Project no. 006463
EuMon
EU-wide **mon**itoring methods and systems of surveillance for species and habitats of Community interest

Instrument: SSP

Thematic Priority: Biodiversity conservation

D2: Recommendations for survey design and data analysis

Due date of deliverable: 31 August 2005
Actual submission date: 31 Octobre 2005

Start date of project: 1.11.2004 Duration: 42 month

National Museum of Natural History (MNHN), Paris, France

Revision : 1

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission	
RE	Restricted to a group specified by the consortium (including the Commission	
CO	Confidential, only for members of the consortium (including the Commission Services)	**CO**

Analysing monitoring data: general methods, general questions

Recommendations for survey design and data analysis

Deliverable 2 of EuMon's Work Package 2.2

By Romain JULLIARD and Pierre-Yves HENRY

in collaboration with Jean CLOBERT, Frank DZIOCK, Klaus HENLE, Piotr NOWICKI and Marek SAMMUL

Table of Contents

(1) Objectives

Monitoring data over space and time have common properties and can be analysed with a typical set of analytical tools independent of the object studied (plants, animals, fungi) and of the type of measure (*cf.* part 3). There are of course exceptions that require particular methods, and these will be treated separately. The goal here is to describe the general method (Figure 1), identifying the biases accounted for, and the underlying assumptions that may be overlooked. Another aim is to review the usefulness, and disadvantages, of the many proposed refinements to the basic method. Appendices summarize the main technical considerations presented in the text. Identification of potential tasks for EuMon's work-packages are highlighted in the text as well as in appendices. General discussions of biodiversity monitoring principles can be found elsewhere (e.g. Noss, 1990, Yoccoz *et al.*, 2001, Balmford *et al.*, 2005, Mace, 2005).

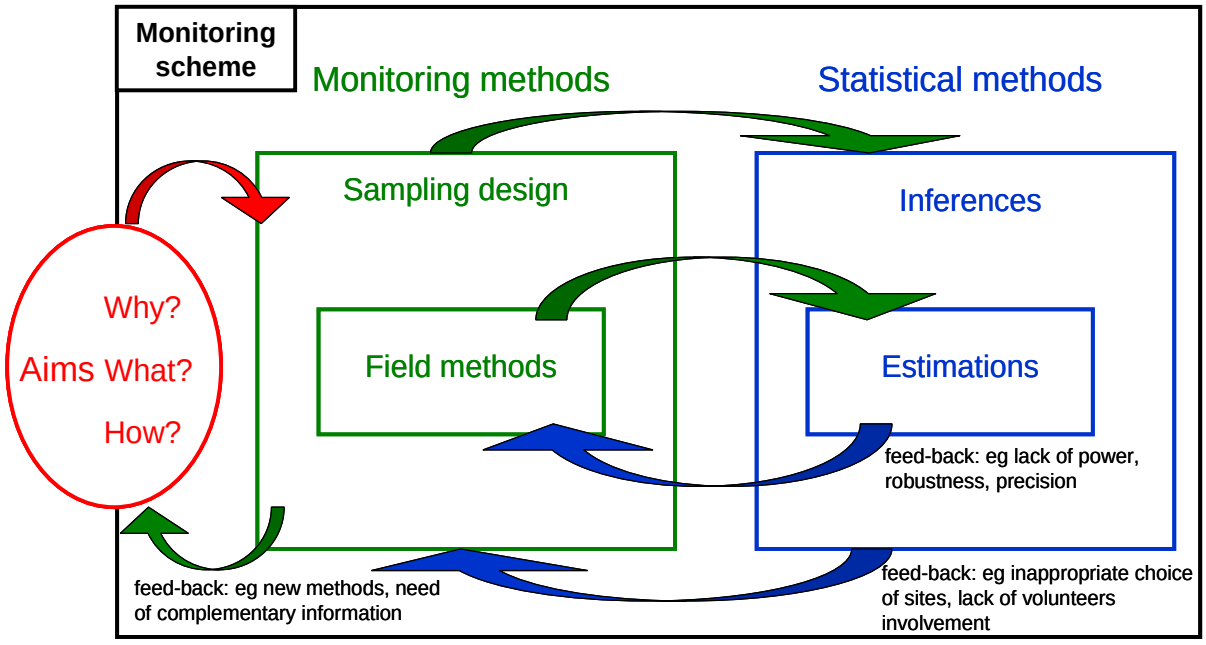

Figure 1: Basic principles of a monitoring scheme. The aim of a monitoring scheme is to answer specific questions, originating from external needs. Sampling designs and field methods (determining field data) are chosen according to objectives. Inferences that can be extracted from the monitoring scheme to answer the general questions of the scheme are directly dependent on the chosen sampling design; the same is true for the biological parameters that can be estimated from the collected data. Then, ideally, a monitoring scheme should be adaptive: monitoring methods are revised according to weaknesses identified during data analysis, and monitoring goals are revised according to new needs or more precisely defined questions.

(2) What is the usual statistical unit of monitoring? The measure, the site, the year

Usual monitoring data are made of three elements: the measure, the site, the year. The measure characterises the state of the entity considered (e.g., actual abundance of a given species; *cf.* part 3) at a particular site, in a particular year. Most analytical methods rely on the assumption that this is indeed the case. The validity of this assumption depends on the field methods, which are not considered here. Afterward, data analysis can hardly compensate for poor field methods.

Statistical analysis of trends are implemented either directly on field measures (e.g. number of individuals counted), or on indices that were computed from field data (e.g. age-ratio, diversity index), or on estimates of biological parameters (e.g. actual density after statistical correction for detection probability).

Indeed, ideal monitoring data should be made of four components: the measure, the site, the year, but also uncertainty. That is, the sampling design should allow the estimation of measurement error. If systematically quantifying uncertainty in the measure is too resource consuming, another solution is to incorporate in the analytical model independent estimates of the error to account for uncertainty in the measure. For instance, if detection probability is known to vary among sites, being twice higher in habitat 1 than in habitat 2, the estimates of detection probability per habitat should be included in the model that tests for among-habitat variation of the abundance index. Such statistical methods, integrating independent error measurements, are to be developed and promoted (task for WP2.2? WP5?).

(3) What to measure?

See Appendix 1 for specific recommendations: What level to monitor for what? What do we need to know to make recommendations on the levels to use?

