, environmental professionals perceived cultural ecosystem services, specifically the aesthetic values of landscape and local ecological knowledge, and soil fertility as vulnerable services. Locals also perceived the effects of drivers of change on ecosystem services. The principal component analysis showed that nature tourists perceived regulating services as vulnerable ecosystem services (Figure 18.1C).

Results from this study suggest that perceptions of the interlinkages between ecosystem services, human well-being and the effect of drivers of change on ecosystem services differ among stakeholder groups in the Adra and Nacimiento watersheds.

Figure 18.1 (A) A dendrogram showing the different stakeholder groups based on a hierarchical cluster analysis, (B) a bar diagram illustrating the perceptions of vulnerable ecosystem services among stakeholder groups, and (C) a biplot of the principal component analysis to find patterns between the associations (between stakeholders and their perceptions of vulnerable ecosystem services) and the effect of drivers of change (Iniesta-Arandia et al. 2014)

as Google's Earth Engine or Earth System Data Cube – see Chapter 24). These developments have helped to see SES interactions and problems in a whole new light (Rejcek 2017). Hampton et al. (2013) showed how the statistical analysis of large ecosystem dynamics data could help understand SES interlinkages. The analysis is done by aggregating, searching, cross-referencing and mining large volumes of data. The resulting outcomes finally generate new understanding that can inform decision-making about emergent properties of complex systems. Dubey et al. (2019) showed how big datasets and predictive analytics could improve social and environmental sustainability by combining data from 205 studies conducted in a region in India.

The analysis of large datasets may lead to erroneous conclusions and unexpected failures if blind trust is placed in the sheer amount of data available. Theoretical understanding and knowledge of the system remain fundamental to good, rigorous analysis. Furthermore, 'big data' are often strongly biased towards developed countries (Reis, Braatz, and Chiang 2016). Satellite or environmental sensor data are, thus, often only available for risk areas in developed Western countries at usable temporal and spatial resolution for conducting profound analyses. In contrast, these types of data are often lacking for particular SES risk regions of interest, especially in many undeveloped countries. Similarly, high-quality social media and trading data are only available in industrialised and rich countries that rely heavily on Internet-mediated communication for social interactions and trading.

Another important emerging statistical field is the use of machine learning. In the wake of big data, artificial intelligence methods are opening new approaches to investigating complex problems. Machine-learning approaches allow for statistical pattern recognition in datasets of very high complexity (see Chapter 17). However, these methods are typically not based on predefined verifiable statistical models and hypotheses, and often act as 'black boxes'. Typical neural network approaches use dynamically adaptive statistical models and therefore do not necessarily provide outcomes that can contribute to the logical understanding of interactions (Kleinberg, Ludwig, and Mullainathan 2016; Knight 2017; Pearl and Mackenzie 2018). However, novel analysis methods based on machine learning enabled by better computational grids allow for much faster analysis of data and near-instantaneous and automated decision-making. This is habitually done in economic trading when it comes to the buying and selling of commodities and stocks. These analysis methods may open up new possibilities for political and other national and international stakeholders to become better informed.

Some studies even suggest that artificial intelligence approaches and machine learning are ushering us into a new era in which abundant data and mathematics will replace classical statistical analysis methods (Anderson 2008). While scientists in the past had to rely on sample testing and statistical analysis to understand a process, computer scientists today have access to the 'entire population' (not only a 'representative sample') and therefore would not need classical statistical analysis methods any more. Thus, some authors speculate that traditional statistical testing approaches could become obsolete (Anderson 2008).

Key readings

- Bors, D.A. 2018. *Data Analysis for the Social Sciences: Integrating Theory and Practice*. Thousand Oaks: Sage.
- Brown, R.B., and M.P. Saunders. 2007. *Dealing with Statistics: What You Need to Know*. Berkshire: McGraw-Hill.
- Crawley, M.J. 2015. *Statistics: An Introduction Using R*. Chichester: John Wiley & Sons.
- Pearl, J. 2003. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.
- Quinn, G.P., and M.J. Keough. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press.