Once objectives and the statistical unit are identified, the next question is: what do we need to monitor? Data type and methods are just a consequence of the goal we pursue (Figure 1). In general, the aim of monitoring is to determine changes in the status (distribution, abundance) for a given taxonomic group. According to our monitoring goal, we identify the monitoring methods to be used, and thus the biological parameters to be monitored (Figure 2; a more general overview of organisational levels of biodiversity can be found in Noss, 1990).

(i) Distribution. Distribution of species can be documented from compiling species lists from several sites. If species presence/absence was collected with an explicit spatial sampling design, data effectively document species distribution. They can also be used as surrogate of species abundance (although abundance indices may allow more robust and powerful inferences; task for WP2.2?).

(ii) Abundance. Monitoring abundance is rather straightforward, and relies on indices of local abundance (as developed hereafter). These indices can be converted in density estimates if information on detection probability is included in the sampling design.

(iii) Demographic processes. The underlying processes of changes in abundance can be disentangled by monitoring demographic parameters (reproduction, survival, emigration-immigration). For this, two main types of data are used: individual follow-up (e.g. capture-mark-recapture) and age (or size) structure of populations. Individual follow-up is to be preferred when possible because it makes less simplifying assumptions for parameter estimation, inferences are more robust, and variations in detection probability can be accounted for (task for WP2.2?). Demographic monitoring is usually time and resource consuming, and can be implemented for a restricted number of taxa and sites only.

(iv) Species assemblage. Multi-species presence/absence data can be used to document community dynamics and composition. The parameters of interest are parameters of community dynamics (colonization, extinction, turn-over) and indices of community composition (e.g. species richness, diversity, originality index, rarity index; cf. part 9). An open-question here is the interest of using closed species-lists (predetermined, restricted number of species to be recorded) *versus* opened species-lists (all sampled species are recorded). What do we gain in restraining the number of species to be recorded? Does it efficiently prevent saturation of the observer (task for WP1? WP5?)?

(v) Environmental parameters. If we are to identify pressures that drive identified trends, we also need to monitor environmental characteristics. We are here at the interface between species and habitat monitoring, and methods may be treated either by WP2 (species monitoring) or WP3 (habitat monitoring). Monitoring of changes in major biological processes, such as pollination, wood degradation, or flows of essential chemical elements, illustrate such species-habitat monitoring approaches (cf. e.g. ALARM project, http://www.alarm-project.ufz.de).

Genetic measures were omitted from the list of standard tools for biodiversity monitoring. Although they are highly relevant for monitoring long-term trends in abundance, population structure, or speciation (e.g. Noss, 1990), they do not seem to be cost-efficient enough for routine monitoring programs.

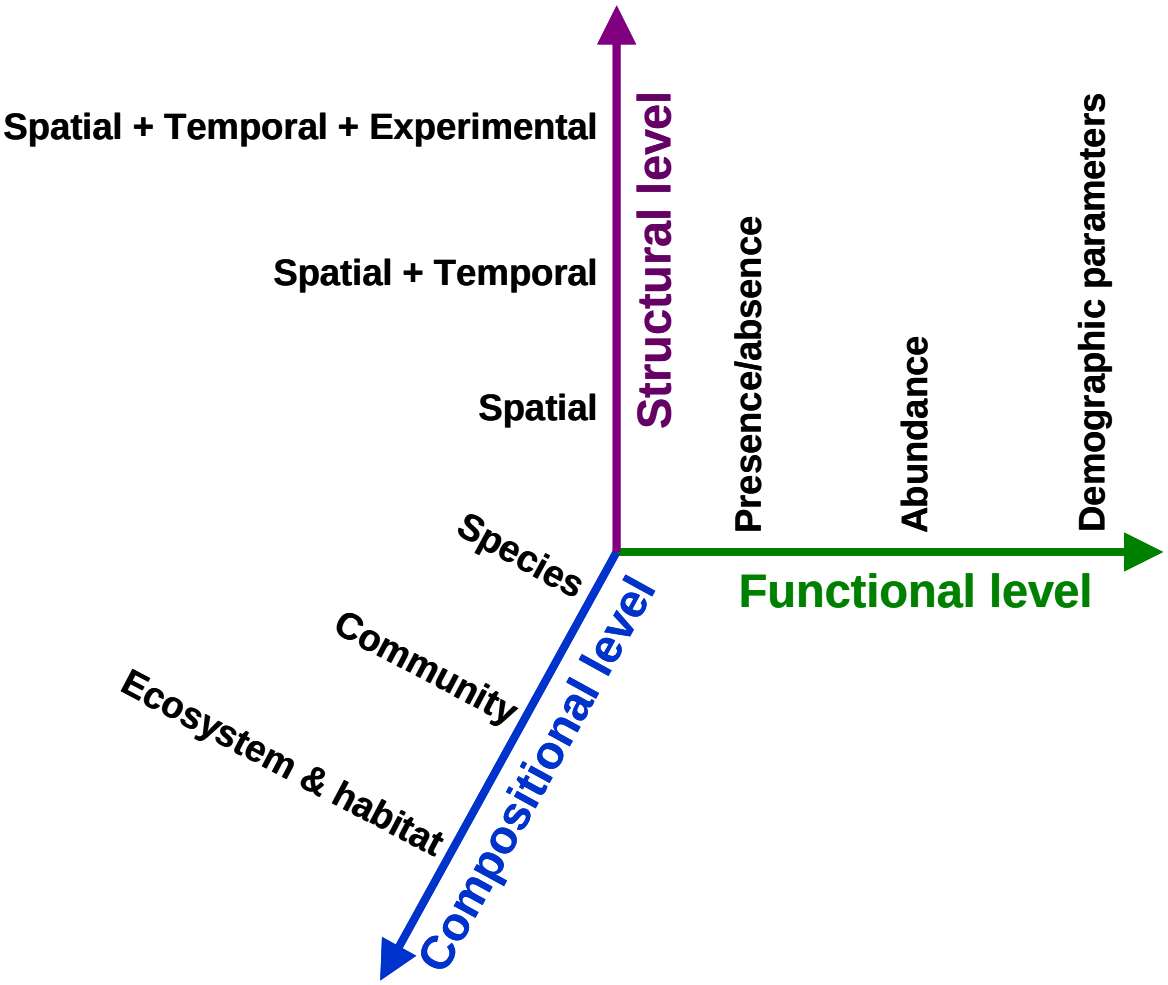

Figure 2. Main levels of biological complexity to be considered when designing a monitoring scheme. Composition level: species monitoring is quite straightforward, and widely implemented. Community monitoring is also common but already more complex as community dynamics involves more biological processes than just species dynamics. Habitat and ecosystem monitoring are the most complex levels of monitoring. Functions monitored are here characterised by the biological parameters we are able to routinely document: presence/absence of a (set of) species is the most basic information collected. For each species detected, the level of abundance can be quantified. The most complex level is monitoring demographic parameters per species. These levels of composition and function can be approached at different levels of structural complexity: just in space (distribution), in space and time (temporal variations in distribution), or complemented with experiments so that observed trends can be linked with explicit processes (usually, driving pressures in the context of biodiversity monitoring).

(4) A good sampling design for representative data

> See Appendix 2: What sampling design to monitor what? What do we need to know to make recommendations on sampling designs?

Sampling designing implies deciding how samples are to be distributed in space and in time, and if control samples have to be involved. Here, we discuss these three steps.

The different sampling sites are generally considered to be a representative sample of a larger area. This is an assumption on which inference methods rely. The validity of such an assumption relies itself on the sampling design, i.e. how sites were chosen to ensure spatial and temporal representativity. If the sampling design was imperfect, a correction may be used by weighting the different sites according to how representative they are. Although technical, this is rather straightforward and will not be further considered (e.g. Van Swaay *et al.*, 2002; task for WP2.2? WP5?). If the sampling design is not specified, then nothing can be done to statistically improve the representativity of the data.

An optimal design for biodiversity monitoring is explicit stratification in space, random choice of sampling sites per stratum, and systematic temporal survey. Let's consider the example of monitoring abundance through time of a species mainly occurring in pine forests, with two possible pine species. We would stratify space in three categories: no pine, pine 1, pine 2. Sampling sites would be randomly chosen within each stratum. The number of sampling sites per stratum would be proportional to the expected variation within each stratum (i.e. if there is twice more variation within pine 1 than within pine 2, the number of samples should be twice higher in pine 1 than in pine 2). Samples would be collected for a fixed number of times per year, distributed according to phenology. By this way, temporal trends are representative for all habitats, independently of their respective coverage, with the same precision of trends in all three habitats. Random choice of sites secures that micro-habitat diversity within strata is well represented. A classical deviation from this design is absence of spatial stratification. If sampling site location is fully randomly determined (without stratification), restricted habitats will be poorly sampled, and consequently trend estimates for these restricted habitats will have a low precision. Free choice of sampling sites by fieldworkers (instead of random choice) is usually – often unconsciously – oriented toward peculiar habitats (e.g. high naturalist interest, occurrence of locally rare species). Then, monitoring data are representative of habitats with these (unknown) specific properties; that is, we do not know what they are representative of!

In principle, temporal distribution of samples should be as refined as spatial distribution. However, despite we stratify and randomly choose the sites to be sampled in space, we generally distribute samples through time with a systematic design - every year, every two years, etc. To optimize sampling designs, and thus sampling effort, temporal structure of samples should also be considered (task for WP2.2?). For instance, long-lived organisms do not need to be monitored each year, since there is a strong inertia in their population dynamics. Or, if a method is manpower-consuming, sites to be sampled could randomly change from year to year to maximize simultaneously spatial and temporal representativity.

Another important question related with sampling designing is: Do we need a control treatment? If we want to monitor changes over large temporal and spatial scales, without *a priori* prediction on when, where or why a change should occur, no control treatment can be designed. Appropriate representativity of the data and statistical tools are sufficient for securing reliable measures of trend. But if the aim of the monitoring scheme is to quantify a change due to an *a priori* identified cause, we will have to distinguish the temporal and spatial

changes due to the investigated cause from normal fluctuations of the index. For instance, if the monitoring scheme aims at evaluating the impact of a change in environmental policy, then a control is required to be able to distinguish temporal changes due to the policy from other temporal trends. When a control treatment is needed, two experimental designs can be used. The actual use of a control, i.e. a set of samples that are not affected by the treatment of interest, is to be preferred. If actual control is not achievable, before/after comparisons, i.e. where the "control" is the time series before the application of the treatment, can be used with the weakness that before/after changes can be confounded by other temporal variations. When no control treatment can be designed (e.g. impact of climate change), experimentation under controlled conditions have to complement monitoring schemes so that changes due to the effect of interest can be separated from confounding changes.

Finally, a specific property of working with volunteers arises when designing sample collection: if a field protocol is too complex or time-consuming, or if sites to be sampled are unattractive, volunteers are naturally tempted to modify the protocol to fit their personal interest. If they depart from the protocol, then representativity of data is dangerously compromised. The work of WP1 will be extremely valuable here for understanding how to communicate with volunteers, and how to reward them, so that they are willing to follow potentially stringent sampling design.

(5) Missing points and among-site variations: classical problems compensated by statistical modelling

Ideally, measures are taken every year at every site. Achieving such a complete coverage of sites and years simplify statistical analyses. This is usually far from being achievable in schemes that rely on a large number of observers, sites and years. All monitoring sites do not start and end on the same year, and some sampling visits cannot be performed for unpredictable reasons. As a consequence, in most cases (thus, this is the general case!), some sites are not sampled in some years. In technical words, monitoring data include missing points (Figure 3). The problem with missing points is that eliminating incomplete time series reduces too much the data set to the expense of precision of the estimate and representativity. Fortunately, statistically accounting for missing points is not that complicated. A solution could be to work on the counts per sampling site rather than the sum over sites. For instance if 100 skylarks are counted on year 1 over 10 sites, and 80 on year 2 over 5 sites, the index would be 10 skylarks per site in year 1 and 16 skylarks per site in year 2. Here, a second problem arises: even with the same number of sites, the level of abundance per sampled sites (i.e., the expectancy on each site) likely varies from site to site. This is because abundance depends on habitat (via, habitat preferences of species and differences in habitat quality), and because the probability to detect species varies among habitats and observers (Figure 3). This makes sites not interchangeable.

A general solution to solve both problems of missing points and of site-specific abundance is to, at first, standardise (i.e. dividing or subtracting) site-specific time series by the average abundance over years observed per site (Figure 4). Then, it is legitimate to average these standardised counts over sites per year (which are year specific deviation from the site specific average, and thus account for local nuisance effects such as variations among habitats or observers). The average temporal trend in abundance can thus be estimated as the regression line of yearly sums of standardized counts against time (Figure 4).