Acknowledgements

Ingo Fetzer acknowledges funding from the European Research Council under EU H2020 (grant no. ERC-2016-ADG 743080). Luigi Piemontese was funded by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning FORMAS (942-2015-740). Juan C. Rocha also received support from the FORMAS grant (942-2015-731). Ingo Fetzer, Luigi Piemontese and Juan C. Rocha especially thank the Stockholm Resilience Centre for support that enabled the writing of this chapter. The authors also especially want to thank Dr Goeff Wells for his fantastic feedback, general help and useful references to this chapter.

References

- Anderson, C. 2008. 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.' *Wired*. www.wired.com/2008/06/pb-theory.
- Angrist, J.D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.
- Bhadani, A., and D. Jothimani. 2016. 'Big Data: Challenges, Opportunities and Realities.' In *Effective Big Data Management and Opportunities for Implementation*, edited by M.K. Singh and D.G. Kumar, 1–24. Pennsylvania: IGI Global. <http://arxiv.org/abs/1705.04928>.
- Brown, R.B., and M.P. Saunders. 2007. *Dealing with Statistics: What You Need to Know*. Berkshire: McGraw-Hill.
- Carvalho, F.P. 2017. 'Pesticides, Environment, and Food Safety.' *Food and Energy Security* 6: 48–60. doi:10.1002/fes3.108.
- Cinner, J.E., C. Huchery, E.S. Darling, A.T. Humphries, N.A.J. Graham, C.C. Hicks, N. Marshall, and T.R. McClanahan. 2013. 'Evaluating Social and Ecological Vulnerability of Coral Reef Fisheries to Climate Change.' *PLoS ONE* 8(9): e74321. doi:10.1371/journal.pone.0074321.
- Crawley, M.J. 2015. *Statistics: An Introduction Using R* (2nd ed). Chichester: John Wiley & Sons. www.wiley.com/en-se/Statistics%3A+An+Introduction+Using+R%2C+2nd+Edition-p-9781118941096.
- Crewson, P. 2006. 'Applied Statistics Handbook.' AcaStat Software. www.acastat.com.
- Cummings, E.M., C.E. Merrilees, A.C. Schermerhorn, M.C. Goeke-Morey, P. Shirlow, and E. Cairns. 2010. 'Testing a Social Ecological Model for Relations between Political Violence and Child Adjustment in Northern Ireland.' *Development and Psychopathology* 22(2): 405–418. doi:10.1017/S0954579410000143.
- Dalgaard, P. 2008. *Introductory Statistics with R* (2nd ed). New York: Springer. doi:10.1007/978-0-387-79054-1.
- Dearing, J.A., R. Wang, K. Zhang, J.G. Dyke, H. Haberl, M.S. Hossain, P.G. Langdon et al. 2014. 'Safe and Just Operating Spaces for Regional Social-Ecological Systems.' *Global Environmental Change* 28: 227–238. doi:10.1016/j.gloenvcha.2014.06.012.
- Dubey, R., A. Gunasekaran, S.J. Childe, T. Papadopoulos, Z. Luo, S.F. Wamba, and D. Roubaud. 2019. 'Can Big Data and Predictive Analytics Improve Social and Environmental Sustainability?' *Technological Forecasting and Social Change* 144: 534–545. doi:10.1016/j.techfore.2017.06.020.
- Fienberg, S.E. 1992. 'A Brief History of Statistics in Three and One-Half Chapters: A Review Essay.' *Statistical Science* 7(2): 208–225. doi:10.1214/ss/1177011360.
- Foster, I., R. Ghani, R.S. Jarmin, F. Kreuter, and J. Lane. 2016. *Big Data and Social Science: A Practical Guide to Methods and Tools*. Boca Raton: CRC Press.
- González-Orozco, C.E., A.H. Thornhill, N. Knerr, S. Laffan, and J.T. Miller. 2014. 'Biogeographical Regions and Phytogeography of the Eucalypts.' *Diversity and Distributions* 20(1): 46–58. doi:10.1111/ddi.12129.
- Hampton, S.E., C.A. Strasser, J.J. Tewksbury, W.K. Gram, A.E. Budden, A.L. Batcheller, C.S. Duke, and J.H. Porter. 2013. 'Big Data and the Future of Ecology.' *Frontiers in Ecology and the Environment* 11(3): 156–162. doi:10.1890/120103.
- Holmes, E.E., E.J. Ward, and K. Wills. 2012. 'MARSS: Multivariate Autoregressive State-space Models for Analyzing Time-series Data.' *The R Journal* 4(1): 30.
- Hon, K. 2010. *An Introduction to Statistics*. Scotts Valley: CreateSpace.