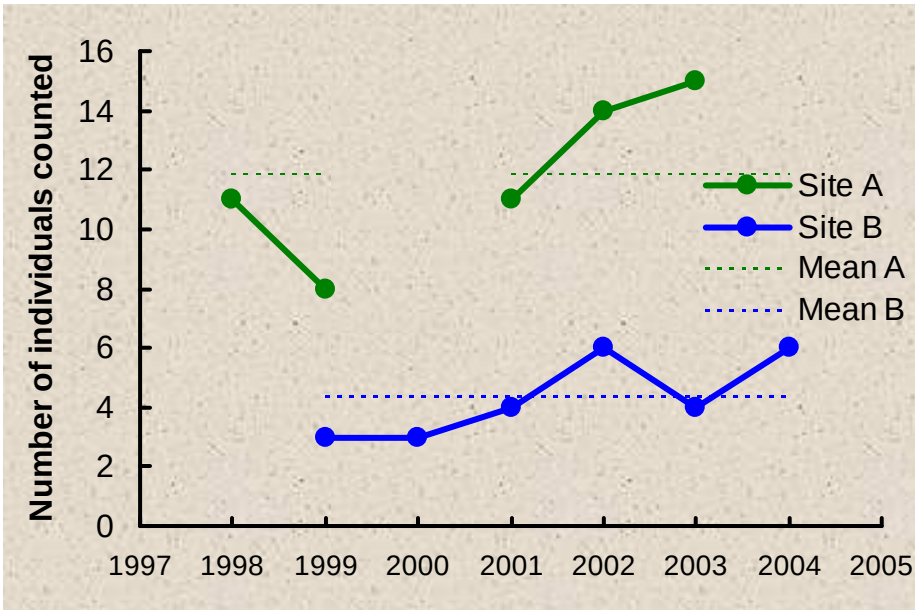

Figure 3. Example of typical monitoring data collected for a species, at two sites (A and B) that differ in average number of counts, and with missing counts.

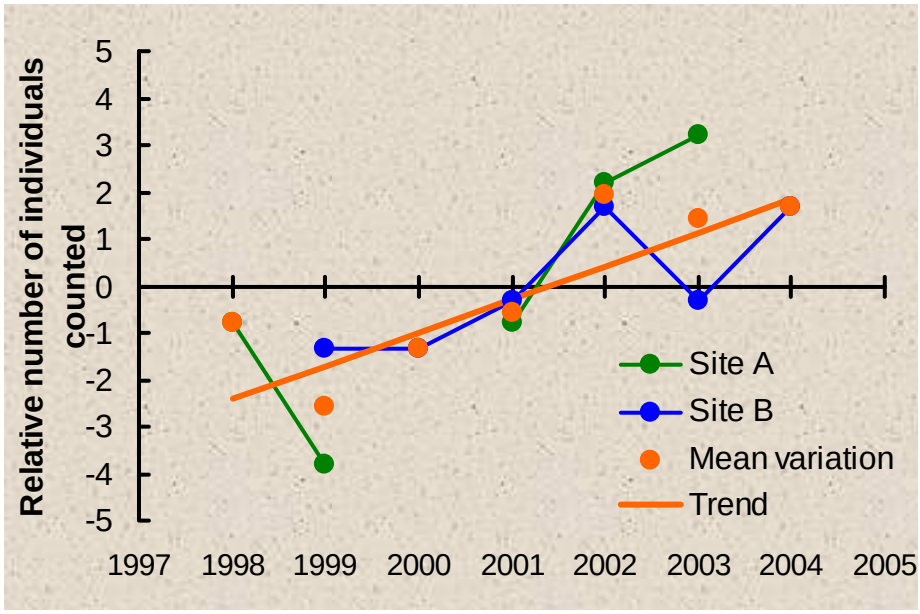

Figure 4. Temporal trend of the abundance index accounting for missing counts and for among-site variations in abundance and detection probability.

From a technical perspective, this means that monitoring data are analysed with a linear model that includes the effects of *n* sites and *t* years as independent variables. The *n* site effects are treated as categorical fixed effects with *n-1* degrees of freedom. The *t* year effects are treated as categorical fixed effects to quantify among-year variations, or as a quantitative linear effect to test for and estimate a temporal linear trend in abundance. Such a statistical treatment is correct when the measures are distributed normally. In many cases, the measure is known to follow another distribution: e.g., presence-absence data may follow a binomial distribution, or counts of individuals may follow a Poisson distribution. In such cases, the

appropriate error has to be defined when building the regression model, and estimation and tests will then rely on Maximum Likelihood methods.

For count data, the appropriate statistical model is log-linear (or Poisson) regression (cf. implementation in software TRIM specifically designed for count data; Pannekoek & van Strien, 2001). This implies the use of a Ln link-function, what has a very interesting property in the context of studying temporal changes in abundance: the dependent variable is no more counts of individuals (as shown in Figure 4) but the inter-annual growth rates of counted populations. The results thus have a straightforward biological meaning: we obtain estimates of population growth rate, rather than arithmetic changes in number of individuals. This is ideal if we assume that most demographic parameters are multiplicative (i.e., rates) rather than additive (e.g., $N_{t+1} = N_t - N_{deaths} + N_{births} - N_{emigrants} + N_{immigrants}$). Another interesting property is that the precision of counts is accounted for (i.e. sampling error): a change from 1 to 2 counted individuals is less informative – i.e. more imprecise – than a variation from 10 to 20, although both changes correspond to the same growth rate.

In summary, linear modelling, with appropriate distribution for data, link-function and parameterization of sites and year effects, intrinsically accounts for the main problems faced when analysing temporal series of monitoring data: heterogeneity among sites, among observers, through time and in precision. Three basic properties are worth highlighting. First, the computation of temporal trends does not require complete time series; missing counts are accounted for. Second, including site effects in the statistical model largely accounts for differences in detection ability among observers and habitats. For this reason, it is strongly recommended that each site is monitored by the same observer, as long as he/she is involved in the monitoring scheme (task for WP5?). Third, the trend in abundance is not the simple difference with the first or the last year of the monitoring, but instead data of all years equally contribute to the trend.

(6) From species monitoring to community monitoring

From counts of individuals, we derived indices of abundance and estimated trends in abundance. From individual follow-up data, we estimated the demographic parameters responsible for observed trends in abundance. The underlying biological processes are explicit and come from a strong theoretical background of population dynamics. From these same counts of individuals, we can also derive indices of community composition and community dynamics. A developing set of tools for community monitoring is the use of capture-mark-recapture modelling (Yoccoz *et al.*, 2001; Pollock *et al.*, 2002). The use of these models allows a direct extrapolation of metapopulation dynamics theory to the understanding of community dynamics; and it takes profit of the methodology developed to account for heterogeneity in detection probability. The unit of community composition is local species richness (i.e. the number of species per site and per year). Trend in community composition is estimated as the trend in local species turn-over. And drivers responsible for the trend are probabilities of local extinction and of local colonization. The advantage of this approach is that biologically meaningful parameters are monitored, with standard, robust statistics and with explicit interpretation for community dynamics theory. A weakness is that information on species abundance is completely ignored (task for WP5?).

Other indices of community composition aim at using simultaneously both information, species richness and species abundance. Classical indices are Shannon's index or Simpson's index. Their weakness is that they are not explicitly linked to community dynamics processes, and are just a standard mathematical way to combine information on species richness and relative species abundance without general biological meaning.

Another common set of community composition indices relies on grouping or weighting of species counts according to traits of interest for the purposes of the monitoring. Such indices were proposed, for instance, for specialization, originality, endemism, or community integrity. But again, their weakness is that they are poorly related to community dynamics theory, and they are generally designed to fit specific monitoring questions and not to fit understandable biological processes. A particular community approach that may escape this lack of biological meaning is monitoring of "functional composition" (e.g. Dolédec *et al.*, 1999). This relies on the theory of ecological niche. Basically, data are the same as for other indices, i.e. species identity and respective abundance indices, but species identity is decomposed into multiple biological traits (i.e. life-history traits, dispersal potential, food regime). Abundance indices are then used to weigh the contribution of each species to community characterisation. Temporal or spatial trends in community functional composition are interpreted as differences in ecosystem integrity.

On the whole, a multitude of indices of community composition have been proposed, loosening possibilities for comparative studies and integration across taxa and countries. An effort should be made to standardize community monitoring methods (task for WP2.2, WP3, WP5?), rooting decisions on indicators to be used in community dynamics theory.

(7) How to combine indices among species or countries?

Up to here, we discussed data for single species or single communities. But for biodiversity monitoring, we are not interested in single trajectories (apart when using bioindicator) but in general trends. We need synthetic indicators of temporal trends that are representative of whole taxa at large spatial scale (across national boundaries). To produce such integrated trends, we want to combine different time series of indices (Gregory *et al.*, 2005), or of estimates if data are too different (e.g. indices with different units but that still document a same process). A common way to do this is to compute the geometric mean of the yearly indices (the geometric mean between 50 and 200 being 100).

There arises the issue of differences in representativity of, e.g., species or countries. Are all species equal? Or should some species have a higher contribution to the global index than others? Methods for weighting species contribution can be developed to account for differences in specialisation, sensitivity, or economic potential. Within species, different countries generally host different population sizes. For instance, there are more brown bears in Slovenia than in Spain. Thus, if we are to compute the European trends for the brown bear, we have to give a higher weight to the Slovenian trend estimate than to the Spanish one. The contribution of each national index can be weighted by the percentage of the total European population size that is present in each country before computing geometric means.

When combining data, we also need to account for differences in precision. For instance, in countries with important manpower involvement, trend estimate will be more precise than in countries with poor participation. Poor representativity can be accounted for by weighting each separate index according to its associated standard error (indeed, the inverse of the squared standard error; formulae given in Gregory *et al.*, 2005).

Data combination is also an intrinsic need of "adaptive monitoring". When new methods become available or new needs are identified (Figure 1), scheme coordinators face a dilemma: is it better to change the protocol or to remain on a suboptimal design but with consistent data through time? To solve this dilemma, the response also comes from methods for data combination (task for WP5?). Basically, it turns to obtaining yearly trend estimates, with their standard deviations for both periods of the time series, i.e. past and new protocol. Then, the analysis of long-term trend is implemented not on raw data but on trend estimates (as for combination across countries).

As presented, data combination may seem rather straightforward. Indeed, current practices come from empirical approaches rather than specifically designed statistical methods. Statistical development of correct practices is thus needed. Several sources of complexity interplay: biological heterogeneity, spatial heterogeneity, and data heterogeneity inducing the need for statistical methods of increasing complexity (Figure 5). Optimal combination methods may not be the same depending on the level at which we need to combine information. Instead of listing combination methods for all levels of complexity, we suggest to, first, characterize what are the common monitoring practices (cf. D8, survey on monitoring schemes in Europe; task of WP2.1). Then, we will be able to identify and make methodological recommendations for specific needs for data combination. On the whole, these weighting issues are directly related to the problem of how to account for national responsibilities (task of WP4), and how to integrate indices or estimates from different schemes, organisms and spatial scales (task of WP5).