- Husak, G.J., C.C. Funk, J. Michaelsen, T. Magadzire, and K.P. Goldsberry. 2013. 'Developing Seasonal Rainfall Scenarios for Food Security Early Warning.' *Theoretical and Applied Climatology* 114(1–2): 291–302. doi:10.1007/s00704-013-0838-8.
- Iniesta-Arandia, I., M. García-Llorente, P.A. Aguilera, C. Montes, and B. Martín-López. 2014. 'Socio-cultural Valuation of Ecosystem Services: Uncovering the Links between Values, Drivers of Change, and Human Well-Being.' *Ecological Economics* 108: 36–48. doi:10.1016/j.ecolecon.2014.09.028.
- Jaramillo, F., and G. Destouni. 2014. 'Developing Water Change Spectra and Distinguishing Change Drivers Worldwide: Worldwide Water Change Spectra.' *Geophysical Research Letters* 41(23): 8377–8386. doi:10.1002/2014GL061848.
- Keane, A., J.F. Lund, J. Bluwstein, N.D. Burgess, M.R. Nielsen, and K. Homewood. 2019. 'Impact of Tanzania's Wildlife Management Areas on Household Wealth.' *Nature Sustainability*. doi:10.1038/s41893-019-0458-0.
- Kim, J.I., and G. Kim. 2018. 'Effects on Inequality in Life Expectancy from a Social Ecology Perspective.' *BMC Public Health* 18(1): 243. doi:10.1186/s12889-018-5134-1.
- Kleinberg, J., J. Ludwig, and S. Mullainathan. 2016. 'A Guide to Solving Social Problems with Machine Learning.' *Harvard Business Review*. <https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning>.
- Knight, W. 2017. 'There's a Big Problem with AI: Even Its Creators Can't Explain How It Works.' *MIT Technology Review*. www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai.
- Kotzee, I., and B. Reyers. 2016. 'Piloting a Social-Ecological Index for Measuring Flood Resilience: A Composite Index Approach.' *Ecological Indicators* 60: 45–53. doi:10.1016/j.ecolind.2015.06.018.
- Maione, C., D.R. Nelson, and R.M. Barbosa. 2018. 'Research on Social Data by Means of Cluster Analysis.' *Applied Computing and Informatics*. doi:10.1016/j.aci.2018.02.003.
- Martín-López, B., I. Iniesta-Arandia, M. García-Llorente, I. Palomo, I. Casado-Arzuaga, D.G. del Amo, E. Gómez-Baggethun et al. 2012. 'Uncovering Ecosystem Service Bundles through Social Preferences.' *PLoS ONE* 7(6): e38970. doi:10.1371/journal.pone.0038970.
- Martín-López, B., I. Palomo, M. García-Llorente, I. Iniesta-Arandia, A.J. Castro, D.G. del Amo, E. Gómez-Baggethun, and C. Montes. 2017. 'Delineating Boundaries of Social-Ecological Systems for Landscape Planning: A Comprehensive Spatial Approach.' *Land Use Policy* 66: 90–104. doi:10.1016/j.landusepol.2017.04.040.
- Meacham, M., C. Queiroz, A. Norström, and G. Peterson. 2016. 'Social-Ecological Drivers of Multiple Ecosystem Services: What Variables Explain Patterns of Ecosystem Services across the Norström Drainage Basin?' *Ecology and Society* 21(1): 14. doi:10.5751/ES-08077-210114.
- Oksanen, J. 2015. 'Multivariate Analysis of Ecological Communities in R: Vegan Tutorial.' <http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf>.
- Oteros-Rozas, E., B. Martín-López, J. González, T. Plieninger, C. López-Santiago, and C. Montes. 2014. 'Socio-cultural Valuation of Ecosystem Services in a Transhumance Social-Ecological Network.' *Regional Environmental Change* 14: 1269–1289. doi:10.1007/s10113-013-0571-y.
- Paumgarten, F., and C. Shackleton. 2011. 'The Role of Non-Timber Forest Products in Household Coping Strategies in South Africa: The Influence of Household Wealth and Gender.' *Population and Environment* 33(1): 108–131. www.jstor.org/stable/41487565.
- Pearl, J. 2009. *Causality: Models, Reasoning and Inference* (2nd ed). Cambridge: Cambridge University Press.
- Pearl, J., and D. Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books. <http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1592572>.
- Piemontese, L., G. Castelli, I. Fetzer, J. Barron, H. Liniger, N. Harari, E. Bresci, and F. Jaramillo. 2020. 'Estimating the Global Potential of Water Harvesting from Successful Case Studies.' *Global Environmental Change* 63. doi:10.1016/j.gloenvcha.2020.102121.
- Pinsky, M.L., and M.I. Fogarty. 2012. 'Lagged Social-Ecological Responses to Climate and Range Shifts in Fisheries.' *Climatic Change* 115(3): 883–891. doi:10.1007/s10584-012-0599-x.
- Pollnac, R., P. Christie, J.E. Cinner, T. Dalton, T.M. Daw, G.E. Forrester, N.A.J. Graham, and T.R. McClanahan. 2010. 'Marine Reserves as Linked Social-Ecological Systems.' *Proceedings of the National Academy of Sciences* 107(43): 18262–18265. doi:10.1073/pnas.0908266107.
- Quinn, G.P., and M.J. Keough. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press.