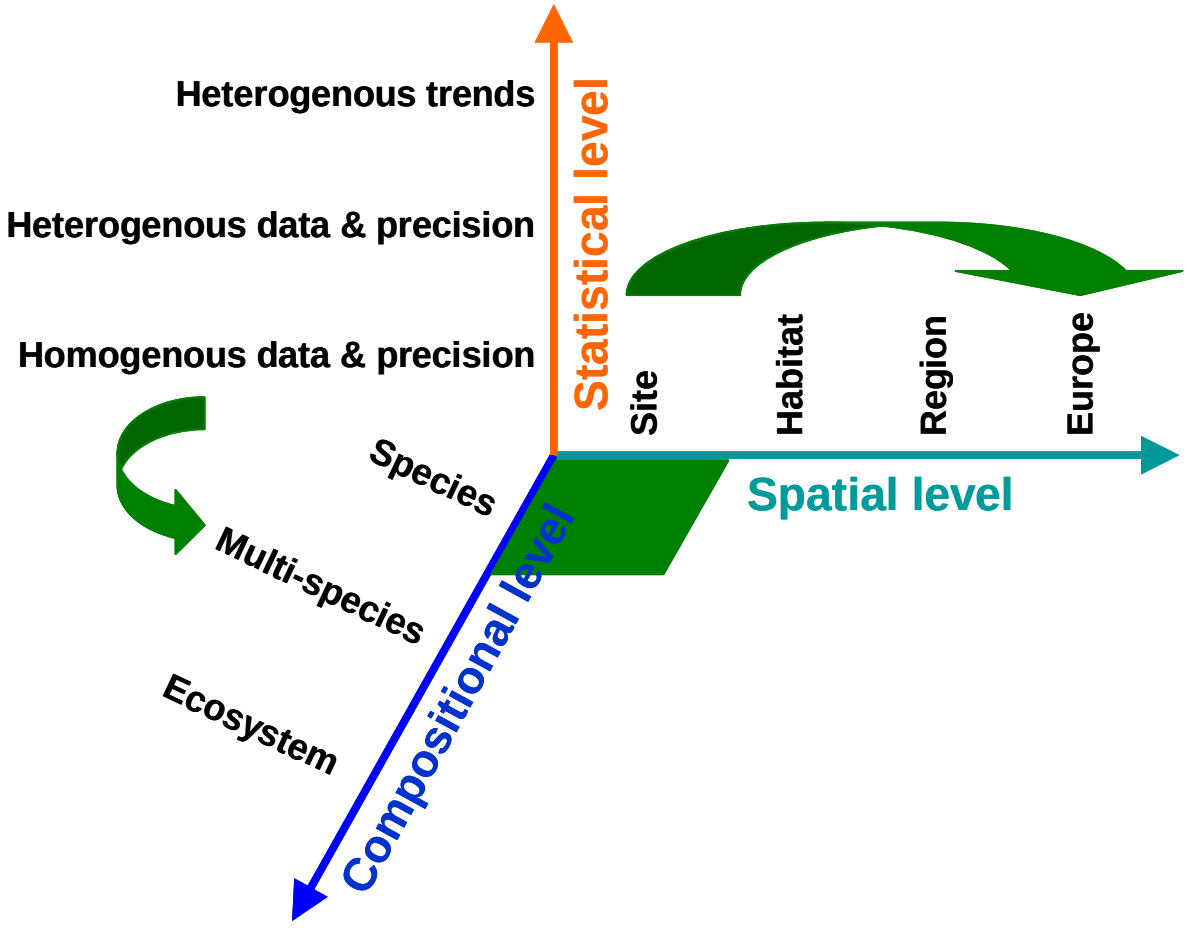

Figure 4. Data combination to produce biodiversity indicators is needed at different spatial and compositional levels. These multi-level needs involve heterogeneity in the data, what will translate in different levels of statistical complexity. Basically, data are collected at the site level and for a given (set of) species (green square). Some classical needs of data combination are aggregating data among species or among countries (green arrows). The corresponding statistical complexity depends on whether data were collected with a same method and same effort (i.e. homogenous data and precision), or with different methods, different sampling efforts, or if the biological signals to be combined among datasets are heterogenous (e.g. opposed trends in different countries).

(8) Some unrecognized problems of monitoring data analysis

See Appendix 3: What method for which statistical problem? What do we need to know to make recommendations on methods to use?

Interaction among explanatory variables. The aim of most monitoring schemes is to identify general temporal or spatial trends. But what generality can we reach when there is an interaction between the effects of site and time? That is when temporal trends vary through space; or spatial distribution varies through time. One may be tempted to ignore the interaction terms, and consider only main effects. But this is statistically incorrect - tests of main effects are spurious - and yields invalid conclusions. In front of a significant interaction, we have to identify the actual biological level at which homogeneity of the trend is verified (where main effects only are significant). For instance, if a species is found to be stable at some sites and decreasing at other sites, we need to identify the appropriate ecological variable(s) that discriminate these two sets of sites, like habitat, geographic or species characteristics. Standard methodologies to identify such levels of homogeneity, and to test main effects in presence of significant interaction terms, are to be designed and more widely used when analysing monitoring data (task for WP2.2 ?).

Using a Poisson distribution for log-linear regression relies on the assumption that counted individuals are independent of each other: the detection of one individual should not change the likelihood of detecting other individuals. This assumption does not hold as soon as individuals are **aggregated**. Consider a species that is usually found in flocks, once an individual is detected, it is likely that the whole flock will be counted. The 'let's count flocks then' approach is usually not satisfying since in most cases flocks have very different sizes, which the observer would like to be taken into consideration. The violation of the independence assumption induces a lack of fit of the Poisson distribution to the data. The consequences of this lack of fit are increased variation of the estimator (the estimator will typically go up and down erratically) and it will bias low error estimates (i.e., statistical significance of the variations are exaggerated). There are well known statistical procedures to correct for the latter problem: the lack of fit of the data is easily quantifiable (it is called over-dispersion of the data) and a correcting factor (often called c-hat) can be applied to the statistics which will increase the standard errors and reduce the significance of statistical tests. However, we are left with unreliable estimates with accordingly large standard errors. Empirically, levelling counts by a maximum appears to reduce over-dispersion and to buffer erratic variations. For instance, an extreme levelling would be to transform counts into presence/absence data (Royle & Nichols, 2003), although this may not be very profitable since population trend detection with presence/absence data suffers from low statistical power (Strayer, 1999). Rank-models may also be an interesting levelling procedure (in Thomas, 1996). Another solution would be to count separately flock number and number of individuals per flock. We are not aware of studies that compared precision and bias in trend estimates over these different solutions to account for individual aggregation. This would be worth investigating (task for WP2.2?).

Imperfect detection of individuals/species when counting is a well known problem in almost all monitoring data (e.g. Rosenstock *et al.*, 2002). If the detection probability is constant per species, for a given observer and a given site, then it should not affect estimates of temporal trends. However, if detection probability varies through time (e.g. temporal increase in experience of observers, decrease in detection ability of observer, changes of

habitat, Link & Sauer, 1998; Rosenstock *et al.*, 2002; Norvell *et al.*, 2003), it dangerously confound the trends we want to document with our monitoring data. Also, spatial variations in the identity of the observer bias inferences of spatial patterns of abundance because of spatial variation in detection probability. The worse case is if detectability varies with population size. Several methods have been designed to account for detection probability. A basic one is to include an observer effect in the analysis. Others are based on sampling designs where information is collected to model detection probability, thus transforming counts into density estimates (e.g. capture-mark-recapture methods, method to account for the distance at which individuals are detected; Pollock *et al.*, 2002; Rosenstock *et al.*, 2002; Bart *et al.*, 2004). General limits to these approaches are important field effort requirements and loss of statistical efficiency (overfitting of parameters, low accuracy of estimates; Thomas, 1996; Link & Sauer, 1997). Imperfect detection of individuals may not be a dramatic problem. It just makes that count data must not be considered (and analysed) as censuses or density estimates. Under explicitly stated and validated assumptions, they remain the best available indices of abundance at large spatial and temporal scales. Particularly, when estimating trends, many components of the detection probability can be controlled for in the sampling design (same sites, same observers, same dates of visit) and we can hope that the number of counted individuals is constantly linearly proportional to actual abundance (Bart *et al.*, 2004). An open-question is: from which level of detection probability variation do benefits overcome the costs of accounting for detection probability in the sampling design and in analysis (task for WP2.2?)?

Many different methods of analysis of temporal trend from monitoring data are available (reviewed in Thomas, 1996). Here, we presented the log-linear regression only. Advantages of other methods should also be mentioned. First, the **chain method** – i.e. chained ratio of successive counts – provide indices that are rapidly updated, and leave the index unchanged after the update. But this method is now regarded as inadequate because of spurious trend generation by random drift (year after year accumulation of errors) and inefficient use of the data (missing points prevent the use of data in the successive year). Second, procedures from **time series analysis** may be useful, particularly to account for temporal autocorrelation patterns (e.g. density-dependence, time lags). It remains to be evaluated if this is a major problem for large scale monitoring (Fewster *et al.*, 2000). Another limitation is that the usually short monitoring data series, with numerous missing points, are not suitable for time series analysis (Fewster *et al.*, 2000). Third, the prevailing trend may be far from linear over time (e.g. strong decrease, followed by steady increase). Then log-linear regressions would not identify the actual trend. In this case, **Generalized Additive Models** (GAM) are a powerful tool to test for non-linear trajectories, and to identify "turning points" in abundance dynamics (Fewster *et al.*, 2000). Nonetheless, a critical issue with these models is that it is hard to distinguish a trend (average rate of change) from a trajectory (an anecdotic pattern of temporal fluctuation). This is difficult because conclusions on the trend are very sensitive to changes in the smoothing parameter (Thomas, 1996). A further application of GAM is the modelling of within-year variations in counts to produce yearly abundance indices that account for phenological patterns of occurrence (Rothery & Roy, 2001). This requires monitoring data to be collected over several sampling sessions within a year (e.g. once per week in butterflies). It remains to be determined up to which degree of seasonal variation it is worth accounting for, and how this can be implemented in the context of among-years trend modelling (task for WP2.2?). Fourth, if sites to be monitored are randomly - instead of subjectively - selected, data could be analyzed with **mixed models**. Mixed models have the property of analysing the site effect no more as fixed unknowns (as in fixed-effects models) but as the realization of a parametric distribution whose parameters are to be estimated (referred to as random effect). The interesting property of such a modelling is that much less

parameters are required to account for among-site variations. Mixed models are more parsimonious than fixed-effects models (Fewster *et al.*, 2000). Applications of mixed models to trend analysis may thus be worth exploring (task for WP2.2?).

(9) References

Balmford, A., Crane, P., Dobson, A.P., Green, R.E. & Mace, G. 2005. The 2010 challenge: data availability, information needs and extraterrestrial insights. *Philos. Trans. R. Soc. Lond. Ser. B-Biol. Sci.* 360: 221-228.

Bart, J., Droege, S., Geissler, P., Peterjohn, B. & Ralph, C.J. 2004. Density estimation in wildlife surveys. *Wildl. Soc. Bull.* 32: 1242-1247.

Dolédec, S., Statzner, B. & Bournard, M. 1999. Special traits for future biomonitoring across ecoregions: patterns along a human-impacted river. *Freshwater Biology* 42: 737-758.

Fewster, R.M., Buckland, S.T., Siriwardena, G.M., Baillie, S.R. & Wilson, J.D. 2000. Analysis of population trends for farmland birds using generalized additive models. *Ecology* 81: 1970-1984.

Gregory, R.D., van Strien, A., Vorisek, P., Gmelig Meyling, A.W., Noble, D.G., Foppen, R.P.B. & Gibbons, D.W. 2005. Developing indicators for European birds. *Philos. Trans. R. Soc. Lond. Ser. B-Biol. Sci.* 360: 269-288.

Link, W.A. & Sauer, J.R. 1997. New approaches to the analysis of a population trends in land birds: comment. *Ecology* 78: 2632-2634.

Link, W.A. & Sauer, J.R. 1998. Estimating population change from count data: application to the North American breeding bird survey. *Ecol. Appl.* 8: 258-268.

Mace, G.M. 2005. *A user's guide to biodiversity indicators*, European Academies Science Advisory Council, The Royal Society, London.

Norvell, R.E., Howe, F.P. & Parrisk, J.R. 2003. A seven-year comparison of relative-abundance and distance-sampling methods. *Auk* 120: 1013-1028.

Noss, R.F. 1990. Indicators for monitoring biodiversity: a hierarchical approach. *Cons. Biol.* 4: 355-364.

Pannekoek, J. & van Strien, A.J. 2001. *TRIM 3 Manual (TRends & Indices for Monitoring data)*, CBS, Voorburg, The Netherlands.

Pollock, K.H., Nichols, J.D., Simons, T.R., Farnsworth, G.L., Bailey, L.L. & Sauer, J.R. 2002. Large scale wildlife monitoring studies: statistical methods for design and analysis. *Environmetrics* 13: 105-119.

Rosenstock, S., S., Anderson, D.R., Giesen, K.M., Leukering, T. & Carter, M.F. 2002. Landbird counting techniques: current practices and an alternative. *Auk* 119: 46-53.

Rothery, P. & Roy, D.B. 2001. Application of generalized additive models to butterfly transect count data. *J. Appl. Stat.* 28: 897-909.

Royle, J.A. & Nichols, J.D. 2003. Estimating abundance from repeated presence-absence data or point counts. *Ecology* 84: 777-790.

Strayer, D.L. 1999. Statistical power of presence-absence data to detect population declines. *Cons. Biol.* 13: 1034-1038.

Thomas, L. 1996. Monitoring long-term population change: why are there so many analysis methods ? *Ecology* 77: 49-58.

Yoccoz, N.G., Nichols, J.D. & Boulinier, T. 2001. Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution* 16: 446-453.

Appendix 1. Link between functional parameters to be monitored (lines) and measures to be taken (columns).