- Reis, M.S., R.D. Braatz, and L.H. Chiang. 2016. 'Big Data: Challenges and Future Research Directions.' www.aiche.org/resources/publications/cep/2016/march/big-data-challenges-and-future-research-directions.
- Rejcek, P. 2017. 'The Farms of the Future Will Be Automated from Seed to Harvest.' *Singularity Hub* (blog). 30 October. <https://singularityhub.com/2017/10/30/the-farms-of-the-future-will-run-on-ai-and-robots>.
- Robinson, B.E., H. Zheng, and W. Peng. 2019. 'Disaggregating Livelihood Dependence on Ecosystem Services to Inform Land Management.' *Ecosystem Services* 36(100902). doi:10.1016/j.ecoser.2019.100902.
- Sabo, J., A. Ruhi, G.W. Holtgrieve, V. Elliott, M.E. Arias, P.B. Ngor, T.A. Räsänen, and S. Nam. 2017. 'Designing River Flows to Improve Food Security Futures in the Lower Mekong Basin.' *Science* 358(6368): eaao1053. doi:10.1126/science.aao1053.
- Scheffer, M., S.R. Carpenter, T.M. Lenton, J. Bascompte, W. Brock, V. Dakos, J. van de Koppel et al. 2012. 'Anticipating Critical Transitions.' *Science* 338(6105): 344–348. doi:10.1126/science.1225244.
- Shongwe, M.E., G.J. van Oldenborgh, B. van den Hurk, and M. van Aalst. 2011. 'Projected Changes in Mean and Extreme Precipitation in Africa under Global Warming. Part II: East Africa.' *Journal of Climate* 24(14): 3718–3733.
- Thai, M.T., W. Wu, and H. Xiong, eds. 2017. *Big Data in Complex and Social Networks*. Boca Raton: CRC Press.
- Varberg, D.E. 1963. 'The Development of Modern Statistics.' *The Mathematics Teacher* 56(4): 252–257.
- Wilminck, F.W., and H.T. Uytterschaut. 1984. 'Cluster Analysis, History, Theory and Applications.' In *Multivariate Statistical Methods in Physical Anthropology: A Review of Recent Advances and Current Developments*, edited by G.N. van Vark and W.W. Howells, 135–175. Dordrecht: Springer. doi:10.1007/978-94-009-6357-3_11.