	Presence/absence	Counts of individuals	Age or size-structure	Individual follow-up (cf. Capture-Mark-Recapture)	Advantages	Disadvantages
Distribution	**Optimal**	Not used	Not used	Not used	Basic information required for status identification	Trends are detected late, after local extinction or colonisation only
Abundance	Appropriate but lower power to detect trends than counts of individuals	**Optimal**	Not used	Ideal but field intensive	Trends detected early, before local extinction or colonisation	No cues on demographic processes driving changes
Demographic processes	Appropriate for estimation of population growth rate inducing range extension / restriction only	Appropriate for population growth rate estimation only	Appropriate	**Optimal**	Detailed understanding of processes driving trends	Data consuming
Community processes	**Optimal**	Appropriate but theory to account for relative abundances in community parameters	To be developed	To be developed	Understanding of changes in biodiversity components across broad taxonomic groups	Community dynamics theory under development (task for WP2.2? WP3? WP5?)
Advantages	Large coverage because easy to implement	Large coverage because easy to implement	Intermediary level of detail	Highest level of detail		
Disadvantages	Poor precision (task for WP2.2?)	Limited information	Usually involves irrealistic simplifications for parameter estimation (task for WP2.2?) Intermediary coverage	Restricted coverage due to field intensivity		

General questions and limits:

- to monitor unbiased functional parameters, differences in detection probability should be accounted for at all levels;
- representativity of inferences remain dependent on sample design limitations (cf. part 4 and Appendix 2);
- when more than one type of measure can be taken for a functional parameter, what are the respective statistical power and sensitivity of each measure type (task for WP2.2?)?
- whenever changes of a functional parameter are to be linked to driving pressures, measures of environmental parameters are needed.

Appendix 2. Link between structure of sampling design in space and time, and expected representativity from the data (for inferences).

	Stratification	Choice of visit	Control treatment	Advantages	Disadvantages
Spatial sampling design	Yes if *a priori* identified main sources of variation among sites (e.g., habitats) Otherwise not achievable	Choice of site: Systematic or Random: appropriate. May differ in manpower requirement. Free choice: inappropriate	Control sites Experimental treatments	An explicit spatial design secures spatial representativity of data (e.g., for distribution analysis, among-habitat comparison)	Imposes constraints that may oppose volunteers' personal interest (task for WP1). Then, a compromise has to be found (mixture between stratification, random and free choice).
Within-year temporal sampling design	Yes if *a priori* identified main sources of variation throughout the year (e.g. phenology) Otherwise not achievable	Choice of time of visit: Systematic: easy to implement but may not cover representatively within-year variation Free choice: appropriate if observers behave independently of each other and of phenology, thus resulting in a similar coverage as random choice Random: appropriate but unusual	Before/after comparison Experimental treatments	Necessary when strong within-year variation in activity (and if no design is implemented to quantify detection probability per visit)	Can be manpower-consuming Requires optimization analysis
Among-year temporal sampling design	Usually no reason for among-year stratification. A counter example: long-lived species may be monitored every second or third year, whereas short lived species are monitored each year.	Choice of time of visit: Systematic: Every year: easy to implement, most usual Every x years: may not cover representatively among-year variation. Random in interaction with sites (some sites visited on some years): may be appropriate to optimize simultaneously spatial and temporal representativity Free choice: not relevant	Before/after comparison Experimental treatments	Unusual: to be developed?	If different from systematic on a year basis, it complicates maintaining volunteer involvement on the long term (WP1)

Advantages	Secures similar representativity and precision of all main sources of variation	Random & Systematic : easy to implement, optimal way to secure within-strata representativity Free choice: maximizes attractiveness for volunteers (task for WP1?)	Control & Before/After comparison: easy to implement Experiments: only way to identify drivers of trends if no control can be designed
Disadvantages	Need to have preliminary information	Systematic: may not provide data as representative as random choice (task for WP2.2?) Random: representativity is dependent on the number of samples relatively to variation expected within strata Free choice: representativity unknown, requiring post-stratification of data to compensate for under-sampling at unattractive sites or periods of year	Before/After comparison: can be confounded by contemporaneous changes Control & Experiments: usually not attractive / too constraining for volunteers (task for WP1?)

General limits and questions:

- If bias of the measure (e.g. detection probability) is to be quantified, the sampling design should include replicates within each sampling unit so that bias can be corrected for at all levels of the sampling design.
- The number of samples is also an issue of sampling designing. It should be proportional to among-strata variation, the more variable a strata, the higher the number of samples.
- Down to which level of pre-identified sources of variation should we go for stratification (task for WP2.2?)?
- Do random *versus* systematic choice of sites differ in terms of spatial representativity (task for WP2.2?)?

Appendix 3. Summary of some unrecognized statistical problems associated with monitoring data analysis

Three common statistical problems:

	Problem	Solution and questions
Interaction site*time	If temporal trends are different across sampling units, no general trend can be identified	Which method should be used to identify the compositional (e.g., species within community, population within species), or spatial (e.g; habitat) level at which trend homogeneity is achieved (task for WP2.2? WP4? WP5?)?
Aggregation	Lack of fit to the underlying distribution due to overdispersion, with artificially high precision and erratic temporal fluctuations of index	Which one of the following methods is optimal (task for WP2.2?): - correction factor for variance inflation (ĉ)? - levelling measures to a maximum? - rank-models? - decomposing into probability to be aggregated and number of individuals in aggregate?
Detection probability	Spatial and temporal variations confound the effects of interest (temporal or spatial trends)	Specific sampling designs and statistical models (e.g. capture-mark-recapture methods, distance sampling) exist but: - field effort is increased - precision is decreased - the number of parameters to be estimated is disproportionally increased From which level of detection probability variation do benefits overcome the costs of accounting for detection probability in the sampling design and in analysis (task for WP2.2?)?

Many statistical models have been proposed to analyse monitoring data (Thomas, 1996). Which method should be used?

	Advantages	Disadvantages
Chained ratios Nt+1/Nt	Rapid update of index Index unchanged for past years after update	Spurious trends due to random drift Missing points cannot be accounted for and entire time series have to be discarded: data use not optimal
Time series analysis	Temporal autocorrelation (e.g. density-dependence, time lag) can be accounted for	Monitoring time series are usually too short Missing points cannot be accounted for and entire time series have to be discarded: data use not optimal
Generalized Additive Model	Non-linear trends can be identified (e.g. phenological pattern, temporal decrease followed by an increase)	Difficulty to distinguish anecdotic trajectories from general trends (task for WP2.2?) Methodological developments are needed to include within-year non-linear patterns (e.g. phenology) in among-year analysis of trend (task for WP2.2?) How do we combine non-linear trends among species or countries (task for WP2.2? WP4? WP5?)?
Mixed Model	If random choice of sites, site effect is modelled as a random effect, what decreases the number of parameters to be estimated	Random effect modelling is not available for all types of models (e.g. under development for capture-mark-recapture models; task for WP2.2?)

On the whole, the prevalence of the previously identified statistical problems should be evaluated so that we can take it into account when making recommendations on optimal monitoring methods. If several methods exist to account for a problem with high prevalence, we should expertise these methods so that we can make recommendations on their use.